Can Observers Predict Trustworthiness?*

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

May 9, 2010

Abstract

We investigate whether experimental subjects can predict behavior in a prisoner's dilemma played on a TV show. Subjects report probabilistic beliefs that a player cooperates, before and after the players communicate. Subjects correctly predict that women and players who make a voluntary promise are more likely to cooperate. They are able to distinguish truth from lies when a player is asked about her intentions by the host. Subjects are to some extent able to predict behavior; their beliefs are 7 percentage points higher for cooperators than for defectors. We also study their Bayesian updating. Beliefs do not satisfy the martingale property and display mean reversion.

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

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

*We are grateful to Joao Santos Silva, several seminar audiences, and in particular an anonymous referee for constructive comments. Many thanks also to Shane Frederick for providing us with the cognitive ability test, to the University of Essex and USE (Utrecht) for financial support, and to CREED for use of their subject database. V. Bhaskar thanks the ESRC Centre for Economic Learning and Social Evolution for support. Jeroen van de Ven thanks the Netherlands Organisation for Scientific Research (NWO) for support under grant Veni 451-05-005.

†Nuffield College, Oxford University, New Road OX1 1NF, Oxford, United Kingdom. E-mail: Michele.Belot@nuffield.ox.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 scope for opportunistic behavior. This highlights the importance of trust and trustworthiness in sustaining efficient economic transactions. Using cross-country data, Knack and Keefer (1997) find that trust raises economic growth, while La Porta et al. (1997) find that it is associated with lower corruption and judicial efficiency.

As important as the average level of trustworthiness or trust is the extent to which an individual can judge how trustworthy is his partner in a transaction. Numerous experiments show that individuals are heterogeneous in their willingness to cooperate (e.g. Blanco et al., 2006). This raises the question, can one trust a particular individual? 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 the overall level of trustworthiness is low and individuals find it hard to distinguish agents with different propensities, then one-shot transactions can often be inefficient and one must rely either on repeated interaction or contractual mechanisms to deter opportunism. This suggests another reason why trust is high between individuals who are socially close (see Glaeser et al., 2000), since they may be able to "read" one another.

This question is also related to a strand of evolutionary game theory that 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 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, the scope of our enquiry is wider than 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 of wether trustworthiness is distinguishable from opportunism, there is little prior work on the subject. The closest is work by psychologists on whether experimental subjects can distinguish true statements from false ones – we discuss this work shortly. 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. We overcome this problem by using DVDs of a game show, where two players play a prisoners' dilemma game. Our experimental subjects watch the show and are asked to predict behavior. This has two advantages. First, players on the game show freely choose their decisions, so that they incur any psychological or moral costs associated with opportunism. Second, their decisions are not anonymous but publicly observable. We observe perfectly whether a player is opportunistic or trustworthy, further enhancing the psychological and moral costs of opportunism. 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.

Prior to playing the prisoner's dilemma, the two players communicate verbally and face-to-face, allowing them to convey the honesty and sincerity of their intentions. Our experimental subjects are shown the game show players until the crucial decision, and asked to report their beliefs, i.e., the probability that a player shares. Subjects make predictions 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 belief, and allows us to study (Bayesian) updating in the context of a natural and complex problem. In addition, we also elicit each subject's *prior belief* 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., 2010).

Our main findings are as follows. Subjects appear to use the right cues forming 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 reveals 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"[1] – our subjects are able to read these responses and distinguish truth from lies. In comparison to situations where players do not make any promise, subjects 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

[1] 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.

involuntary truth-telling hypothesis seems to be valid for elicited promises.

By using these signals of cooperation, our subjects can to some extent distinguish cooperators from defectors – the average player who shares induces subject beliefs that are 7 percentage points higher than the average player who grabs.[2] This estimate is significant, in view of the fact that our subjects are untrained and only exposed to the players for a short time, and since predictions by non-professional subjects have historically generally been only barely 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 on how subjects perceive player 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 deception. Gneezy (2005) studies the factors that makes a person more likely to deceive. Wang et al. (2006) study pupil movements to see whether deceivers send involuntary signals. The flip side of this question, whether deception is detectable, has been the preserve mainly of psychologists. In the typical psychology experiment,[3] an actor is instructed to lie or to tell the truth, and an observer assigns a truth value to the actor's statement. Ekman and Friesen (1974) showed nurses a movie that could be

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

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

pleasant or nasty; in either case, the nurse was instructed to tell an interviewer that the movie was pleasant. These interviews were videotaped, and shown 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.[4]

The general finding is that observers are not able to detect lies. Their success rate in classifying statements is usually not significantly higher than 50% (chance). DePaulo et al. (1985, p. 327), conclude that "deception accuracy usually exceeds chance, although rarely by an impressive margin." For instance, the widely cited study by Ekman and O'Sullivan (1991) reports an accuracy of 3 percentage points above chance. Estimates range from about 42% accuracy to 68% accuracy, but only a small minority of about 10% of the studies find an accuracy of at least 10 percentage points above chance (Vrij, 2008). In his review of the literature since 1980, Vrij (2008) finds an average accuracy of around 54% among nonprofessionals. However, people with training, such as federal officers and secret service agents, manage to do better (Ekman et al., 1999).

Since the "deceivers" in these experiments have been instructed to lie, this may reduce guilt and make detection harder. This problem is avoided by Mann et al. (2002), who 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 Dawes et al. (1977), Frank et al. (1993) and Brosig (2002), who let subjects communicate before playing a prisoner's dilemma game, and asked 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. Dawes et al. (1977) report an accuracy of 3 percentage points above chance in their exper-

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

iment 1, and an accuracy below chance in experiment 2. Frank et al. (1993) find an accuracy of 11 percentage points above chance, and Brosig (2002) 8 percentage points above chance. However, their attribution of what constitutes "chance" when the true cooperation rate differs from 50% is questionable. If the cooperation rate differs from 0.5, subjects could outperform this chance measure by simply predicting cooperation or defecting every time. Using this benchmark of "chance", the observers in these studies do not do better than chance.[5] 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, they argue that this success can be explained by objective features, such as the length of the bargaining negotiations, rather than the ability to detect involuntary signals.

Our setup has several methodological advantages over existing studies. The main methodological differences are as follows:

1. Players on the game show freely choose their decisions, so that they incur any psychological or moral costs associated with opportunism.

2. We observe perfectly whether a player is opportunistic or trustworthy. Compared to a fully experimental design, using the game show has the advantage that the stakes for the players are high, giving real incentives to appear trustworthy.

3. We ask subjects to report probabilistic beliefs rather than a binary variable, and provide incentives such that reporting true beliefs is optimal. This enables us to measure more precisely which cues subjects use to predict sharing, and how accurate their predictions are.

4. Each subject sees a random sample of episodes; we do not select for a certain proportion of cooperators (or truth-tellers). Subjects could therefore have a prior based on how likely it is that players share in such a large stakes environment on TV.

[5] In existing studies (e.g. Frank et al. 1993; Brosig, 2002), chance is defined as $\gamma = pq+(1-p)(1-q)$, where p is the reported fraction of cooperators and q is the actual cooperation rate. If, for instance, $q > 0.5$, then subjects can achieve a success rate of q by always predicting cooperation, and this outperforms γ, since then $q > (1-q)$, hence $q(1-p) > (1-q)(1-p)$ and thus $q > \gamma$. Intuitively, if the actual probability is above one half, then without any further information, the best prediction is cooperation, and mixing between predicting cooperation and defection yields a lower accuracy.

5. Each subject makes predictions for a sample of players. 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.

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

3 Background and description of the data

3.1 The game show

We showed our subjects edited episodes of a TV game show, focusing on the final stage of this game where two players play a prisoner's dilemma game.[6] 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 his opponent's action are depicted in Fig. 1.

[INSERT FIGURE 1 ABOUT HERE]

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 value of X is €1,683, so that the stakes are substantial. Tables 1 and 2 present summary statistics of the player characteristics and outcomes of the game. We distinguish between lead and chosen players. Prior to the final stage, players earn their score by answering quiz questions. The *lead* player is the player with the highest score, and that player selects the *chosen* player from the other remaining players to enter the final stage. A third of players is female and the mean age is 33 years. Women tend to answer less quiz questions and are for that reason less likely to end up as lead player. Players share 43% of the time, and their decisions are uncorrelated, so that in 19% of the episodes it happens that both players share.

[6] 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?'. For a more detailed description of the game show, see Belot et al. (2009, 2010).

[INSERT TABLE 1 ABOUT HERE]

[INSERT TABLE 2 ABOUT HERE]

3.2 Experimental set-up

We ran a total of seventeen sessions, 5 of them in Utrecht in May 2006 with 89 subjects, and 12 in Amsterdam in May 2007, with 80 subjects. 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. Most of the show may not be relevant for predicting the decisions made by the final two players. For this reason, and 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. An analysis of the data confirmed that watching the entire game show did not make a difference to the predictions made by subjects.[7] Therefore, 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.

The show was paused several times, at which points we asked subjects to make predictions (see Fig. 2). The first time was just before the selection decision is made, where we asked subjects to predict which player would be chosen for the final.[8] They were asked to assign a probability of each of the two players shares, at two points – before the players made speeches to each

[7] For interested readers this analysis is available upon request from the authors.

[8] These selection predictions are used for another project (see Belot et al., 2009) and we do not discuss them further here.

other, and after.[9] Subject choices were restricted to a discrete grid, with step size 0.1, i.e. the set $\{0, .1, ..., .9, 1\}$. They 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.[10] Subjects were not given any feedback on the actual decisions made by players, so as to prevent learning. We are also confident that the subjects are very unlikely to have seen the episodes they were asked to make predictions about.[11]

[INSERT FIGURE 2 ABOUT HERE]

At the end of the session, we ascertained personal details of the subjects and asked them to estimate the average cooperation rate over all the 69 episodes of the show. They were paid €1 if their prediction lay within 5 percentage points of the true value. Answers and earnings were private and not divulged to other subjects. 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.[12] We asked the Amsterdam subjects if they would like to donate a part

[9] 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. To some extent, this is akin to the strategy method. Brandts and Charness (2009) review the literature comparing decisions with and without strategy method, and report that most studies find that the strategy method does not lead to different behavior than the more standard direct-response method. Comparing the Utrecht and Amsterdam sessions, we find no difference in the average prediction rate, it is equal to 50% in both locations.

[10] 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.

[11] There was a five-year lag between the experiment and the game show. The show was not very popular (it was broadcast in the afternoon, and lasted only one season). We also asked subjects if they had seen the show – very few subjects said yes, and these could not remember any specific show or contestants.

[12] Cooperation rates were higher in the Amsterdam sample (59%, against 45% in Utrecht). We cannot tell wether this is due to, e.g., different incentives or because they are matched with a player on the show (in Utrecht) or with another participant of the experiment (in Amsterdam).

of their earnings to a charity, Warchild.[13] 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 letting subjects choose between a fixed amount (€2) and a series of lotteries with varying stakes. The complete instructions are in the Appendix.

4 Perceived cooperative traits

Players on the game show share 43 percent of the time. This is exactly what the subjects expect on average when asked to report the average sharing probability across all 69 episodes (see Table 3). However, their average prediction is 9 percentage points 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. Possibly, this effect is even larger for players on the show, as they experience the presentation first-hand.[14]

As we will see, the content of communication is strongly predictive of predictions, but consistent with Bayesian updating, the average prediction is the same before and after communication, so that on average subjects do not revise their estimates in a systemically biased way. However, we find strong violations of Bayesian updating if we compare predictions before and after communication for different levels of the predictions (see Section 5).

[INSERT TABLE 3 ABOUT HERE]

[13] 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.

[14] We explored this exposure effect by comparing predictions of subjects that saw short or long versions of episodes, but found no difference. We also note that different elicitation methods were used for the overall prior and predictions: quadratic scoring rule for reported beliefs, and a fixed amount if within 5 percentage points of the correct answer for the prior. While we cannot exclude the possibility that this can account for the difference, this seems unlikely because the scoring rule gives minimal distortions in our setup (see also footnote 20).

4.1 Cooperative cues

We now investigate how subjects form their beliefs and how they update their prior in response to information. Our empirical strategy is based on the random assignment of episodes to groups of subjects, and on the panel aspect of the data. Subjects were asked to report predictions for several players, with different characteristics. We can therefore identify how subjects update their beliefs in response to this information, using the within-subject variation. This may be compared with the analysis of how actual behavior varies with characteristics, conducted in et al., 2009).

We model the prediction of subject i regarding player j as:

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

where p_{ij} is the prediction of subject i regarding player j,[15] 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 4 reports the results from the analysis based on the data from the game show (see et al., 2009). The reported coefficients are the marginal effects of the probit estimates. Women are almost 20 percentage points more likely to share than men,[16] and those who contribute relatively little to the final prize money are more likely to share. A larger stake slightly increases sharing, while age and attractiveness 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 correctly 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

[15] This is the final prediction, after communication. This makes it easier to compare the results with those from our analysis on the decisions in the game show (since the decisions are taken after communication). The results are similar if we use predictions made before communication

[16] Such gender effects are not new. See Belot et al. (2009) for a discussion how they relate to the existing literature.

there is a strong relationship in the actual data. Finally, they correctly do not associate age or attractiveness with cooperativeness.

[INSERT TABLE 4 ABOUT HERE]

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. It has been established that communication is very effective in fostering cooperation (Sally, 1995). In particular, promises are informative about intended behavior, because subjects are reluctant to lie (Gneezy, 2005).

In an analysis of the decisions of players, et al. (2010) find that promises

are very informative about the behavior of players. However, not all promises are alike in this respect. *Voluntary* promises, i.e. those that are at the player's initiative, are highly correlated with sharing behavior. In some cases, the presenter explicitly asks a player if he or she intends to share. Subjects invariably respond affirmatively to this question (with one exception), and we label these as *elicited* promises. Elicited promises are uncorrelated with actual sharing decisions. Table 5 provides descriptive statistics on promises, and the percentage of players that share by types of promises. 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". Out of all promises, about 30% are elicited promises. The most striking fact is that those who make a voluntary promise are almost *50 percentage points more likely* to share than those who do not.

Do the predictions by subjects reflect the above findings? Table 6 reports summary statistics on predictions. Individual predictions are reported for both before and after communication by type of message and decision of

players. From this, we see that mean beliefs are somewhat higher after voluntary promises than after no promises or elicited promises, but the difference is small. Interestingly, beliefs are similar before and after communication when players make no promise or a voluntary promise. This is very different after elicited promises. In that case, we see that the predicted probability of sharing increases by 10 percentage points after a true elicited promise, while it decreases by 4 percentage points for players who make false elicited promises.

Table 6 also shows the accuracy level of predictions before and after communication, where accuracy is measured as p if a player shares, and $1-p$ if a player grabs. By this simple measure, the accuracy rate is higher for true promises than false promises, and the effect is biggest for voluntary promises.

To summarize this, we find that beliefs are more optimistic after voluntary promises, and the accuracy is highest after voluntary true promises. By contrast, comparing before and after communication, accuracy of predictions increase most after elicited promises, suggesting that subjects can identify liars better in this case. This is possibly a consequence of the "Columbo effect"; if subjects are surprised by the question of the presenter, they may find it harder to disguise lies.

[INSERT TABLE 5 ABOUT HERE]

[INSERT TABLE 6 ABOUT HERE]

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. We estimate the following equation:

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

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 7 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 when a player make a promise, but the effect is only significant for voluntary promises. This suggests that subjects understand that lying after making a voluntary promise is psychologically more costly than after an elicited promise. Possibly, subjects anticipate stronger feelings of guilt in this case (as for instance in Charness and Dufwenberg, 2006), but in any case this is evidence against a pure cost-of-lying *per se* story that does not differentiate between types of lies. The effect is relatively small though - predictions increase by about 5 percentage points for voluntary promises in comparison to the average increase of 50 percentage points in the game show data. So, overall, subjects fail to capture the magnitude of the effect of voluntary promises.

Since some players who make a promise choose to grab, column 2 investigates whether subjects are able to identify liars. Subjects are not able to distinguish truth from lies when they see a voluntary promise, but are able to do this very well when the promise has been elicited. There could be several reasons for this. Those who initiate a promise may be better liars, or lies prompted by a surprise question 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.

These results are related to those of Charness and Dufwenberg (2006, 2010) who conduct experiments on a trust game where subjects can create and send open format messages, finding that personalized promises are effective, while impersonal promises are not.

Since communication is sequential, there 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, the second player's speech or behavior does not depend on whether the first player makes a promise or not. We find no systematic differences in the predictions made for players 1 and 2; these results are not reported for the sake of brevity. However, we do find a significant correlation (.25) in predictions made for both players, so that subjects believe

that the players' decisions are correlated.

[INSERT TABLE 7 ABOUT HERE]

4.3 Overall quality of predictions

We now consider how well subjects predict the sharing decision. Overall, they correctly predict an average cooperation rate of 43%. However, it can well be that subjects have learned what the average cooperation rate is, but still find it very difficult to predict behavior for any specific individual. As they tend to underestimate the importance of several cues, the accuracy of predicting a specific player's behavior is lower than what is possible. To examine how well subjects predict, we regress final beliefs upon the sharing decision, while including subject fixed effects and player random effects, in the form of the following equation:

$$p_{ij} = \alpha + \delta_i + \beta share_j + \varepsilon_{ij},$$

where α is a constant, δ_i is a subject-specific fixed effect and $\varepsilon_{ij} = \eta_j + \xi_{ij}$ (that is, we cluster standard errors at the player level). Because subjects make multiple predictions, we can filter out the subject's prior and directly estimate how beliefs differ for players who end up sharing in comparison to players who end up grabbing. The null hypothesis corresponding to random reports (chance) is $\beta = 0$. Thus, $\widehat{\beta}$ provides a direct measure of accuracy. Our estimated value for β is .07, and this is significant at the 1% level. Thus, if a player shares, this is associated with a 7 percentage points increase in subject beliefs.

Our subjects are untrained and exposed only briefly to the players. In addition, there may be some measurement error if some subjects do not truthfully report beliefs. To the extent that some subjects report random beliefs instead of truthful beliefs, this will bias the results towards chance levels. These factors suggest that our estimate is a lower bound on what subjects can achieve.

Because our context differs from existing studies in important respects, it is not straightforward to compare findings to the existing literature. As pointed out before, previous studies rely on binary reports. Subjects will only report different beliefs if, upon observing a certain cue, their posterior crosses the

threshold level of reporting 0 versus reporting 1. With these coarser data, valuable information is lost needed to identify the cues that subjects use, and will limit the preciseness of accuracy measurements. If we coarsen our belief variable by classifying a belief greater than 0.5 as 1, and that less than 0.5 as 0, the accuracy level of our subjects is 0.52, i.e. no different than chance by any reckoning of chance. This suggests that previous studies may have underestimated the ability of subjects to predict behavior, by using coarse predictions.[17]

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, reports 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 forms a subjective prior belief μ_i about the probability 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 $\pi_i(s)$:

$$\pi_i(s) = \frac{\mu_i \ell_i(s)}{\mu_i \ell_i(s) + (1 - \mu_i)}, \tag{3}$$

[17] On the other hand, we have argued in section 2 that some studies (e.g. Frank et al., 1993, and Brosig, 2002) have overestimated the ability of subjects, by using a wrong benchmark of what constitutes chance. On balance, these concerns point to the importance of a suitable methodology.

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:

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

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:

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

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:

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

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 8 shows the results. For the game show data we can estimate the coefficient on promises by comparing sharing behavior between subjects that do and do not make a promise (cross-section), and for beliefs by subjects we can either estimate the coefficient based on predictions reported after communication only

(cross-section) or by comparing final versus interim beliefs reported by the same subject and for the same player.

The results confirm what we have found with the linear specification. Overall, subjects do perceive the correct signals, but underestimate their magnitude. However, the gender coefficient we estimate from subjects beliefs (0.62) is not significantly different from the gender coefficient estimated from the players' actual decisions (0.78). Therefore, the update of beliefs in response to gender is consistent with Bayesian updating.

[INSERT TABLE 8 ABOUT HERE]

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 estimate a linear specification, while allowing the coefficient on characteristics to vary depending upon the prior. These results are reported in Table 9. 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 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.[18]

[18] 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.

[INSERT TABLE 9 ABOUT HERE]

5.2 Testing the martingale property of beliefs

Under Bayesian updating beliefs satisfy the martingale property: the prior must equal the weighted average of posteriors, and can be written as:

$$\mu_i = \sum_s \Pr(s)\pi_i(s), \tag{4}$$

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

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

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.[19] In other words, we have mean reversion in beliefs, and a violation of the martingale property.

[INSERT FIGURE 3 ABOUT HERE]

The hypothesis of equality of priors and average posteriors can also be tested by comparing interim beliefs (i.e. predictions before communication) and final beliefs (after communication). Since the equality in (4) 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

[19] 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.

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 almost *any* value of interim belief. For low interim beliefs, the average change is positive, while for high interim beliefs, it is negative. That is, final beliefs are systematically higher than interim beliefs when interim beliefs are low and systematically lower when the interim belief is high.[20] 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 changes in updates are larger when subjects see a promise.[21]

[INSERT FIGURE 4 ABOUT HERE]

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 subject i in instance k is given by

$$p_{ik} = \hat{p}_{ik} + \varepsilon_{ik}, \tag{5}$$

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

[20] 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).

[21] Note that we chose to pay subjects for all predictions, rather than choosing one at random. The advantage of paying for all predictions is that distortions due to probability weighting are minimized (see Offerman et al, 2009). Moreover, the existing experimental evidence does not find support for hedging behavior (Blanco et al, 2008).

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), 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.[22] 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. A Bayesian subject 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? A second question is whether subjects who are better at predicting differ in a systematic way from others.

[22] 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.

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 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 percentage points of the true value.

The prior and the subject mean prediction are correlated, with a correlation coefficient of 0.38 that is significant at the 1% level. We show the results in Table 10. Columns (1) and (2) are based on the whole sample, columns (3) and (4) are based on the Amsterdam sample only. We find 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.[23] That is, subjects who are more cooperative also believe that others are more cooperative, and is consistent with findings in Glaeser et al. (2000), using survey data.

[INSERT TABLE 10 ABOUT HERE]

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; Charness and Rabin, 2002). Second, people may project their preferences onto others, thinking that the average person behaves like them. (see e.g., Messe and Sivacek, 1979). This 'false consensus effect' is possibly a result of rationalizing behavior stemming from cognitive dissonance (Festinger, 1957) if subjects can make themselves feel better by believing that others would behave similarly.[24]

[23] 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.

[24] 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.

The second question is, who are the subjects who are better at predicting behavior. Since a subject's earnings is a measure of the quality of her prediction, we estimate the following equation:

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

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

Table 11 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[25]), 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 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.

[INSERT TABLE 11 ABOUT HERE]

7 Conclusion

We examine 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 somewhat predictable. 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

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

is made in response to an explicit question by the presenter, in line with the involuntary truth telling hypothesis. Overall, our untrained subjects assign a 7 percentage points higher probability of cooperation to cooperators as compared to defectors, suggesting that opportunism is indeed to a certain extent detectable even by naive subjects. There may be several reasons why subjects underestimate the magnitude of several signals. One possibility is that subjects are only briefly exposed to the players on the show, and not all facial and non-facial expressions are visible to the audience. This makes it harder to observe certain types of signals, such as those related to lying. In addition, subjects are relatively unfamiliar with the environment. How the accuracy of subjects vary with these and other factors is left for future research.

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 up to the point where the lead player chooses the other finalist (see Fig. 2). They then predict the selection decision, and also predict the sharing decision for every player. Since 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. They then watch the show up to the point where the two finalists had to make their decision to share or grab, and make predictions again.

In the Amsterdam treatment, we only asked subjects to predict which of the two contestants would be selected to play the final round – sharing decision predictions were not asked for at this point, but only after the selection decision was made. They watch the show up to the communication stage, and make predictions again.

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.)

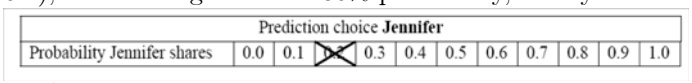
Instructions You will see 6 shortened episodes. In 2 episodes you'll see one round of trivia questions, in the other 4 we skip all trivia rounds. We start by showing the beginning where the 3 relevant candidates introduce themselves.

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., "The aggregate burden of crime," *Journal of Law and Economics* 42: 2 (1999), 611–642.

[2] Belot, M., V. Bhaskar, and J. van de Ven, "Promises and Cooperation: Evidence from a TV Game Show," *Journal of Economic Behavior and Organization* 73 (2010), 396-405.

[3] Belot, M., V. Bhaskar, and J. van de Ven, "Beauty and the Sources of Discrimination," available at SSRN: http://ssrn.com/abstract=956600 (2009).

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

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

[6] Blanco, M., D. Engelmann, A. Koch, and H. Normann, "Belief elicitation in experiments: is there a hedging problem?," IZA Discussion Paper No. 3517 (2008).

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

[8] Brandts, J., and G. Charness, "The Strategy Method: A Survey of Experimental Evidence," mimeo (2009).

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

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

[11] Charness, G. and M. Dufwenberg, "Promises & Partnership," *Econometrica,* 74 (2006), 1579-1601.

[12] Charness, G., and M. Dufwenberg, "Bare Promises: An Experiment,", forthcoming *Economics Letters* (2010).

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

[14] Charness, G. and M. Rabin, "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics* 117:3 (2002), 817-869.

[15] 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 (1977), 1-11.

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

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

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

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

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

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

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

[23] Ekman, P. and M. O'Sullivan, "Who can catch a liar?," *American Psychologist* 46:9 (1991), 913-920.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[40] 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.), *Heuristics and Biases* (New York: Cambridge University Press, 1992).

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

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

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

[44] Offerman, T., J. Sonnemans, G. van de Kuilen, and P. Wakker, "A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies* 76:4 (2009), 1461-1489.

[45] Sally, D., "Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992," *Rationality and Society* 7 (1995), 58-92.

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

[47] Vrij, A., "Detecting Lies and Deceit," 2nd ed., (Chichester: John Wiley and Sons, 2008).

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

Table 1: Individual characteristics of players

	lead player	chosen player
N obs.	69	69
Mean age	33.9	32.1
Share women	22%	52%

Table 2: Distribution of outcomes and stakes

Outcome	Frequency	Median stake (€)
Percentage sharing	43%	1,683
Both share (S,S)	19%	3,090
One shares (G,S)	48%	1,533
Both grab (G,G)	33%	1,850

Table 3 - Sharing probabilities - Summary statistics

Sharing probability	Mean	sd
Actual game show data ($N = 138$)	.43	.50
Prior ($N = 169$)	.43	.21
All predictions before communication ($N = 1,672$)	.52	.26
All predictions after communication ($N = 1,672$)	.52	.29

Table 4 - Actual cues and perceived cues

Dependent variable: Sharing decision	Actual realizations (1)		Beliefs (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 5% and 1% significance levels respectively. Col. (1): bivariate probit estimates, marginal effects. Col. (2): OLS estimates, with standard errors clustered by player.

Table 5 - Communication and Sharing - Summary statistics

	% players in category	fraction sharing
Contents communication		
All		.43 (.50)
No promise ($N = 74$)	54	.28 (.45)
Promise ($N = 64$)	46	
Of which:		
voluntary ($N = 19$)	70	.73 (.45)
elicited ($N = 45$)	30	.26 (.45)

Notes: elicited promises are promises in response to question by presenter. Standard deviation in parentheses.

Table 6 - Predictions by subjects- Summary statistics

	Belief			
Prior ($N = 169$)	.43 (.21)			
Individual predictions	Before communication		After communication	
	average belief	rate of accuracy	average belief	rate of accuracy
All predictions ($N = 1,672$)	.52 (.29)	52	.52 (.30)	52
No promise - shares ($N = 268$)	.53 (.25)	53	.51 (.31)	51
No promise - grabs ($N = 642$)	.52 (.26)	48	.50 (.29)	50
Voluntary promise - true ($N = 367$)	.56 (.25)	56	.56 (.30)	56
Voluntary promise−false ($N = 247$)	.55 (.26)	45	.53 (.29)	47
Elicited promise − true ($N = 35$)	.43 (.28)	43	.53 (.31)	53
Elicited promise − false ($N = 113$)	.56 (.25)	44	.52 (.28)	48

Notes: Belief is prediction of sharing. Prior is average reported belief over all episodes. Stand. dev. in parentheses. Accuracy is measured as p (*/1-p*) if a player shares (/grabs)

Table 7 - Communication and beliefs

Dependent variable:	Difference in sharing 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)***	-.03	(.01)***
R-squared	.008		.014	
N. obs	1672		1672	

OLS estimates; Standard errors are reported in parentheses. ** and *** correspond to 5% and 1% significance levels respectively

Table 8 - Actual and perceived cues (log likelihood ratios)

Dependent variable: sharing decision	Actual realizations		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 < median score	-.09	(3.61)	-.35	(.17)*		
Attractiveness > 4	-.18	(.26)	-.13	(.19)		
Voluntary promise (cross-section)	2.12	(.33)***	.21	(.17)		
Voluntary promise (final vs interim beliefs)					.36	(.10)***
N. obs	138		1672		1672	

Column (1): Log ratios derived from probit estimates and standard errors computed with delta method. Column (2) OLS estimates with subject fixed effects. Standard errors in parentheses, * and *** denote 10% and 1% significance levels respectively. Standard errors are clustered by show contestant in col (2). Voluntary promise: *cross-section* based on comparison players that make or do not make promise; *final vs interim* beliefs based on subjects reporting player specific beliefs before and after communication.

Table 9 - Actual cues and perceived cues depending on the prior

Dependent variable: sharing decision	Actual realizations		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, *, ** correspond to 10% and 5% significance levels respectively.Standard errors are clustered by player in col (2).

Table 10 - Subject characteristics, prior and actual predictions

	All sample				Amsterdam sample			
Dependent variable:	Beliefs		Prior		Beliefs		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	

OLS estimates; beliefs are individual predictions; prior is the estimated average reported for all players in the game show, Standard errors are reported in parentheses *, ** and *** denote 10%, 5% and 1% significance levels respectively. Standard errors are clustered by show contestant in col. (1) and (3).

Table 11: 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	-.040	(.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	

OLS estimates including player fixed effects. Standard errors are reported in parentheses, * and ** correspond to 10% and 5% significance levels respectively and are clustered at the subject level.

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

Figure 1: Monetary Payoffs

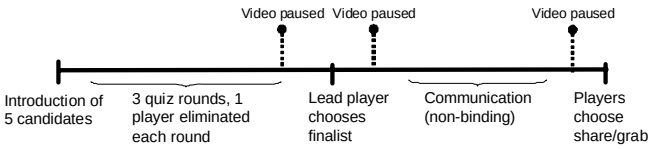


Figure 2: Timeline game show

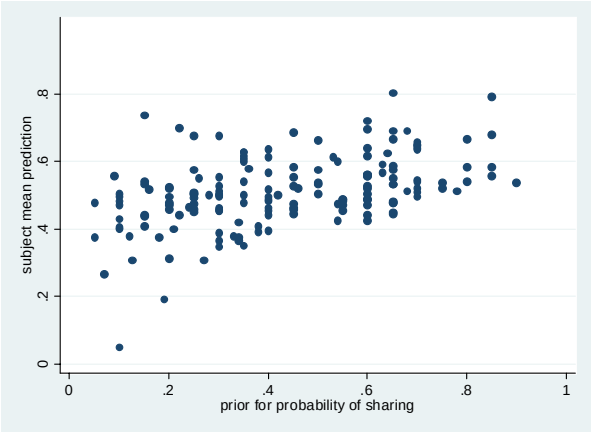


Figure 3: Prior and subject mean prediction

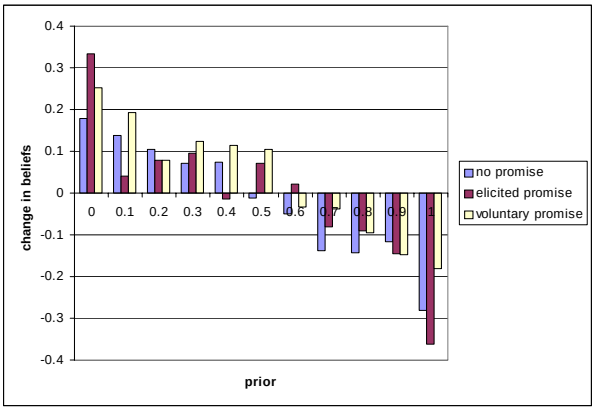


Figure 4: Prior and mean change predictions