Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect[1]

Miguel A. Costa-Gomes (University of Aberdeen) Steffen Huck (University College London)
Georg Weizsäcker (DIW Berlin & University College London)

January 2010

Abstract: In many economic contexts, an elusive variable of interest is the agent's expectation about relevant events, e.g. about other agents' behavior. Recent experimental studies as well as surveys have asked participants to state their beliefs explicitly, but little is known about the causal relation between beliefs and other behavioral variables. This paper discusses the possibility of creating exogenous instrumental variables for belief statements, by shifting the probabilities of the relevant events. We conduct trust game experiments where the amount sent back by the second player (trustee) is exogenously varied by a random process, in a way that informs only the first player (trustor) about the realized variation. The procedure allows detecting causal links from beliefs to actions under plausible assumptions. The IV estimates indicate a significant causal effect, comparable to the connection between beliefs and actions that is suggested by OLS analyses.

Keywords: Social capital, trust game, instrumental variables, belief elicitation

JEL Classification: C72, C81, C91, D84

[1]We thank participants of the CEMMAP/ELSE workshop on distorted beliefs as well as seminar audiences at Autonoma, DIW Berlin, Exeter, Jena and Paris I for their comments. We are grateful for the financial support of the U.K. Economic and Social Research Council (ESRC-RES-1973) and the ELSE centre at UCL. The experimental sessions were conducted with the excellent support of Rong Fu, Tom Rutter, Brian Wallace and Mark Wilson. Contact: m.costagomes@abdn.ac.uk, s.huck@ucl.ac.uk, g.weizsacker@lse.ac.uk

1 Introduction

In subjective expected utillity theory and related models, the agent's expectations can be viewed as a pure *as if* construct, meaning that the expectations are no more than an elegant, low-dimensional way of summarizing choice data. According to this view of expectations, choice is represented by a hypothetical optimization problem that involves maximizing a function of expectations — for example, the expected utility function. But choice is the fundamental concept, and any additional assumption that one may make about expectations is really an assumption about what choices are made. A much more literal interpretation of expectations is that they are *real*, meaning that they are independent entities that have some physical incarnation and that can in principle be accessed directly, for example, by asking people to state them. Much can be said in favour of such an interpretation, for instance that humans are able to express expectations even about variables that are irrelevant for their choices. But if expectations are independent entities, one should be able to influence them and thereby measure their effect on choices. This leads to the straightforward empirical question we shall address in this study: are choices driven by beliefs?

This question has important consequences for policy interventions because its answer determines whether one can induce efficient outcomes by way of changing people's expectations, as attempted in many policy campaigns. Especially interventions that are geared towards creating trust or optimism have this rationale, relying on the self-fulfilling powers of such sentiments: if the policymaker can induce the agents to be optimistic and trusting about future outcomes, their subsequent choices may, collectively, justify the optimism and repay the earlier trust. By instigating an initial increase in trust and optimism, the policymakers may therefore create a shift from a less desirable outcome to a better one.

But the role of beliefs first needs to be affirmed. Researchers have turned to belief elicitation procedures where the agents state their expectations explicitly. Especially trust game experiments (following Berg, Dickhaut and McCabe, 1995) provide a frequent context for such methods. Fehr *et al.* (2003), Bellemare and Kröger (2007), Sapienza, Toldra and Zingales (2007) and Naef and

Schupp (2009), among others, ask the participants in their experiments to state expectations on how much money other players will return if trusted, and find a strong correlation, as well as much explanatory power, when regressing the level of trusting behavior on stated expectations. We note that the trust game lends itself well to investigations of the role of beliefs, because there is substantial behavioral variance both in the level of trust and in the stated expectations. Yet it remains unanswered whether the variance in trusting behavior arises *because of* the variance in stated beliefs, or whether the co-variation in the two variables is driven by other, omitted variables that capture unobservable differences between the participants.

Under the latter hypothesis, one natural candidate for an omitted variable is the perception of social norms. Take a simple two-player trust game experiment: the trustor invests in a joint venture, and the trustee can either appropriate the investment and its return, or repay the first player's trust by sending some money back. Among the experimental participants who are assigned the role of trustors, presumably some view a high investment as the "right" thing to do, since this is the choice that maximizes the social surplus and thus opens up the possibility of efficient, mutually beneficial exchange. Whether or not such social norms influence the investment choices may depend on multiple factors that are unobservable, e.g. the participants' education, cultural factors, or indeed the framing employed in the experiment. But such unobservables will influence actions *and* beliefs. In particular, it may be that the same participant who invests a large amount also predicts that the other particpant will return a large amount because that, too, is arguably the "right" thing to do. That is, the unobservable perception of whether or not a social norm of mutual cooperation is relevant can generate a correlation between the belief statement about the opponent's behavior and the player's own investment choice — without implying anything about a causal influence of one variable on the other.

Such a correlation is not necessarily an artefact of some norm-sensitive behavioral rule but it can arise as an equilibrium phenomenon in a quite natural game of incomplete information. We develop a simple example to illustrate this in the appendix of this paper. Players interact in a mini trust game with just two actions for each player, trust or not, and reciprocate or not. Both players know that there is social norm that prescribes trust and its reciprocation, while self-interested

behavior would induce the opposite, inefficient outcome. There prevails some uncertainty about whether self-interested choices will be sanctioned — for example by the experimenter who may not stick to his promise to treat observations anonymously. Players receive signals about the likelihood of sanctions, e.g. from clues in the instructions or the experimenter's general demeanor. As both players read the same instructions and observe the same experimenter, these signals are correlated. It is shown that even with relatively little correlation between the players' expectations about norm enforcement, the Bayesian Nash equilibrium predicts a strong correlation between the trustor's own actions and her belief about the opponent's action. Both variables are driven by an underlying ommitted variable, which is the trustor's perception of the likelihood of norm enforcement. Our example also shows that despite the strong correlation between beliefs and actions, an exogenous shift of the trustor's beliefs about the opponent's action would have a relatively small effect on his action. In other words, in the example it would be misleading to interpret the strong correlation between beliefs and actions as saying that one drives the other, as both are indeed driven by the possibility of sanctions.

This example evidently only suggests one kind of potentially relevant omitted variable. Multiple omitted variables may be effective apart from social norms, and social norms may not even be the most relevant one. Our intended insight is merely that the participants playing the game may well have good reasons (here, play an equilibrium strategy in a more general game) to exhibit strong correlations between beliefs and actions that the researcher may easily mis-interpret as a causal relation. To measure the effect of a belief change on actions, one needs more powerful observations than the simple correlations.

In Section 2 we introduce a technique to measure the effect, presented in two separate experimental design variations. Both variations involve the artificial creation of instrumental variables in trust game experiments. These variables can be varied in a purely exogenous way, and thereby are suitable instruments for the endogenous belief statements.

The first instrument applies to continuous-choice situations, like Berg, Dickhaut and McCabe's (1995) trust game. It is a random "shift" that exogenously increases or reduces the trustee's level of re-payment. The realization of the random shift is known to the trustor, who as a result might

change her belief statement about the trustee's final level of re-payment, and might accordingly change her own action. The trustee is informed of the existence of the shift and of its distribution. However, she is not informed about the realization of the shift, and her behavior remains unaffected by the realization. Therefore, the trustor's belief about the trustee's behavior (her chosen level of re-payment prior to its manipulation through the shift) should also be unaffected by the realization of the shift. Our data conforms to these predictions. At the same time, the beliefs about the payoff-relevant event – the level of re-payment including the shift – react strongly to the exogenous variation, which is necessary to apply IV estimation. Regarding the second, "exclusion restriction" requirement of IV, that the instrument influences the actions only via the beliefs about the level of re-payment, we argue that it is natural to make this assumption because the instrument is an element of the statistic that the belief is formed about, and does not enter the interaction in any other way. The exogeneity of the shift also rules out that it is affected by any omitted variable and, conversely, it plausibly does not affect potentially influential variables like personal characteristics or perceptions of social norms.

The second instrument is an exogenously varied population of artificial agents, "robots", whose actions replace a trustee's action with some probability. This instrument is designed for use in the context of discrete-choice situations like matrix games (but would in principle also be applicable for continuous-choice problems) and we examine it in a separate experimental game. For both instruments, we conduct a series of tests to investigate whether the overall distributions of belief statements and actions changes significantly between treatments with and without the instrument. For the continuous-choice version of the trust game, we find that none of the relevant statistics are affected by the introduction of the instrument in any problematic way. For the discrete-choice game, however, we find that some caution is in place as the belief statments indicate some subjects might not have fully comprehended the procedure. We therefore put the main emphasis on the analysis of the continuous-choice version of the game.

After the discussion of invasiveness and appropriateness of the instruments, Section 3 proceeds with the analysis of the paper's main question, whether a variation in beliefs drives choice behavior. The analysis shows that the exogenous belief variation does indeed have a significant impact on

choices, in both experiments. In the continuous-choice game, the IV-estimated effect of beliefs on actions is not quite as strong as the naive OLS analysis suggests (which we carry out on a control treatment without an instrument, producing similar results to the previous literature). But as in OLS, the IV analysis finds an effect that is significantly different from zero, and the point estimate not too dissimilar from the OLS estimate (0.56 with IV versus 0.76 with OLS).[2] In the discrete-choice game, the IV estimate is much larger than the OLS estimate, with a point estimate of about tripled size. A further check for the influence of omitted variables, we conduct analyses where we include personal background variable that we collected in the laboratory — demographic variables as well as calculation skills and some attitudinal questions about trust. Including these controls does not affect any of the results, indication that at least in the limited set of controls that we have, there was no strong omitted variable that would artificially drive the correlation of beliefs and actions.

The findings constitute, to our knowledge, the first laboratory evidence that beliefs play a causal role in the determination of actions. As such, the results confirm the important role of beliefs not only in the trust game, but more generally in decisions under uncertainty. This role was implicitly ascribed to beliefs in experiments that use stated expectations (McKelvey and Page, 1990, Offerman, Sonnemans and Schram, 1996, Croson, 2000, Huck and Weizsäcker, 2002, Nyarko and Schotter, 2002, and many later studies) as well as survey evidence that solicits expectations about relevant market variables (see Manski, 2004, and Attanasio, 2009, and the literature cited there). An important set of close relatives to this paper are field experiments that vary informational conditions in different economic contexts, see e.g. Jensen (2008) and Dupas (2009). Under some additional assumptions about how information maps into beliefs and actions, one can interpret these studies as evidence for an effect of beliefs on actions. Our experiments complement this by

[2] A separate regression projects the participants' actions directly on the exogenous shift variable. There, we also find a significant effect of the shift on the action, see Section 3. The procedure of replacing one player's choice by an exogenous random move has been done in several experimental studies that investigate whether positively reciprocal actions appear only when a certain action is played by a human agent. See, in the context of the trust game, Cox (2004) among others. These studies, however, replace the trustor's action by a random move, whereas we manipulate the trustee's move.

allowing for a consistent estimate of the size of the effect and by offering results in a clean laboratory setting.

While the results are positive and provide supporting evidence to the previous supposition of many economists that beliefs do influence actions, one needs to be aware that results may be different in different settings, of course. As a general methodological point we therefore suggest openly addressing the difficulty of ascribing differences in behaviors to observed differences in stated expectations. Exogenous variation of artificially introduced instuments may be a way to go in other contexts as well, as the designer of experiments and surveys often have the freedom to create an appropriate variation. This is further discussed in the conclusion of the paper, appearing in Section 4.

2 Experimental design: Instrumental variables for belief statements

The experimental design revolves around two simultaneous-choice trust games involving two players. In one of them, both players have a continuum of possible actions (Continuous Trust Game, henceforth CTG), while in the other players only have two actions each (Mini Trust Games, henceforth MTG). For each game, and in addition to the choice data, we collect trustors' beliefs about the actions played by the trustees. In separate treatments both games are played with and without an instrumental variable that serves as a driver for the trustor's beliefs about the level of repayment. The games played without the instrument are control games that are needed for two reasons. First, they allow for an important empirical check of the validity of each instrument: whether or not the instrument affects behavior in undesirable ways. In field studies that involve IV methods, this is less of a concern because the instrument is usually part of the natural decision making environment. But with an artificial instrument, we should check that the instrument is neutral in the sense that its presence does not distort the data generating process. Second, and no less important, the control treatments provide a data source for OLS estimates and similar methods that do not make use of

the instrumenting technique and that serve as the comparison benchmark for our IV results.[3]

With two baseline games, CTG and MTG, and two conditions regarding instrument (I)/no instrument (NI), we create four games altogether, CTG/I, CTG/NI, MTG/I and MTG/NI, as in a 2x2 factorial design. In each experimental treatment each of the two baseline games is played once; one of them under the intrument condition, the other one without it. Thus, in the first part of each treatment we randomly and anonymously matched pairs of participants to play one of the baseline games under one of the instrument conditions, and in the second part we rematch the participants randomly and anonymously, and ask them to play the other baseline game under the other instrument condition. For example, subjects who first play CTG/I would then switch to MTG/NI. The participants receive no feedback between the two games. In the second match, we also switch the player role assigned to each participant, so that a trustor in the first game becomes a trustee in the second game and vice versa. Thereby, we can make sure not only that every participant is in exactly one treatment that involves an instrument but also that she is a trustor exactly once and is therefore asked to state exactly one belief about the opponent. Our balanced design allows us to check whether order effects influence play of the games for each of the player roles because each of the four games appears as the first game being played in one treatment, and as the second game in another treatment.

For the collection of belief statments, we employ a quadratic scoring rule, which is incentive compatible in the sense of theoretically eliciting the mean of the subjectively expected distribution, under the assumption that subjects are risk neutral.[4]

[3] Again, this is different from the field where both OLS and IV are normally run on the same data set. This would, however, not yield the desired insights in our context. Under the hypothesis that an omitted variable is at work in the trust game, an OLS analysis on the data with instrument does not identify the co-variation between the relevant beliefs and actions, as the beliefs are artificially perturbed. That is, in the treatment with instrument, the beliefs are formed about a different statistic, which is the return including the shift and is thus not the relevant statistic in the trust game without the instrument. The belief distributions and the nature of the beliefs may therefore differ between the treatments. If the true data generating process involves an omitted variable, then the OLS results may thus also differ between the two treatments. The analyst's interest arguably lies in the validity of OLS as a diagnostic tool in the treatment without instrument.

[4] The quadratic scoring rule has been used by numerous researchers and — although not all studies agree (see e.g.

We conducted our experimental sessions at University College London and at the University of York, with a roughly equal number of subects in each treatment at each location, as reported in Table 1. The appendix contains a sample of the instructions. In all, 434 experimental subjects participated in our sessions. Subjects earned points by playing two games and one belief elicitation task, which were then converted into money at an exchange rate of 40 points per £1. They were also paid a show-up fee of £5. Sessions lasted about 90 minutes from the moment they were seated until leaving the laboratory after collecting their payments.

Treatment	# York Participants	# UCL Participants	# Total Participants
CTG/I – MTG/NI	62	62	124
MTG/NI – CTG/I	62	58	120
CTG/NI – MTG/I	46	50	96
MTG/I – CTG/NI	48	46	94

Table 1: Overview of treatments

2.1 The Continuous Trust Game (CTG) and the shift instrument

Under the rules of the CTG, each of the two players initially receives an "account" that contains 100 points. The trustor, here labelled "participant X", chooses the share a_1 of her points that are to be transferred to the trustee, "participant Y". The transfer is productive — every point that the the trustor sends is tripled on the way to the trustee. Simultaneously, i.e. without knowing the trustor's transfer, the trustee decides how much to transfer back from the total that she has in her account after X's transfer. Both transfers are chosen simultaneously, so the trustee, like the

Croson, 2000, and Rutstrom and Wilcox, 2009) — it is usually not found to be intrusive in the sense of affecting the players' actions in the games (see e.g. Blanco et al, 2008, Costa-Gomes and Weizsäcker, 2008). In our experiment, the danger of such an intrusion appears minimal, given that we elicit beliefs *after* the choices. Either the participant is done with his or her decisions at this point of the experiment, or he or she proceeds to be a trustee in the second match – but the trustees' choices are not the focus of our study. Moreover, any possible effect on trustees' choices cannot even enter the trustors' expectations: the trustors in the second match were trustees in the first match and were not made aware of the belief elicitation task that their opponents faced in the first match.

trustor, makes a decision about a relative "transfer share" a_2, not an absolute amount.

The transfer shares (a_1, a_2) are restricted to lie in the interval $[0.2, 0.8]$. Thus, effectively, the trustor can transfer between 20 and 80 points, which are tripled and added to the trustee's amount, resulting in an account balance for the trustee between 160 and 340 points. Of these points, the trustee can transfer back a share of between 0.2 and 0.8 but has to do so without knowledge of the exact account balance that she has available. This simultaneous version of the trust game has the advantage that the trustor's beliefs about the trustee's transfer share are a comparatively simple object — a distribution over $[0.2, 0.8]$. (In the sequential version the trustor's beliefs would specify such a distribution for each possible action of his own.)

The instrumental variable is a shift z that increases or decreases the trustee's transfer share by a value between -0.2 and 0.2, which is randomly drawn from a uniform probability distribution over the 41 values on the grid $\{-0.2, -0.19, ..., 0, ..., 0.19, 0.2\}$. The participants are both informed that the trustee's transfer share a_2 will be randomly changed by adding the random variable z to it. The trustor is, in addition, informed about the value of z, while the trustee is not. For example, suppose that upon being informed that the realization of the shift z is 0.15, the trustor transfers a share $a_1 = 0.4$ of her intitial balance of 100 points, thus leading to intermediate account balances of 60 and 220 points for the trustor and trustee, respectively. In addition, suppose that the trustee decides to transfer $a_2 = 0.3$, which leads to an actual transfer to the trustor of $0.45 = 0.3 + 0.15$ of the trustee's intermediate balance, leading to final balances of 159 and 121 points for the trustor and trustee, respectively.

The experimental participants receive instructions about the instrument that begin with the following sentences.

> "There is one important detail about the transfer out of Participant Y's account. The computer adjusts the share that is actually transferred from Participant Y's account to Participant X's account. More specifically, the computer will adjust Y's transfer share in a random way, increasing or reducing it by up to 20 percentage points. That is, the computer will generate a number that we call "CHANGE TO Y's TRANSFER SHARE" by picking a random percentage number among -20%, -19%, ..., 0%, ...,

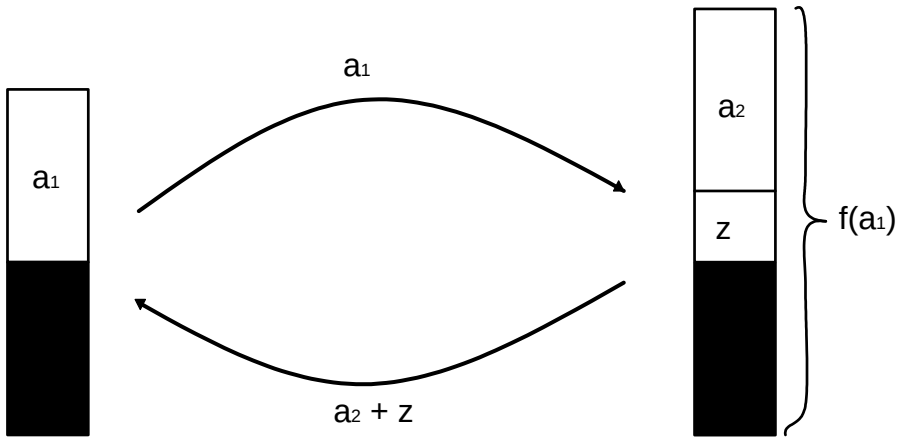

Figure 1: Illustration of the continuous trust game with instrument. Player 2 knows only the distribution of z and chooses action a_2. Player 1 knows the distribution of z and the value of z before choosing action a_1 and belief statement b_1. $f(a_1)$ indicates that player 2's account balance depends on a_1.

+19%, +20%. Each of the whole-numbered percentages in this range is equally likely."

The instructions continue by giving a repetitive illustration of the instrumental variable and its effects on payoffs. The game's rules, including the role of the instrument, are illustrated in Figure 1.

After making their choices, the trustor is asked to report her belief statement about the trustee's "adjusted transfer share", i.e. about the sum $\widetilde{a}_2 = a_2 + z$. The belief statement is rewarded according to the quadratic scoring rule

$$\pi_b = A^{CTG} - c^{CTG}(\widetilde{a}_2 - b_1)^2,$$

where b_1 is Participant X's belief statement about Participant Y's transfer share, and the parameter values are $A^{CTG} = c^{CTG} = 250$ points. This elicitation procedure applies both when the game is played with and without the instrument — when played without the instrument, the trustor is simply asked about the trustee's transfer a_2. At the time when participants choose the actions in the game, none of them is made aware of the subsequent belief elicitation task.

Importantly, the shift instrument z is generated independently of all other relevant random variables. This property justifies the conditional exogeneity assumptions that are required for IV: in the bi-variate linear projection of the trustor's transfer share a_1 on her stated beliefs b_1,

$$a_1 = \beta_0 + \beta_1 b_1 + u, \tag{1}$$

the exclusion restriction requires that while the error term u can, in general, be confounded with b_1 due to omitted variables, the instrumental variable needs to be orthogonal to u. (If other controls are included, the analogous statement with the corresponding error term of the regression on beliefs and controls is required.)[5] Since z is independently generated in the laboratory, we can rule out that u has an influence on z , or that any omitted variable may co-determine u and z. It remains an assertion that z does not influence u. We regard this as a reasonable assertion because z is simply a component of $\widetilde{a}_2$, which is the statistic that beliefs b_1 are formed about, and z does not enter the interaction in any other way.[6]

This discussion also relates to the interpretation of the explanatory variable b_1. For the IV estimate of β_1 to be consistent, we interpret the connection of b_1 and a_1 in (1) as a causal link between stated beliefs b_1 and actions a_1, and assume that the exogenous information z does not enter a_1 other than through b_1. The stated beliefs are thus interpreted as containing all choice-relevant assessments of the re-payment rate on behalf of the agent. This requires a rather consistent pair of responses from the experimental participants in the two tasks — e.g. we rule out that they report a random belief statement but are influenced by z in their actions. But we note that the monetary incentives to report optimal belief statements is designed to minimize such random belief statments. Section 3 will contain results that demonstrate that belief statements are indeed strongly responsive to z (in fact, they respond in a way that is consistent with the hypothesis that participants simply add z to their beliefs about a_2), lending support to the validity of the instrument.

[5] See e.g. Angrist and Pischke (2009) for a useful wider discussion of exclusion restrictions.

[6] Of course, the functional form assumption is never innocuous but this is a general property of regression analyses. We also note that if one is willing to maintain linearity assumptions, the exogeneity of z and u should hold true for much wider classes of preferences than then money-maximizing-agent model. All preferences over distributions of players' earnings (e.g. inequity-averse preferences) that induce a linear model are compatible with the orthogonality of z and u.

2.2 The Mini Trust Game (MTG) and the robot instrument

Under the rules of the MTG, the participants play the following matrix game.

		Column	
		$left$	$right$
Row	top	$120, 120$	$120, 120$
	$bottom$	$270, 210$	$30, 360$

Row acts as the trustor in this game, having the choice between the safe action *top* and the uncertain action *bottom*, whereas Column simultaneously chooses between $left$ and $right$. The trusting (also the cooperative) outcome is the action pair $(bottom, left)$ — involving less inequality than the outcomes following *top*, but maximizing the sum of earnings by a considerable margin. Both players' cooperative actions are socially desirable in that they increase the pie size.

After the game is played, Row's belief about Column's probability of choosing $left$ is elicited, again with a quadratic scoring rule: every participant in the role of Row is asked to report a probability $b_l \in [0, 1]$ of the event that her Column opponent chooses $left$. From this task she earns an additional payoff of

$$\pi_b = A^{MTG} - c^{MTG}(I_l - b_l)^2,$$

where I_l is an indicator function that takes on the value 1 if the participant's opponent actually chooses $left$, and the payoff parameters are $A^{MTG} = c^{MTG} = 300$ points.

The instrumental variable is introduced to the participants as detailed in the following passage from the experimental instructions.[7]

> "There is a twist in this decision scenario. The decision that the COLUMN participant makes is not always COLUMN's FINAL DECISION between $left$ and $right$. In one out of two cases, that is, with a probability of 50%, COLUMN's DECISION is made by a ROBOT. If this happens, then the robot that makes the choice will be selected randomly by the computer from a pool of 10 pre-programmed robots."

[7] The quote is changed from the original instructions in the labelling of actions; we used the abstract symbols # and & instead of the words $left$ and $right$.

The instructions continue by giving more detail on this random procedure. They explain that there are ten pre-programmed robots, each of which chooses either $left$ or $right$. The precise distribution of the robots' actions is unknown to the trustee who only knows that this distribution is itself a random variable: all 11 possible distributions (all 10 choose $left$), (9 choose $left$, 1 chooses $right$), ..., (all 10 choose $right$) are equally likely. Only the trustor is informed about the actual distribution of robots' actions. We denote this distribution of the robots' actions by z_l, a random variable that is, just like the instrument in the CTG, independent of all potential relevant background variables in our statistical analysis. At the same time, the realization of z_l may of course be relevant for Row's choice, to the extent that her choice is driven by the expectation about Column's final decision.[8]

A few comments about this instrumental variable are in order, as one may notice two potential pitfalls. First, the instructions for this instrument are relatively complicated and involve a two-step random procedure: Nature determines the robots' pool composition (drawn uniformly among eleven possible values) and also whether the robots are relevant for the final choice (with 50% chance). This may lead to confusion among the participants. The problem can partially be overcome by instructing the participants with much care, but the complexity could limit the possible use of this instrument and create disturbances in the data. Section 3 will show that certain data patterns show suspicious deviations from theoretical predictions. Second, the robot instrument potentially distorts the trustees' choices because they can realize that the robots will on average choose each action with probability one half. This affects the players' payoffs in a specific direction and the participants may therefore change their choice probabilities in response (for example, they may want to counteract the behavior of the robots). For these reasons, the bulk of the data analysis in the next section will focus on the CTG, where the above pitfalls do not apply.

[8] Theoretically, a Row participant who believes that a proportion b_l of Column participants choose $left$ and who receives the information that a proportion z_l of the robots choose left would update her assessment of Pr(opponent chooses $left$) by adding the difference $\frac{1}{2}(b_l + z_l) - b_l = \frac{1}{2}(z_l - b_l)$ to her initial belief b_l of Column. Thus for each additional $left$ playing robot that is added to the distribution of robots, the belief increase by five percentage points.

3 Results: Causal effects of beliefs

3.1 Preliminaries: Data pooling, descriptives, checks for invasiveness and first-stage regressions

Data Pooling. We first determine if there are any statistically significant differences between the data collected at UCL and at York, and between data that were generated in the participants' first versus second matchings (recall that every participant played two versions of the trust game). We do so for each of the games and consider both actions and stated beliefs. The absence of major differences allows us to pool the data and simplify the subsequent analysis.

We start with the CTG data and compare the distributions of transfer shares of each of the players and the distributions of the trustors' stated beliefs. Initially, we pair the two treatments in which the CTG game was played under the same instrument condition at each of the locations, thereby testing for order effects. The absence of such order effects leads us to pool the data and test for laboratory effects, by comparing the data collected at the two different locations. We apply Kolmogorov-Smirnov's two-sample exact test to the transfer shares and belief statements and find no statistically significant order or laboratory effects, for any of the player roles or for any instrument condition.[9]

We follow the same steps for the MTG data. Here, we apply Fisher's exact test to compare the distributions of each of the players' binary choices and again Kolmogorov-Smirnov's two-sample exact test for the distributions of the trustors' stated beliefs. Once again, initially we test for order effects by pairing the two treatments in which the MTG game was played with the same instrument condition at each of the locations. No order effects appear, and we pool the data of the participants' two matchings and test for laboratory effects. We again find no statistically significant order or laboratory effects, neither in the actions data nor in the beliefs data.[10] Therefore, in the subsequent

[9] The twelve tests on the order and laboratory effects all produced p-values above 0.1, with the exception of the test of order effects on the trustee's transfer share in the instrument condition at for the sessions run at York, which had a p-value of 0.003. Since our data analysis focusses entirely on the trustors, the rejection of the null nypothesis for trustees at York is not problematic. Appendix C has more detail on this and other features of the data.

[10] The lowest p-value is 0.24. In addition, we use Exact Chi-square tests to test the hypothesis that the pooled data was generated randomly. The hypothesis is rejected for both player roles in both instrument conditions with

data analysis we use the pooled data played under each instrument condition.

Descriptives. Before turning to the question of whether the instrument was distortive in undesirable ways, we first decscribe the marginal distributions of the data, for each of the four games. We focus on the trustor's data, as the trustee's role in this study is only to generate an uncertain re-payment.

In the continuous-choice game without instrument, CTG/NI, the transfer shares of the trustor follow a familiar tri-modal pattern that has been observed in many other trust game experiments, with substantial proportions of participants transfering the lowest possible amount (here, a transfer share of 0.2, chosen by 32.6%), or the midpoint of the action space (0.5 transfer share, chosen by 19.0%) or the highest possible transfer share (0.8, chosen by 14.7%). The remaining data are dispersed between these three modes. The average transfer share in CTG/NI is 0.427 (std. dev. 0.218). The belief statements of the same participants have an average of 0.350 (std. dev. 0.132). This is not too far from the target of the belief statement excercise, i.e. the trustee's actual mean transfer share, which is 0.306 (std. dev. 0.144).

The corresponding numbers for CTG/I with instrument are of comparable size, with the exception that the presence of the instrument induces additional variance in the belief statements — as it should because the shift is random. The frequencies of transfer share that lie on the points of the simplest three-point grid $\{0.2, 0.5, 0.8\}$ are 30.3%, 7.4% and 18.9%. The average transfer share is 0.435 (std. dev. 0.226). The belief statements in treatment CTG/I have an average of 0.330 (std. dev. 0.185),[11] which again makes for a fairly accurate mean prediction as the mean transfer share of trustees lies at 0.303 (std. dev. 0.150).

---

the exception of the Row player in the instrument condition, as the split between *top* and *bottom* is fairly uniform in these data.

[11] The larger variance of belief statements under CTG/I relative to CTG/NI is consistent with the fact that z adds variance to the target of the belief statement. Under the assumption that the participants in CTG/I arrive at their belief statements by simply adding the shift variable to their belief about a_2, the two variables "belief about a_2" and z are independent and one can thus simulate the predicted variance in CTG/I under the null hypothesis that beliefs about a_2 are constant between the treatments — the variance of the stated belief in CRG/I is the sum of the variance of z and the variance of the beliefs about a_2 in CTG/NI. This is almost precisely confirmed in the data: empirically, the sum of the variance of z and the variance of the stated beliefs in CTG/NI is 0.03292, the square root of which is 0.181, very close to the standard deviation of stated beliefs in CTG/I (0.185).

In the discrete-choice game without instrument, MTG/NI, a proportion of 0.262 chooses the trusting action *bottom*. The average belief statement about the trustee's probability of $left$ is 0.359 (std. dev. 0.320). This, too, is not far from the true value of the target: 31.2% of trustee's actually choose $left$. In the MTG with instrument, MTG/I, the trustor's average belief statements is significantly more optimistic, at 0.449 (std. dev. 0.293) but this change in belief is largely justified by the robots' bias towards playing $left$ with fifty-fifty. Empirically, the average behavior of trustees in MTG/I is similar to MTG/NI, in that 26.3% of trustees choose $left$, but adding the effect of the robots yields an average of "column's final decision" of 0.396.

Checks for invasiveness. We now examine whether the presence of an instrument has undesired effects on how subjects play the games. There are two types of undesired invasiveness that one would like to rule out. First, the mere introduction of the instrument should not affect the behavioral variables, except through the channel of influencing the beliefs. In particular, in the CTG, where the shift's expectation is 0, it is natural to require that none of the averages of the behavioral variables exhibit a significant variation between the treatments with and without the shift. The NI/I comparisons made in the two previous paragraphs all pass this test, as the corresponding hypothesis test for differences in means never registers a rejection at a significance level of 5%. There is also no significant difference in the transfer shares' variances.[12]

For the NI/I comparison in the MTG, matters are more complicated because the introduction of the robots affect "column's final decision" on average, and hence should also affect the average belief. What we would like is that the underlying beliefs about the opponent's probability of playing $left$ (before the effect of the instrument) are equally distributed under both instrument conditions. Although we cannot test this hypothesis directly (since we do not have access to the true underlying beliefs) we can reconstruct what the underlying beliefs of a rational decision maker with the specific value of z_l would have to be, given the observed belief statement.

An important check is whether these hypothetical, underlying beliefs are "admissible" in the

[12] For the NI/I comparison of action data in CTG, we conducted for each player role a Mann-Whitney test as well as a variance ratio test. For the comparison of belief statements in CTG, we conducted a Mann-Whitney test. None of the tests rejected at any conventional level. For the comparison of variances in stated beliefs between CGT/NI and CTG/I, see footnote 11.

sense that they are in the unit interval. Other belief statements would indicate a potential confusion on behalf of the participant. For example, with an instrument value of $z_l = 0.9$ (i.e. nine out of ten robots would choose $left$), it is "inadmissible" to state the belief 0.3, as the underlying belief about the opponent's probability of choosing $left$ would have to be a negative number. A similar calculation can be made for treatment CTG/I, where it is required that $b_1 - z$ lies within the interval $[0.2, 0.8]$, i.e. that the stated belief could have been derived from an "admissible" underlying belief, by simply adding the shift value. The results of these calculations indicate that under treatment MTG/I, a substantial number of subjects may have been confused: 24% state "inadmissible" beliefs, given their instrument values. In contrast, under CTG/I, only 4% state "inadmissible" beliefs, leading us to conjecture that the instructions of CTG/I are easier to understand than those of MTG/I. We therefore put the emphasis of our subsequent analysis on the CTG data. Also, from here onwards, we exclude from the analysis all observations where a participant stated an "inadmissible" belief.

First-stage regressions. A final preliminary check is whether the belief statements react to the instrument, as is necessary for IV analyses. The answer is a clear yes, for both games. In the CTG/I game, the slope coefficient in a binary regression is estimated at 0.828 (robust std. dev. 0.131). The coeffient is insignificantly different from 1, the theoretical value under the assumption that the participants simply add the shift to their belief about the trustee's action. Further support for such a rational way of belief formation can be found by inspection of the location of (b_1, z)-pairs, in Figure 2. In the figure, the dashed line represents the target, which is given by the mean behavior of trustees plus the trustor's specific value of z. This line's slope is equal to one, and points on the line correspond to the set of optimal belief statements, conditional on the subjects' information. The solid line is the regression line generated from the depicted data. As can be seen, a prominent feature of the data is that many (b_1, z)-pairs lie on straight lines. This is consistent with the hypothesis that the participants add their value of z to a belief that lies on the grid $\{0.2, 0.3, ...\}$. Altogether, 64% of the observations have this property, suggesting that the majority of participants may have simply added z to their belief about a_2.

A very similar picture emerges for the MTG/I game, at least for the 76% of data where the

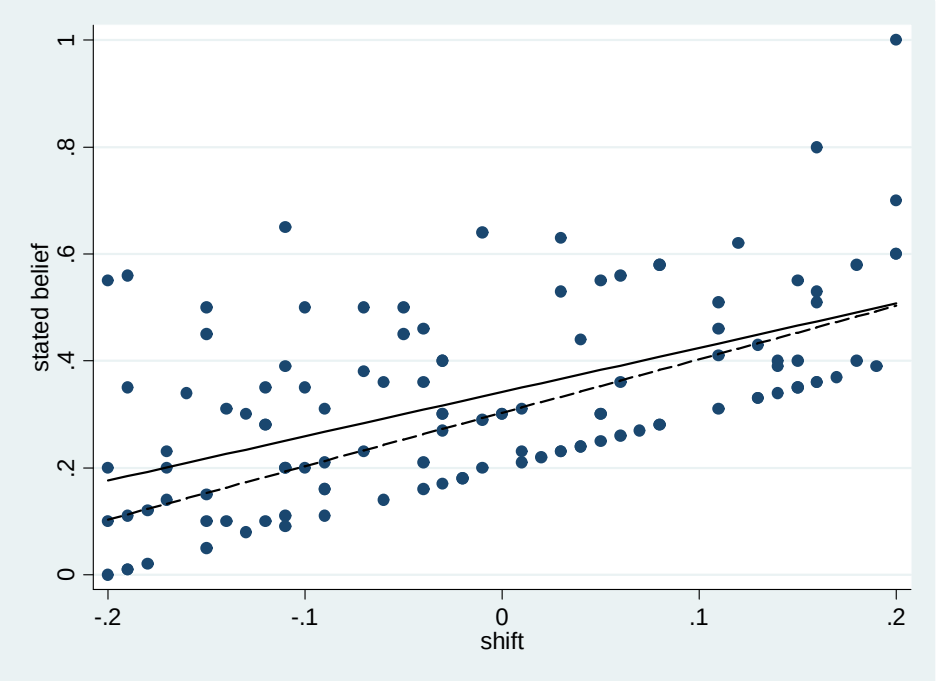

Figure 2: Stated beliefs versus shift in treatment CTG/I. Dashed line: average target + shifter. Solid line: OLS predicted values, coefficient 0.828 ($p < 0.001$, $R^2 = 0.31$, # of obs.= 117)

belief statements are "admissible" given the instrument values. Under the assumption that the participants report, as would be optimal, the average of the probability of $left$ for the robot and their estimate of the human opponent's probability of $left$, we would observe a regression coefficient of 0.5 (as indicated by the dashed line, representing the target after adding the robot proportion). The empirical coefficient is 0.570 (robust std. dev. 0.055), which is strongly different from 0 and insignificantly different from 0.5. Also, Figure 3 shows the analogous pattern as Figure 2, that the large majority of points (93%) lie on a grid of straight lines. Here, however, this is less surprising because the instrument is also distributed on a grid. In sum, the data analyses in this subsection show that — with the exception of the 24% of MTG participants who state "inadmissible" beliefs — introducing the instruments has no undesirable side effects on the data distributions and generates strong instrumental variables with the predicted effects on belief statements.

3.2 Regression analysis of the CTG data

Table 2 shows the results of the OLS analysis carried out on the CTG/NI data. (Remember, this is the analysis that one would carry out in order to establish causality, in the absence of omitted-variable problems.) The OLS estimates show a strong correlation of stated beliefs and the trustors' transfer shares. In the regression that does not include personal background characteristics, an increase in the belief by 10 percentage points translates into an increase by 7.6 percentage points in the transfer share. In the regression with controls, the coefficient is even larger, corresponding

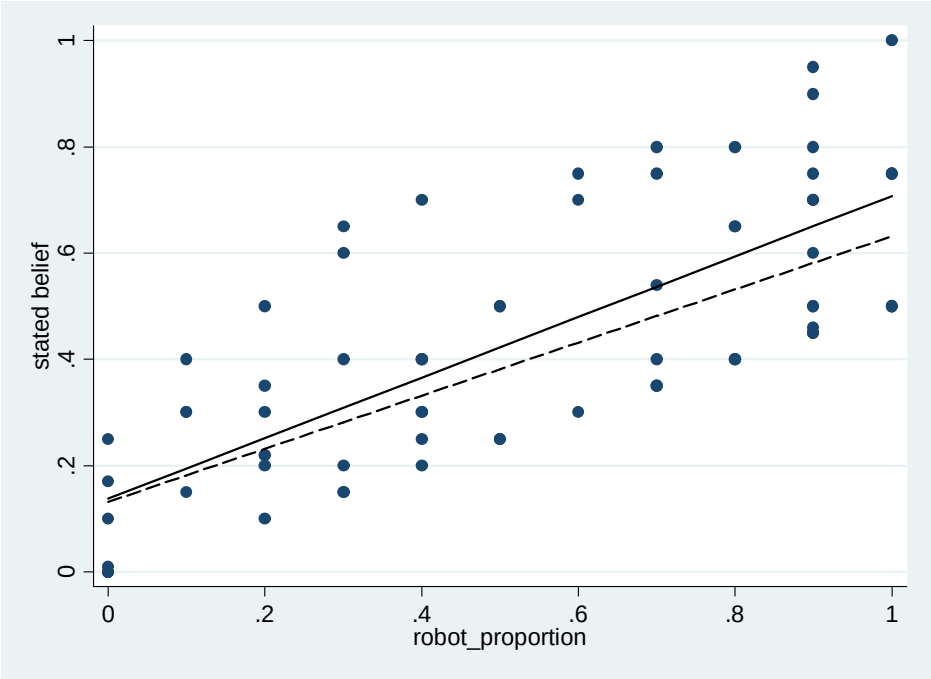

Figure 3: Stated beliefs versus robot proportion in treatment MTG/I. Dashed line: $\frac{1}{2}$∗average target $+\frac{1}{2}$ robot proportion. Solid line: OLS predicted values, coefficient 0.570 ($p < 0.001$, $R^2 = 0.58$, # of obs.= 72)

to a 9.5 percentage points increase in the transfer share.

	Transfer share, Treatment CTG/NI	
	(1)	(2)
	OLS	OLS
Belief statement	0.764 (0.186)	0.949 (0.207)
Constant	0.159 (0.065)	0.495 (0.292)
Personal controls	no	yes
# of Obs.	95	91
R^2	0.214	0.431

Table 3: Transfer shares in treatment CTG/NI. Note: Robust standard errors in parentheses.

A "naive" attribution of these statistical connections to a causal effect would thus suggest that beliefs are a strong driver of trust. The paper's main question is whether this attribution can be corroborated by the IV results. Table 4 has the IV results, showing that the answer is affirmative:

	Transfer share, Treatment CTG/I			
	(1)	(2)	(3)	(4)
	OLS	OLS	IV	IV
Belief statement	0.492 (0.099)	0.537 (0.133)	0.559 (0.195)	0.544 (0.268)
Constant	0.277 (0.036)	0.431 (.238)	0.255 (0.065)	0.428 (0.262)
Personal controls	no	yes	no	yes
# of Obs.	117	113	117	113
R^2	0.149	0.279	0.146	0.279

Table 4: Transfer shares in treatment CTG/I. Note: Robust standard errors in parentheses.

As indicated in column (3) of Table 4, the IV coefficient without control variables is estimated at 0.559. This is insignificantly smaller than the OLS coefficient from treatment CTG/NI.[13] The

[13]It is also insignificantly larger than the OLS coefficient from the data with instrument, but this in itself is not an

important observation about the IV results is that the coefficient is significantly different from zero. To our knowledge this is the first evidence in favor of the hypothesis that the correlation between beliefs and actions in an experimental game is indeed causal.

The results also show that within treatment CTG/I, there is no discernible difference in the results of OLS versus IV.[14] This is another indication that there cannot be a strong omitted-variable problem. However, one may worry about the observation that the OLS coefficients differ between CTG/NI and CTG/I. The difference is insignificant, however, in a regression that includes all main and interaction effects of treatment (NI/I) and belief statement.[15] A further indication that omitted variables play no large role is that the regressions with personal control variables yield essentially unchanged results. The beliefs statement coefficients in the odd-numbered columns of the tables do not appear to simply reflect the influence of the available background characteristics.

As a separate regression analysis, we also report the direct effect of the instrument on the trustor's transfer share. The results confirm directly that the exogenous variation has a significant effect on the trustor's transfer share. The impact of an increase in the shift is comparable to the change associated with a corresponding increase in the stated belief in treatment CTG/NI (0.46 versus 0.56, in regressions without controls). However, the coefficient here does not indicate a measure for the size of the effect of beliefs on actions, merely the size of the effect of the artificial

interesting observation because the belief statements in the treatment with instrument have been affected, as they should, by the introduction of the instrument. (Cf. footnote 3.)

[14] Hausmann tests do not reject OLS assumptions at any conventional level.

[15] The significance level of a difference in slope coefficients between the two treatments is $p = 0.199$. To the extent that there is a difference between the two treatments, this could be generated by reciprocity: under treatment CTG/NI, trustors may want to be kind to their opponents if they expect them to be kind as well. In treatment CTG/I, part of the belief is driven by the computer draw, so a reciprocal agent may respond less to this belief.

instrument on actions.

	Transfer share, Treatment CTG/I	
	(1)	(2)
	OLS	OLS
Shift	0.462 (0.174)	0.397 (0.204)
Constant	0.446 (0.021)	0.646 (0.250)
Personal controls	no	yes
# of Obs.	117	113
R^2	0.059	0.187

Table 5: Transfer shares in treatment CTG/I. Note: Robust standard errors in parentheses.

3.3 Regression analysis of the MTG data

Table 6 reports the essential regressions for the MTG data. The regressions are conducted without control variables, and only on observations with "admissible" belief statements.

	$Pr(bottom)$		
	(1)	(2)	(3)
Treatment	MTG/NI	MTG/I	MTG/I
	OLS	OLS	IV
Belief statement	0.383 (0.121)	-	1.123 (0.296)
$\frac{1}{2}$*(robot proportion)	-	1.280 (0.314)	-
Constant	0.125 (0.058)	0.192 (.101)	0.038 (0.145)
Controls	no	no	no
# of Obs.	122	72	72
R^2	0.077	0.193	0.072

Table 6: Trustor's probability of *bottom* in treatments MTG/NI and MTG/I

Note: Robust standard errors in parentheses. Variable "$\frac{1}{2} *$ (robot proportion)" is scaled for comparability with belief statement.

The regressions confirm a significant causal role for beliefs in the determination of actions. The variable "$\frac{1}{2}$*(robot proportion)" is scaled to cover the interval $[0, \frac{1}{2}]$, in order to increase the comparability of coefficients across the table's column's: given the construction of the instrument, an increase of of "$\frac{1}{2}$*(robot proportion)" by one unit would theoretically translate into an increase of the belief by one unit (cf. footnote 8).

We see from the regressions that the regressor coefficients in the data from MTG/I are much larger than for MTG/NI. This may be due to the select sample of respondents who make "admissible" belief statements, or to the fact that the robot intervention is not neutral on the actions, as described in the previous subsection as well as in Appendix C. These competing explanations cannot easily be separated, and we thus regard the evidence from the MTG as less definitive than that from the GTG. But the fact that both the IV coefficient of the belief statement and the coefficient of "$\frac{1}{2}$*(robot proportion)" are significantly positive is nevertheless an independent confirmation of

the causal role of beliefs.

4 Conclusion

Our intended contribution to the experimental literature is to introduce specially designed instrumental variables into the laboratory. In previous experiments, researchers have of course used their control of the design to manipulate directly the explanatory variables of interest. This allows causal insights and is of the main reason why experiments are so popular. But in some contexts, the explanatory variable of interest is by its very nature an endogenous variable, and thus cannot be fully controlled by the experimenter. In particular, the agents' expectations about other agents have this property. In such contexts, we point out that one can at least influence the explanatory variable of interest *to some degree*, by way of using instrumental variables. Under standard linearity assumptions, this suffices to measure causal links. Similar procedures may be applied in studies where the explanatory variable of interest is of a different nature, but is likewise endogenous to the process that determines the actions. Variables in this set may be responses to attitudinal questions, happiness reports or even neurological data.

An unusual feature of our study is that we are explicitly asking about causal link of expectations on actions — yet traditionally expectations are, at least under subjective expected utility, not viewed as a concept that is separate from actions. We acknowledge that we do not offer an alternative definition of expectations, but simply take the belief statements as our data. We hope that future research will allow us to give a more satisfying conceptual treatment of belief statements.

References

Angrist, J., and S. Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

Attanasio, O. (2009), "Expectations and perceptions in developing countries: Their measurement and their use", American Economic Review (Papers and Proceedings) 99, 87-92.

Bellemare, C., and S. Kröger (2007), "On representatitve social capital", European Economic Review 51, 181-202.

Berg, J., J. Dickhaut and K. McCabe (1995), "Trust, reciprocity, and social history", Games and Economic Behavior 10, 122-142.

Blanco, M., D. Engelmann, A.K. Koch and H. Normann (2008), "Belief elicitation in experiments: Is there a hedging problem?", IZA Discussion Paper 3517.

Costa-Gomes, M., and V. Crawford (2006), "Cognition and behavior in two-person guessing games: An experimental study", Americcan Economic Review 96, 1737-1768.

Costa-Gomes, M., and G. Weizsäcker (2008), "Stated beliefs and play in normal form games", Review of Economic Studies 75, 729-762.

Cox, J. (2004), "How to identify trust and reciprocity," Games and Economic Behavior 46, 260-281.

Croson, R. (2000), "Thinking like a game theorist: Factors affecting the frequency of equilibrium play," Journal of Economic Behavior and Organization 41, 299-314.

Dupas, P. (2009), "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya", mimeo, UCLA.

Fehr, E., U. Fischbacher, B.v. Rosenbladt, J. Schupp and G.G. Wagner (2003), "A Nation-Wide Laboratory Examining trust and trustworthiness by integrating behavioral experiments into representative surveys", IEW Working Paper 141.

Huck, S., and G. Weizsäcker (2002), "Do players correctly estimate what others do? Evidence of conservatism in beliefs", Journal of Economic Behavior and Organization 47, 71-85.

Jensen, R. (2008), "The perceived returns to education and the demand for schooling", mimeo, Brown University.

Manski, C.F. (2004), "Measuring Expectations", Econometrica 72, 1329-1376.

McKelvey, R.D., and T. Page (1990), "Public and private information: An experimental study of information pooling", Econometrica 58, 1321-1339.

Naef, M., and J. Schupp (2009), "Measuring trust: Experiments and surveys in contrast and combination", mimeo, Royal Holloway.

Nyarko, Y., and A. Schotter (2002), "An experimental study of belief learning using real beliefs," Econometrica 70, 971-1005.

Offerman, T., J. Sonnemans, and A. Schram (1996), "Value orientations, expectations and voluntary contributions in public goods," Economic Journal 106, 817-845.

Rutström, E.E., and N.T. Wilcox (2009), "Stated beliefs versus inferred beliefs: A methodological inquiry and experiemtal test", Games and Economic Behavior 67, 616-632.

Sapienza, P., A. Toldra and L. Zingales (2007), "Understanding Trust", mimeo, University of Chicago.

5 Appendix A: An example of naive inference under omitted variables and equilibrium play

In this section we give an example of how the correlation between belief statments and actions can be misleading in the presence of ommitted variables. To arrive at a "misleading" effect, we imagine that a researcher observes the full data (choices and belief statements about the opponent's choices) but ignores the possibility of a social norm, or any other unobserved variable, that could drive behavior and belief statements. The players, in contrast, are aware of the full model and play the unique Bayes-Nash Equilibrium (BNE) of the game.

The example builds on a 2x2 mini trust game, where player 1 can either trust ($a_1 = 1$) or not ($a_1 = 0$) and player 2 can reciprocate ($a_2 = 1$) or not ($a_2 = 0$). The players are aware of a social norm that prescribes trust and reciprocation ($a_1 = a_2 = 1$). A random event specifies whether violations of the social norm are sanctioned: in state $\omega = 1$, violations are sanctioned, and we assume that this state arises with probability $\frac{1}{2}$. If $\omega = 1$ occurs, player i is penalized by a term γ_i if she does not comply with the norm. The punishment parameter γ_i is known to the player herself but not to her opponent, who only knows the distribution of γ_i to be uniform over $[0, 1]$. If $\omega = 0$, no punishment applies.

A possible justification for such a probabilistic social norm enforcement is that with probability $\frac{1}{2}$ the interaction does not remain anonymous. For example, an outside observer (say, the experimenter) may identify each player's action and impose a punishment γ_i on non-cooperative play. Or, the players meet afterwards and may be compelled to reveal their play in the game. In this case, the punishment parameter γ_i would reflect the extent of embarrassment. The payoffs (π_1, π_2) in the two states are as follows.

$\omega = 0$		Player 2	
		$a_2 = 0$	$a_2 = 1$
Player 1	$a_1 = 0$	$0, 0$	$0, 0$
	$a_1 = 1$	$-1, 2$	$1, 1$

$\omega = 1$		Player 2	
		$a_2 = 0$	$a_2 = 1$
Player 1	$a_1 = 0$	$-\gamma_1, -\gamma_2$	$-\gamma_1, 0$
	$a_1 = 1$	$-1, 2 - \gamma_2$	$1, 1$

We assume that the two punishment terms γ_1 and γ_2 are *i.i.d.* uniformly distributed on the interval $[0, 1]$. The worst feasible punishment, $\gamma_i = 1$, makes the non-cooperative action $a_i = 0$

weakly dominated for player i, under state $\omega = 1$. The smallest possible punishment for player 2, $\gamma_2 = 0$, makes player 2's non-cooperative action be weakly dominant (independent of ω). Player 1's optimal action depends on ω, too, but as usual in the trust game it also depends on her belief about a_2 — for a large expected return, it pays off to trust.

While players do not know the true state ω for sure, they each receive a signal s_i that has precision $\frac{2}{3}$. That is, $\Pr(s_i = 1|\omega = 1) = \Pr(s_i = 0|\omega = 0) = \frac{2}{3}$, for $i = 1, 2$. Their information about ω is therefore correlated: players know that it is more likely than not that the opponent receives the same signal. The probability of the opponent having the same signal is $\frac{5}{9}$ (and the correlation coefficient between the two players' signals is $\frac{1}{9}$).

In this Bayesian game, a player's type is given by her signal s_i and her punishment payoff γ_i. We assume for simplicity that the punishments (γ_1, γ_2) are independent of the signals (s_1, s_2). It is then straightforward to determine the players' optimal choice probabilities: for any signal s_i and any belief about the opponent's strategy, we first ask what values of γ_i make it optimal for the player to choose the cooperative action $a_i = 1$. The answer yields a cutoff value $\widehat{\gamma}_i(s_i)$, such that for $\gamma_i \geq \widehat{\gamma}_i(s_i)$, the player chooses $a_i = 1$. Each player i employs two such cutoffs, one for each signal realization, $s_i \in \{0, 1\}$. Player i also entertains a belief about the opponent's cooperations: $\Pr(a_j = 1|s_i) = \sum_{\widetilde{s_j} \in \{0,1\}} \Pr(s_j = \widetilde{s_j}|s_i)(1 - \Pr(\gamma_j < \widehat{\gamma}_j(\widetilde{s_j})))$. This belief determines player i's two cutoffs, and the BNE solution is then found by solving for a set of four cutoffs that form a fixed point. In particular, denote the choice probabilities under the players' equilibrium strategies by $r = \Pr(a_1 = 1|s_1 = 0) = 1 - \widehat{\gamma}_1(s_1 = 0)$, $s = \Pr(a_1 = 1|s_1 = 1) = 1 - \widehat{\gamma}_1(s_1 = 1)$, $t = \Pr(a_2 = 1|s_2 = 0) = 1 - \widehat{\gamma}_2(s_2 = 0)$, and $u = \Pr(a_2 = 1|s_2 = 1) = 1 - \widehat{\gamma}_2(s_2 = 1)$. To find e.g. the cutoff value $\widehat{\gamma}_1(s_1 = 1)$ that makes player 1 indifferent upon signal $s_1 = 1$, we solve

$$\begin{aligned} E[\pi_1(a_1 &= 0|s_1 = 1, \widehat{\gamma}_1(s_1 = 1))] = E[\pi_1(a_1 = 1|s_1 = 1, \widehat{\gamma}_1(s_1 = 1))] \\ \frac{2}{3}(-\widehat{\gamma}_1(s_1 &= 1)) = \Pr(a_2 = 0|s_1 = 1) \cdot (-1) + \Pr(a_2 = 0|s_1 = 1) \cdot 1 \end{aligned}$$

which can be rewritten as:

$$s = \frac{3}{2}(\frac{8}{9}t + \frac{10}{9}u) - \frac{1}{2}$$

Formulating analogous expressions for r, t and u allows to solve for the unique equilibrium values

$\{r = 0, s = \frac{3}{5}, t = \frac{1}{5}, u = \frac{1}{2}\}$. We see that in equilibrium, both players react strongly to their signals.[16]

Now consider a naive researcher who wants to infer the causal effect of player 1's beliefs on her actions. We define a naive researcher as one who is not aware aware that the information structure determines the players' beliefs and actions. Rather, to allow for the possible observation of heterogeneity in belief statments, the researcher views the players' beliefs as exogenous and does not require that they are in equilibrium. The researcher collects player-1 data on actions and belief statements about player-2 actions, which we assume are reported truthfully, generated by the full model with social norms. The researcher will therefore observe two different belief statements: first, when player 1 receives the signal $s_1 = 1$, she reports the belief that her opponent cooperates with probability

$$\Pr(a_2 = 1|s_1 = 1) = \frac{5}{9}u + \frac{4}{9}t = \frac{11}{30}.$$

Under this signal realization $s_1 = 1$, we saw above that her actions are cooperative with probability $\frac{3}{5}$. Second, when player 1 receives the signal $s_1 = 0$ she reports that her opponent cooperates with probability

$$\Pr(a_2 = 1|s_1 = 0) = \frac{4}{9}u + \frac{5}{9}t = \frac{1}{3},$$

and her actions under this signal realization are cooperative with probability 0. The data on player 1 that the researcher observes can therefore be summarized in the following table (where the cell entries indicate the relative frequency of the four possible belief-action pairs):

Player 1		Belief statements	
		$bs_1 = \frac{11}{30}$	$bs_1 = \frac{1}{3}$
Actions	$a_1 = 0$	$\frac{2}{10}$	$\frac{1}{2}$
	$a_1 = 1$	$\frac{3}{10}$	0

As the naive researcher ignores the existence of the social norm, he will also wrongly assign

[16] The equilibrium is in (essentially) pure strategies, as a player with a given punishment parameter γ_i and a given belief about the opponent's play has a strict best response, except for the zero-probability event that her paramter γ_i makes her indifferent.

causal effects: we assume that he attributes any change in actions exclusively to changes in beliefs. (We also assume that the researcher is not puzzled by the fact that not all actions are best responses to stated beliefs. One could write down a simple error model of what the researcher has in mind, but this would not add much beyond the verbal statement in the sentence before these parentheses.) He therefore believes that if he could intervene and influence players' beliefs, he would also influence players' actions as prescribed by the frequencies in the data matrix. In particular, let us suppose that he thinks he could convince all members of the player-1 population who hold the belief of $\frac{1}{3}$ (i.e. one half of the population) to increase their belief by $\frac{1}{30}$. These player 1s would then hold the same belief as the other half of the population. After such an intervention, the naive researcher would expect the actions to change in accordance to the difference between the columns of the above data matrix. He would thus expect the following data after the intervention:

Player 1		Belief statements	
		$bs_1 = \frac{11}{30}$	$bs_1 = \frac{1}{3}$
Actions	$a_1 = 0$	$\frac{2}{5}$	0
	$a_1 = 1$	$\frac{3}{5}$	0

But what would the actual effects be of such an intervention, given the true model? To find the answer, the researcher could use a simple announcement (related to what is called the "robot instrument" in Section 3): he could address all player 1s whose belief statement is $\frac{1}{3}$, explaining to them that in one out of 20 times, their opponent would be replaced by a robot that always cooperates.[17] In the above equilibrium, and starting from the belief $\frac{1}{3}$, a player with signal $s_1 = 0$ would indeed arrive at a belief that the opponent cooperates with probability $\frac{11}{30}$, as one can easily check:

[17] To be precise, the announcement must be made after the researcher observes the player 1's intended actions and belief statements, but before the game is played. Importantly, for this example, the researcher must not inform player 2 about this intervention, because she would otherwise change her equilibrium behavior. Here in the theoretical example such trickery may be acceptable for the sake of exposition. In our experiments, both players are told about the possibility of intervention, so that no deception is used.

$$\begin{aligned}\Pr(\text{opponent cooperates}|s_1 &= 0) = \frac{19}{20}\Pr(a_2 = 1|s_1 = 0) + \frac{1}{20} \\ &= \frac{19}{20}\frac{1}{3} + \frac{1}{20} = \frac{11}{30}\end{aligned}$$

Under the true model, what would be the effect of the announcement on player 1's cooperation rate? What the naive researcher misses is that even under the above announcement, a player 1 with signal $s_1 = 0$ would still assign a low probability to the event that a non-cooperative action would be penalized. She would therefore still find the non-cooperative action $a_1 = 0$ relatively attractive — the omitted variable thus reduces the beneficial effect of the belief shift. But importantly, if the researcher makes the above announcement, then he does not need to know the mechanism that determines beliefs and actions and would still measure the correct causal link between them.

To find the size of the effect, we consider the relevant cutoff $\widehat{\gamma}_1(s_1 = 0)$, after the announcement. The indifference condition is:

$$\begin{aligned}E[\pi_1(a_1 &= 0|s_1 = 0, \widehat{\gamma}_1(s_1 = 0))] = E[\pi_1(a_1 = 1|s_1 = 0, \widehat{\gamma}_1(s_1 = 0))] \\ \frac{1}{3}(-\widehat{\gamma}_1(s_1 &= 0)) = \frac{19}{20}(\Pr(a_2 = 1|s_1 = 0)1 + (1 - \Pr(a_2 = 1|s_1 = 0))(-1)) + \frac{1}{20}1 \\ \frac{1}{3}(-\widehat{\gamma}_1(s_1 &= 0)) = \frac{19}{20}(\frac{1}{3} - \frac{2}{3}) + \frac{1}{20}1 \\ \widehat{\gamma}_1(s_1 &= 0) = \frac{4}{5}\end{aligned}$$

Thus only a proportion of $\Pr(\gamma_1 \geq \frac{4}{5}) = \frac{1}{5}$ of the players with $s_1 = 0$ would cooperate and the new data matrix after the announcement is

Player 1		Belief statements	
		$bs_1 = \frac{11}{30}$	$bs_1 = \frac{1}{3}$
Actions	$a_1 = 0$	$\frac{3}{5}$	0
	$a_1 = 1$	$\frac{2}{5}$	0

We conclude that by looking at the frequencies instead of measuring the effect, the naive researcher would considerably overestimate the causal link between beliefs and actions. Under the true model, only one fifth of the announcement's recipients would change their actions.

6 Appendix B: Instructions of treatment CTG/I[18]

WELCOME!

PLEASE WAIT UNTIL THE EXPERIMENTER TELLS YOU TO START!

You are about to participate in an experiment in decision making. Universities and research foundations have provided the funds for this experiment.

In this experiment we will ask you to read instructions that explain the decision scenarios you will be faced with. We will also ask you to answer questions that test your understanding of what you read. Finally, you will be asked to make decisions that will allow you to earn money. Your monetary earnings will be determined by your decisions and the decisions of other participants in the experiment. All that you earn is yours to keep, and will be paid to you in private, in cash, after today's session.

It is important to us that you remain silent and do not look at other people's work. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., you will be asked to leave and you will forfeit your earnings. Thank you.

The experiment consists of two parts, part I and part II. In each part you will anonymously interact with another participant in the room. The participant with whom you will interact in part I will be different from the participant with whom you will interact in part II. These two participants will be randomly chosen by the computer. Your identity and the identities of the other participants will not be revealed during or after the experiment.

Neither you nor the other participants will learn anyone else's decisions until the entire experiment (i.e., parts I and II) is over.

In the instructions below all earnings are described in points. At the end of the experiment all points will be converted into money. Each point is worth 2.5 pence. That is, 40 points are worth £1 (equivalently, 100 points are worth £2.50).

This handout contains the instructions for part I. These are the same instructions that the

[18] The instructions of the other treatments are available upon request. We conducted at least two sessions of each treatment at each of the two locations (UCL/York). Some of the treatments at York were conducted in parallel as part of a large session. In these, the participants received different instructions, unbeknownst to them.

participant with whom you are matched in part I will receive.

PART I INSTRUCTIONS

In this part you will be interacting anonymously with another participant in this room. The decision scenario thus involves two participants called "Participant X" and "Participant Y". We will inform you whether you are "Participant X" or "Participant Y" at the end of the instructions but before the interaction begins.

At the start of part I we will create an account for each of the participants in our experiment, and deposit 100 points into each account. At the end of the experiment, all points in the accounts will be converted into money at the exchange rate mentioned earlier. By interacting with the other participant in part I's decision scenario you can change the balance in your account, as follows.

First, Participant X decides how many points s/he wants to transfer from her/his account to Participant Y's account. The points transferred from Participant X's account will be tripled by the computer when deposited into Participant Y's account (in other words, Participant Y receives three times the amount that Participant X sends).

Second, Participant Y decides how many points, out of the points that are in her/his account, s/he wants to transfer into Participant X's account. The number of points transferred from Participant Y's account will be equal to the number of points deposited into Participant X's account (in other words they will not be tripled). This concludes the interaction, and both participants will later exchange the points in their accounts for money.

Both participants will be asked to announce a transfer share (as a percentage) of points in their account that they want to transfer to the other participant's account, instead of deciding on the number of points that they want to transfer. That is, Participant X will announce the share of her initial balance of 100 points that s/he wants to transfer to Participant Y. Participant Y will also announce the share of the number of points in her/his account that s/he wants to transfer to Participant X. However, when making her/his decision, Participant Y will not know what share Participant X has transferred. Hence, Participant Y will not know the precise balance in her account (which will be equal to her/his initial balance of 100 points plus three times the number of points transferred by Participant X) when making her/his decision.

Both participants have to announce transfer shares that lie between 20% and 80% of the balance in their accounts. Since Participant X's account has an initial balance of 100 points her/his transfer share will correspond to a number of points between 20 and 80. These points will leave the account of Participant X and will be tripled when deposited into Participant Y's account. Participant Y will therefore receive a number of points between 60 and 240, which will be added to the 100 points in her account. In sum, Participant Y will have between 160 and 340 points in her account, of which s/he can transfer a share between 20% and 80%. These points will be transferred from Participant Y's account and will be deposited into Participant X's account.

Both participants will be prompted by the computer to enter their decisions, expressed as percentages anywhere between (but including) 20% and 80%. We will refer to the decisions as "X's TRANSFER SHARE" and "Y's TRANSFER SHARE".

When Participant X chooses X's TRANSFER SHARE s/he will not know Y's TRANSFER SHARE. Equally, when Participant Y chooses Y's TRANSFER SHARE s/he will not know X's TRANSFER SHARE.

There is one important detail about the transfer out of Participant Y's account. The computer adjusts the share that is actually transferred from Participant Y's account to Participant X's account. More specifically, the computer will adjust Y's transfer share in a random way, increasing or reducing it by up to 20 percentage points. That is, the computer will generate a number that we call "CHANGE TO Y's TRANSFER SHARE" by picking a random percentage number among -20%, -19%, ..., 0%, ..., +19%, +20%. Each of the whole-numbered percentages in this range is equally likely.

The number drawn by the computer cannot be influenced by the participants.

Therefore, the total share that is sent out of Participant Y's account, and that we call "Y's ADJUSTED TRANSFER SHARE", is equal to:

Y's ADJUSTED TRANSFER SHARE = Y's TRANSFER SHARE + CHANGE TO Y's TRANSFER SHARE

Y's ADJUSTED TRANSFER SHARE is a percentage number between 0% and 100%. (Please note that if the CHANGE TO Y's TRANSFER SHARE is a negative number, e.g. -20%, then

its absolute value (in the example, 20%) will be subtracted from Y's TRANSFER SHARE even though it is "added" (+) in the above formula. Adding a negative number is like subtracting its absolute value.)

Before the interaction starts, the computer will inform Participant X about the randomly drawn value of the CHANGE TO Y's TRANSFER SHARE. S/he will see an announcement on the screen, stating,

"The computer's randomly drawn CHANGE TO PARTICIPANT Y's TRANSFER SHARE is XX%."

(XX is the randomly chosen number between -20 and 20.)

Therefore, Participant X will know the CHANGE TO Y's TRANSFER SHARE drawn by the computer before making her/his decision. However, participant Y will not her/himself be informed of the CHANGE TO Y's TRANSFER SHARE drawn by the computer, before making her/his decision.

Importantly, keep in mind that it is Y's ADJUSTED TRANSFER SHARE that determines the exact transfer from Y to X. When making her/his decision, Participant X will know one of the two components of this share (the random draw by the computer) but s/he will not know Participant Y's TRANSFER SHARE.)

If you have any questions about the instructions please raise your hand.

(END OF PART I HANDOUT)

Understanding Test:

Before we proceed we ask you to answer the following five questions. Once you have answered all of them correctly, you will move on to the decision stage of Part I.

Please note that we make a calculator available to you on the screen. You can access the calculator by clicking on the Calculator icon. The calculator will remain available throughout the experiment.

You will receive immediate feedback when you submit your answer to each of the questions. If your answer is incorrect you will be asked to try again, and as many times as you need. However, after several failed attempts please raise your hand and we will come to your desk to explain any

open questions.

1. The initial balance in both participants' accounts is 100 points. Suppose that you are Participant X, and you choose

X's TRANSFER SHARE = Q1X %[subject specific random number].

How many points will be in the account of Participant Y, available for him/her to transfer to you? _____

Please click OK.

[In case of a mistake an error screen appears, saying "Your answer is not correct. Please try again. If you need help, raise your hand and an experimenter will come to your desk." Likewise for all other questions.]

2. The initial balance in both participants' accounts is 100 points. Suppose that you are Participant Y, and Participant X chooses

X's TRANSFER SHARE = Q2X%[subject specific random number].

How many points will be in your account, available to transfer to him/her? _____

Please click OK.

3. The initial balance in both participants' accounts is 100 points. Suppose that you are Participant X, and you choose

X's TRANSFER SHARE = Q3X%[subject specific random number].

How many points will be in the account of Participant Y, available for him/her to transfer to you? _____

Please click OK.

4. The initial balance in both participants' accounts is 100 points. Suppose that you are Participant X and you choose

X's TRANSFER SHARE = Q4X%[subject specific random number].

Suppose further that the other participant (Y) chooses

Y's TRANSFER SHARE = Q4Y%[subject specific random number],

and that the computer's random adjustment is

CHANGE TO Y's TRANSFER SHARE= T4%[subject specific random number].

How many points will you have in your account after both transfers? _____

Please click OK.

5. The initial balance in both participants' accounts is 100 points. Suppose that you are Participant X and you choose

X's TRANSFER SHARE = Q4X%[Q4 subject specific random number].

Suppose further that the other participant (Y) chooses

Y's TRANSFER SHARE = Q4Y%[Q4 subject specific random number],

and that the computer's random adjustment is

CHANGE TO Y's TRANSFER SHARE= T5%[subject specific random number].

How many points will you have in your account after both transfers? _____

Please click OK.

You have completed the understanding test successfully. Please note that none of the numbers that were given in the above questions are meant to be suggestive of what anyone may want to decide in this experiment. They only serve as an illustration, for the sake of the understanding test.

Please click OK.

This is the DECISION STAGE - Part I.

You are PARTICIPANT X.

The computer's randomly drawn CHANGE TO Y's TRANSFER SHARE is

CHANGE TO Y's TRANSFER SHARE = DX% [subject specific random number]

Please enter your transfer share (a percentage between 20% and 80%):

X's TRANSFER SHARE = _____%

If for some reason you want to change your decision, simply re-enter a new number. You have to confirm your decision (by clicking the OK button) to make it final. Once you confirm your decision you will not be able to change it.

[Screen for the Trustee, with instrument:]

This is the DECISION STAGE - Part I.

You are PARTICIPANT Y.

Please enter your transfer share (a percentage between 20% and 80%):

Y's TRANSFER SHARE = _____%

If for some reason you want to change your decision, simply re-enter a new number. You have to confirm your decision (by clicking the OK button) to make it final. Once you confirm your decision you will not be able to change it.

7 Appendix C: Detailed data summary

In this appendix we discuss subjects' compliance with Nash equilibrium and simple dominance in each of the games, as well as their best response rates and other data patterns. The CTG game under the no instrument condition has a unique Nash equilibrium, where both the trustor and the trustee transfer the minimum allowed, 20% of the pie. Compliance with equilibrium is 32.6% (30.3%) for the trustor and 50.5% (51.6%) for the trustee in the no instrument condition, CTG-NI (in the instrument condition, CTG-I) in the pooled data. Therefore, we can say that compliance with equilibrium is higher for the trustee (for whom the Nash equilibrium strategy is also dominant) than for the trustor. When the CTG game is played under the instrument condition and the shift z weakly exceeds 0.14, it is a dominant strategy for the trustor to transfer a share of her endowment as large as she is allowed, 0.8, since the amount that she transfers is multiplied by three, and she is guaranteed to receive at least 0.34 of this total amount, thus a larger amount than she intially sent. In the 22 instances with shifts greater than or equal to 0.14, the trustor transfes 0.8 only 7 times, making compliance around one third. Although the play data reveals a very high level of violations of dominance, it is well known that the trust game invokes non-pecuniary preferences, which makes it hard to draw conclusions about subjects' rationality by looking at the play data. Nevertheless, it is still informative to evaluate compliance with dominance in strategic situations that invoke non-pecuniary preferences in order to assess the strength of dominance arguments. Although under the no instrument condition all the beliefs that can be stated (in the interval $[0.2, 0.8]$) are rational, the introduction of the shift z in CTG/I limits the range of beliefs which are rational to the interval $[0.2 + z, 0.8 + z]$. Given that trustor subjects can state any belief in the interval $[0, 1]$, we can count how many beliefs fall outside the "admissible" interval. We find that trustor subjects stated "inadmissible" beliefs in only 5 out of 122 cases, a low frequency that is line with the frequencies with which dominated actions are played in games that do not invoke augmented preferences (see for example, Costa-Gomes and Crawford (2006) and Costa-Gomes and Weizsacker (2008)). In this, rationality is not thrown out of the window in trust games, even if the subjects' choices are often most easily explained by non-pecuniary preferences.

The MTG has a unique (weak) Nash equilibrium, $(top, right)$. Looking at the pooled data we

find that Column subjects play *right* 68.9% and 73.7% of the time and that Row subjects play *top* 73.8% and 49.5.% of the time in the no instrument (MTG-NI) and in the instrument (MTG-I) conditions, respectively. Although none of the players has a dominated action in the MTG game when the game is played without the instrument, the addition of the instrument makes *top* a dominated action for Row whenever the number of robots playing $left$ is larger than or equal to 9. In this case, playing *bottom* yields an expected payoff larger than 120, which makes *top* dominated. Row players choose *top* 8 out of the 27 times in these cases. This high frequency of violations of dominance is cause for concern, but it may well be driven by non-pecuniary preferences. As in the CTG game, violations of rationality in the MTG are better assessed by analyzing subjects' stated beliefs. In the instrument condition, the presence of robots imposes constraints on the interval of beliefs that Row can hold about the actual play of Column. When the number of robots playing $left$ is x, the probability with which $Left$ is played has to lie in the interval $[0.5(x/10), 0.5 + 0.5(x/10)]$. We observe that our subjects state beliefs outside this interval in 23 out of 95 cases, thus violating dominance considerably more often than it is usually observed. This is the main reason why we focus on the CTG data in the analysis.

We now consider some other features of subjects' stated beliefs, namely their level of accuracy in predicting the play of opponents. In the CTG/I and CTG/NI, respectively, the average expected actual transfer shares stated by the trustor are 35.02% and 33.02%, which differ from their corresponding average actual transfer shares, 30.54% and 30.29%, by less than five percentage points. Thus, we can say that the population of the trustor subjects predict the behavior of the population of trustee subjects with considerable accuracy. In the MTG, a similar level of accuracy is observed: in both instrument conditions, the average stated belief of actual play of $left$ by the Column subjects differs from actual play by approximately five percentage points, 35.93% versus 31.15% in MTG/NI and 44.91% versus 39.60% in MTG/I (after adding the effect of the robots).

We now examine the consistency between trustor subjects' stated beliefs and their choices of transfer shares by determining best-response rates. In the CTG best-response rates are rather low, with transfer shares being best responses to stated beliefs in 41 out of 95 cases in the no instrument condition, and even lower in the instrument condition, with only 39 out of 122 cases. Best-response

rates are substantially higher in the MTG, with trustors playing best reponses in 61 ouf ot 95 instances in the instrument condition, and 76 out of 122 in the no instrument one.

An interesting feature of the trustors' stated beliefs in the CTG is that the re-payment percentage that she expects is a multiple of 5%. When the game is played without the instrument this is the case 90.5% (86/95) of the time. When the instrument is present the frequency of such beliefs drops to 38.5% (47/122). However, that drop can be explained by the fact that the shift is not a multiple of 5 percentage points, but a percentage value between -20% and 20%. If we subtract z from the belief statements (thereby generating a hypothetical belief about the opponent's intended re-payment), the frequency increases to 78.7% (96/122). This feature of the data is reassuring because in another strategic setting Costa-Gomes and Weizsäcker (2008) have oberved that subjects very often state beliefs as multiples of 5%. The CTG/I data show that subjects are able to make more precise probability statements, depending on the infomration that is given to them.

In MTG, we observe like in Costa-Gomes and Weizsäcker (2008) that the relative frequency of the stated beliefs that are multiples of 5% is quite high, at 92.6% (113/122) when there is no instrument and 91.6% when the instrument is used.