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 (UCL)
Georg Weizsäcker (UCL & DIW Berlin)

January 2012

Abstract: In many economic contexts, an elusive variable of interest is the agent's belief about relevant events, e.g. about other agents' behavior. A growing number of surveys and experiments ask participants to state 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 informing the agent about exogenous manipulations of the relevant events. We conduct trust game experiments where the amount sent back by the second player (trustee) is exogenously varied. The procedure allows detecting causal links from beliefs to actions under plausible assumptions. The IV-estimated effect is significant, confirming the causal role of beliefs. It is only slightly and insignificantly smaller than in estimations without instrumentation, consistent with a mild effect of social norms or other omitted variables.

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

JEL Classification: C72, C81, C91, D84

[1]We thank Orazio Attanasio, Charles Bellemare, Jörg Breitung, Syngjoo Choi, Costas Meghir, Lars Nesheim, Thomas Siedler and audiences at Autonoma de Barcelona, DIW Berlin, Exeter, Glasgow, Innsbruck, Jena, Paris I, Royal Holloway, UCL and WZB for their comments. We are grateful for the financial support from the U.K. Economic and Social Research Council (ESRC-RES-1973), the European Research Council (ERC-263412) 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. E-mail addresses: m.costagomes@abdn.ac.uk, s.huck@ucl.ac.uk, g.weizsacker@ucl.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 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 the nature of choice. 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 a literal interpretation of expectations, 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 through changing people's expectations. Many policy campaigns attempt such influencing—from asserting the reality of climate change to bolstering consumer confidence—where beliefs are targeted to bring about behavioral change. In particular, many policy campaigns are geared towards creating trust or optimism, relying on their self-fulfilling powers: if the policy maker 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. The campaign may thus instigate a shift from a less desirable outcome to a better one.

But the role of beliefs first needs to be affirmed. Researchers have increasingly turned to belief elicitation procedures where the agents state their expectations explicitly. Trust game experiments (following Berg, Dickhaut and McCabe, 1995) provide a frequent context for such methods. Fehr, Fischbacher, Rosenbladt, Schupp and Wagner (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 trust on stated expectations. Yet it remains unanswered whether the variance in trust 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.

A 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, given that it maximizes social surplus and thus opens up the possibility of mutually beneficial exchange. Whether or not such social norms influence the investment choices may depend on multiple unobservable factors, e.g. the participants' education, cultural influences or even the framing employed in the experiment. But such unobservables will influence both beliefs *and* actions. In particular, it may be that the same participant who invests a large amount also predicts that the other participant 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 a "behavioral" phenomenon but can arise as an equilibrium outcome of a natural game of incomplete information. We develop a simple illustration of this in Appendix A. Players interact in a mini trust game with just two actions for each player: whether to trust or not, and whether to reciprocate or not. Both players are aware of the social norm that prescribes trust and its reciprocation. There prevails some uncertainty about whether deviations from the social norm will be sanctioned—for example, by the possibility that the players' anonymity be lifted. Players receive signals about the likelihood of sanctions, e.g. from clues in the description of the choice environment. As both players receive the same description, these signals are correlated.

The appendix shows that even with relatively little correlation between the players' signals the Bayesian Nash equilibrium involves a strong correlation between the trustor's own action and her belief about the opponent's action. The driver of both variables is the trustor's perception of the likelihood of sanctions (a variable that is omitted in most empirical analyses). The example also shows that despite the strong correlation between the trustor's belief and action, an exogenous shift of the trustor's beliefs about the opponent's action would have a relatively small effect on her action. It would therefore be misleading to interpret the strong correlation between beliefs and actions as saying that one drives the other.

This example only suggests one particular omission in the analyst's model—yet many other omitted variables apart from social norms might have an effect on actions, and social norms may not even be the most relevant one. The example's message is merely that the players may well have good reasons (here, play an equilibrium in a larger game) to exhibit strong correlations between beliefs and actions that the researcher may mis-interpret as a causal relation. To measure the effect of a belief change on actions, one needs more powerful observations than simple correlations.

In Section 2, we introduce a technique to measure the effect in the context of a trust game, involving the artificial creation of an instrumental variable. The game we examine is a simultaneous version of Berg, Dickhaut and McCabe's (1995) trust game and the instrument is a random zero-mean shift that exogenously increases or reduces the trustee's level of re-payment. The realization of the random shift is known to the trustor, thus affecting her belief about the final level of re-payment, and potentially affecting her 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.[2] The trustor's belief about the trustee's behavior (her chosen level of re-payment prior to its manipulation through the shift) should therefore also be unaffected by the realization of the shift. Our data confirm these predictions. At the same time, the

[2]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. In addition to addressing a different question, these studies also differ from ours because they replace the trustor's action by a random move, whereas we manipulate the trustee's move.

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 an IV estimation. Regarding the "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 (the level of re-payment), 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 cannot affect potentially influential variables like personal characteristics or perceptions of social norms.

To check for the validity of the design, the trust game is played under two different conditions—with and without the instrument. The no-instrument condition is a control that serves two key purposes: it allows checking whether the introduction of the instrumentation technique has any undesirable influences on the data generating process and it generates the benchmark "naive" estimate of the connection between beliefs and actions. Consistent with the previous literature, we find a strong correlation between the two variables.

The IV results, using the data from the instrument condition, indeed establish a causal link between beliefs and actions. The results reported in Section 3 show that the exogenous belief variation has a strong and significant impact on choices. The average marginal (proportional) effect of beliefs on actions is 0.5, that is, if beliefs about the opponent's action increase by ten percentage points, investments increase by five percentage points. However, the IV-estimated effect of beliefs on actions is not quite as strong as the non-instrumented analysis suggests—the coefficient is roughly by one third (yet insignificantly) smaller than the "naive" estimate.

These findings constitute, to our knowledge, the first laboratory evidence supporting that beliefs are causal for actions in games. From a methodological point of view, our paper emphasizes the difficulty of ascribing differences in behaviors to observed differences in stated expectations. Causal links between beliefs and actions were implicitly suggested not only in experiments with belief elicitation (McKelvey and Page, 1990, Offerman, Sonnemans and Schram, 1996, Croson, 2000, Huck and Weizsäcker, 2002, Nyarko and Schotter, 2002, and many later studies) but also in survey studies that use stated expectations about relevant market variables (see Manski, 2004, and

Attanasio, 2009, for useful overviews). Both literatures contain rich sets of observations that are consistent with a causal influence of beliefs on actions, but the endogeneity of beliefs and actions is rarely addressed in the analyses. Notable expceptions are the papers by Bellemare, Kröger and van Soest (2008, 2011) and Bellemare, Sebald and Strobel (2011) who estimate structural econometric models that include covariance between beliefs and actions.[3] These models allow the parameters of an agent's other-regarding preference to be jointly determined with beliefs—an endogeneity that is confirmed in the data. In the model of Bellemare, Kröger and van Soest (2011), the quantitative effect of second-order beliefs on actions is identified even without exogenous randomization. We view our method of generating instrumental variables as complementary to these structural approaches. The latter approaches can address endogeneity without interfering with the decision environment, but the structural formulations are specific to the games that are studied.

A further important set of close relatives to our paper are field experiments that vary informational conditions in different economic contexts, see e.g. Jensen (2010) and Dupas (2011).[4] Under some assumptions about how information maps into beliefs and actions, one can interpret these field studies as evidence for an effect of beliefs on actions. Our experiments complement them by allowing for a consistent estimate of the size of the effect and by offering results in a clean laboratory setting. Exogenous variation of artificially introduced instruments may be attractive in other contexts as well, as the designers of experiments and surveys will typically have the freedom to create such variations. This is further discussed in the paper's conclusion in Section 4.

[3] Bellemare, Kröger and van Soest study first-order beliefs of proposers (2008) and responders (2011) in the ultimatum game. Bellemare, Sebald and Strobel (2011) study second-order beliefs in a sequential game akin to the trust game.

[4] A related literature is summarized in Guiso, Sapienza and Zingales (2006, 2009) showing evidence of a causal role of culture on both actions and beliefs.

2 Experimental design: Instrumental variables for belief statements

Our experimental design revolves around a continuous trust game with two players. We study two versions of this game, the game with instrument (Condition I) and the game without the instrument (Condition NI). In addition to the choice data we collect for each game trustors' beliefs about the actions played by the trustees.

The game without the instrument serves as a control. First, it allows for an important empirical check of the validity of the instrument: whether or not it affects behavior in undesirable ways. In field studies that involve IV methods, this is less of a concern as the instrument is usually part of the natural decision making environment. But with an artificial instrument we must check that the instrumentation technique is neutral in the sense that its presence alone does not distort the data generating process.

Second, and no less important, the control condition without instrument provides the comparison benchmark for our IV results: it is the usual laboratory environment for trust games. It would be misleading to compare the IV estimates with non-instrumented estimates—e.g. OLS—using the data collected under the instrument condition. This is because the instrument affects the beliefs in a random way so that the beliefs exhibit additional variance. Under the hypothesis that an omitted variable is at work, the OLS analysis on the data with instrument therefore yields a biased (attenuated) estimate of the relationship of interest.

We note that in all experimental sessions subjects also played a second type of trust game with binary actions. This game, too, was played in a variant with and a variant without an instrument. However, as documented extensively in the paper's previous version (Costa-Gomes, Huck, and Weizsäcker, 2010), the instrument employed in the binary trust game failed our tests for invasiveness and hence we focus here on the continuous trust game.[5]

[5] The instrument used there is different from the instrument used in the continuous trust game. In the earlier version we also carefully examine whether the continuous game data can be analysed separately from the binary data. The experimental design involves four types of trust games (either continuous (CTG) or binary/mini (MTG) and either with (I) or without an instrument (NI)) and the protocol was such that each subject played just two

For the collection of belief statments, we employ a quadratic scoring rule that is incentive compatible in the sense of theoretically eliciting the mean of the subjectively expected distribution, under the assumption that subjects are risk neutral.[6]

We conducted our experimental sessions at University College London and at the University of York, with a roughly equal number of subjects in each treatment at each location, as reported in Table 1.[7] In all, 434 experimental subjects participated in our sessions. Subjects earn points by playing two games and one belief elicitation task (see footnote 5), which are then converted into money at an exchange rate of 2.5 pence per point, resulting in an average variable payment of £13.12.[8] Sessions lasted about 90 minutes from the moment the subjects were seated until leaving

games (without feedback) ensuring that she would play once in a binary and once in a continuous game, once with and once without an instrument, and once as a trustor and once as a trustee. Therefore, we had four treatments: CTG/I&MTG/NI, MTG/NI&CTG/I, CTG/NI&MTG/I and MTG/I&CTG/NI. We detected no spillover from the binary game onto the continuous trust game data.

[6] The quadratic scoring rule has been used by numerous researchers and—although not all studies agree (see e.g. 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. In principle, subjects could use their belief statement to hedge their position (that is, they could bet on low returns if they have invested a lot) but such behavior would require extreme sophistication and strong risk aversion.

[7] The breakdown of participants in the different treatments was as follows (cf. footnote 5):

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

[8] They were also paid a show-up fee of £5 at UCL and £4 at York, chosen in each case so as to coincide with the show-up fee of a different experiment being run at the respective lab at the same time. The overall average payment including the show-up fee was £17.60.

the laboratory after collecting their payments.

Condition	# York Participants	# UCL Participants	# Total Participants
I	124	120	244
NI	94	96	190

Table 1: Overview of experimental conditions

2.1 The continuous trust game and the shift instrument

Each of 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 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. The trustee, like the trustor, makes a decision about a relative "transfer share" a_2, not an absolute amount.

The transfer shares a_1 and a_2 are restricted to lie in the interval $[0.2, 0.8]$. Thus 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 knowing the exact balance in her account. This simultaneous version of the trust game has the advantage that the trustor's belief about the trustee's transfer share is 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 her 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, drawn from a uniform probability distribution over the 41 values on the grid $\{-0.2, -0.19, ..., 0, ..., 0.19, 0.2\}$. Both participants are informed that the trustee's transfer share a_2 is added to the zero-mean random variable z. The trustor is, in addition, informed about

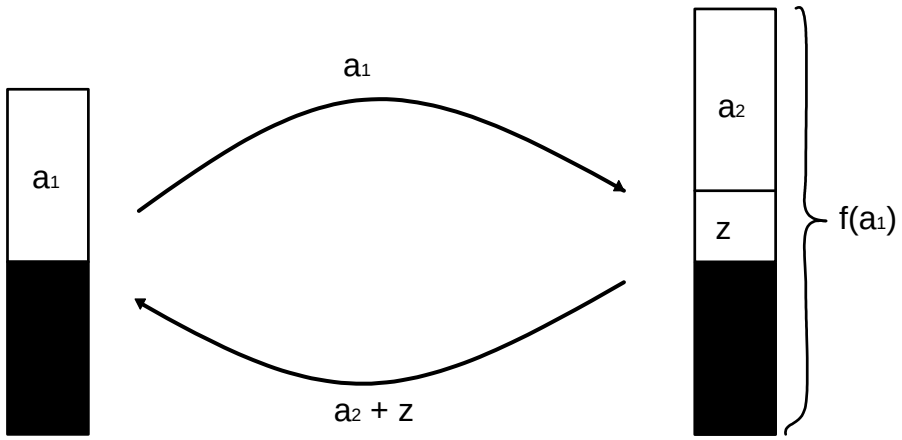

Figure 1: Illustration of the continuous trust game with instrument. Player 2 knows only the distribution of z and chooses action a_2. Player 1 knows the distribution of z and the value of z before choosing action a_1 and belief statement b_1. $f(a_1)$ indicates that player 2's account balance depends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

the realized value of z, while the trustee is not.

For example, suppose that upon being informed that the realization of the shift z is 0.05, the trustor transfers a share $a_1 = 0.5$ of her initial balance of 100 points. This would lead to intermediate account balances of 50 and 250 points for the trustor and trustee, respectively. Suppose further that the trustee decides to transfer $a_2 = 0.25$. Hence, the actual transfer to the trustor would be a share of 0.3 $(= 0.25 + 0.05)$ of the trustee's intermediate balance, leading to final balances of 125 and 175 points for the trustor and trustee, respectively. The game's rules are illustrated in Figure 1.

We explained the shifter z to participants as follows (for full instructions, see Appendix C):

> "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 instructions continue by giving a further illustration of the instrumental variable and its effects on payoffs.

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

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

where b_1 is Participant X's belief statement about Participant Y's transfer share, and the parameter values are $A = c = 250$ 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 instrument z is generated independently of all other relevant random variables. This property justifies the exogeneity assumption required for IV. Consider the bivariate 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, e.g. due to omitted variables, the instrumental variable z needs to be orthogonal to u.[9] To see how this property helps in finding the causal link between b_1 and a_1, consider the simple logic of two-stage least squares regression: the analyst regresses a_1 on z, resulting in a slope coefficient $\beta_{a,z}$, and also regresses b_1 on z, resulting in a coefficient $\beta_{b,z}$. If the only way in which z influences a_1 is through its effect on b_1 (i.e. z and u are orthogonal), it follows that the effect of b_1 on a_1 must be $\frac{\beta_{a,z}}{\beta_{b,z}}$

[9] If control variables like demographic, socio-economic, and cognitive skills are included, the analogous statement with the corresponding error term of the regression on beliefs and controls is required. See e.g. Angrist and Pischke (2009) for a wider discussion of exclusion restrictions. In the regressions of the next section, we use Tobit instead of OLS models, but the same logic applies for Tobit.

times as large as the effect of z on b_1. That is, the causal effect of b_1 on a_1 is consistently estimated by $\frac{\beta_{a,z}}{\beta_{b,z}}$. The exclusion restriction is therefore key for the causal inference—indeed we designed the experiment to make it maximally plausible. 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 a summand of $\widetilde{a}_2$, which is the statistic that beliefs b_1 are formed about, and because z does not enter the interaction in any other way.[10] Section 3 will contain results that demonstrate that belief statements are indeed strongly responsive to z. In fact, participants respond in a way that is fully consistent with the hypothesis that they simply add z to their beliefs about a_2.

3 Results

3.1 Preliminaries: Data pooling, descriptive summary and checks for invasiveness of the instrument procedure

Data Pooling. We first determine if there are any statistically significant differences between the data collected at UCL and at York, and whether there are any order effects on either actions or stated beliefs (recall that each participant played two versions of the trust game). The absence of major differences allows us to pool the data and simplify the subsequent analysis.

Initially, we pair the two treatments in which the 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 both players' transfer shares and to the trustor's belief statements and find no statistically significant order or laboratory effects, for any of the player roles or for any instrument condition.[11] Therefore, in the

[10] Of course, the functional form assumption is never innocuous but this is a general feature 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 over payoff distributions than the money-maximizing-agent model.

[11] The twelve tests on the order and laboratory effects all produced p-values above 0.1, with the exception of the test for order effects on the trustee's transfer share in the instrument condition run at York, for which we obtain a

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

Data summary and checks for invasiveness. As part of the data summary, we examine whether the presence of an instrument has undesired effects on how subjects play the games. More specifically, we want to make sure that the mere introduction of the instrument does not affect the behavioral variables except through the channel of influencing the beliefs. We focus on the trustor's data as the trustee's role in this study is accessory and only serves the purpose of generating an uncertain re-payment.

Our first step is to use the beliefs stated under the instrument condition (the beliefs about the trustee's "adjusted" transfer share after manipulation through the instrument) to construct the *underlying* beliefs about the behavior of the human opponent. We then check whether these inferred underlying beliefs are "admissible", i.e., whether one could hold such beliefs about the trustee's transfer share. More concretely, suppose that upon being informed that the shift is equal to z the trustor states that her expectation of $\widetilde{a}_2$ is a share equal to b_1. Her underlying belief is then inferred to be $b_1 - z$ and the stated belief b_1 is deemed "admissible" if the underlying belief $b_1 - z$ is in $[0.2, 0.8]$, the interval of transfer shares the trustee can choose from. A stated belief whose underlying belief falls outside this interval is "inadmissible" and indicates a potential confusion on behalf of the trustor subject. We find that only 5 subjects' beliefs (4% of trustors in Condition I) are "inadmissible", a low percentage.[12] For consistency, we exclude these 5 subjects from the analysis, unless mentioned otherwise.

Next, we check whether the instrument has any undesirable effect on the behavioral variables. The shift's expected value is zero, and thus non-invasiveness requires that none of the behavioral variables exhibits a significant change in means between the treatments with and without the shift. We first summarize the distribution of transfer shares. In the game without instrument (Condition

p-value of 0.003. Since our data analysis focusses on the trustors, the rejection of the null nypothesis for trustees at York is not problematic.

[12] Note that expressing an inadmissible belief is a strictly dominated decision, and that their low frequency is actually lower than than the frequencies of dominated actions in games, see e.g. Costa-Gomes and Crawford (2006). Additional evidence of a high level of logical consistency of trustors' stated beliefs is provided by the frequencies of observing multiples of 5% in the data, referred to at the end of this subsection.

NI) the transfer shares of the trustors follow a familiar tri-modal pattern that has been observed in many other trust game experiments, with substantial proportions of participants choosing the lowest possible transfer (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 observations are dispersed between these three modes. As can been seen in Table 2, the sample mean of the trustor's transfer share is 0.427, with a standard deviation of 0.218. For comparison, in the game with instrument (Condition I) the frequencies of transfer shares that lie on the points of the simple three-point grid $\{0.2, 0.5, 0.8\}$ are 29.9%, 7.7% and 19.7%, and the mean transfer share is 0.435 (std. dev. 0.226).[13] We conclude that with the single exception of observing fewer transfer shares at level 0.5, the features of the action data under both conditions are very similar. In particular, their means and standard deviations are close to identical and statistical tests cannot reject the hypothesis that the distributions are held constant. The same

[13]Note that in CTG/I, if the shift z weakly exceeds 0.14, it is a dominant strategy for the trustor to transfer a share of 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. In the 22 instances with shifts greater than or equal to 0.14, the trustors comply with this prediction 7 times.

holds true for the trustees' transfer shares.[14]

Condition	NI	I (All)	I ("admissible" data)
Trustor's transfer share	0.427	0.435	0.440
	(0.218)	(0.226)	(0.227)
Trustee's transfer share	0.306	0.303	0.303
	(0.144)	(0.150)	(0.150)
Trustor's belief about	-	0.330	0.331
adjusted transfer share $\widetilde{a}_2$		(0.185)	(0.178)
Trustor's belief about	0.350	-	-
transfer share a_2	(0.132)		
Shift z	-	-0.008	-0.013
	(-)	(0.121)	(0.119)

Table 2: Summary of behavioral variables and shift

Now consider the trustor's belief about the trustee's transfer share. In Condition NI, its mean is 0.350 (std. dev. 0.132), away from its target by less than 5 percentage points. The corresponding numbers for the game with the instrument, Condition I, are close in terms of mean (0.330) but the presence of the instrument induces additional variance in the belief statements—as it should because the shift is random and a rational subject adds the shift to her belief about the opponent. Even the size of the variance difference is very close to the predicted effect under the assumption that the subjects add the shift to their beliefs.[15] Further evidence on this is given by the fre-

[14] For the NI/I comparison of action data we conducted for each player role a Mann-Whitney test as well as a variance ratio test. None of the tests rejects the corresponding null at any conventional level.

[15] Under the assumption that the participants in Condition I arrive at their belief statements by simply adding the shift variable z to their (exogenous) belief about a_2, the two variables "belief about a_2" and z are independent and their sum of variances is thus equal to the variance of their sums. Counterfactually assuming that subjects in Condition NI were to observe the same realisations of z and add them to their beliefs, one can analogously construct a variance of "hypothetical beliefs about $\widetilde{a}_2$" in Condition NI, and compare it to the observed variance of beliefs about $\widetilde{a}_2$ in Condition I. The comparison supports the hypothesis that beliefs about the underlying a_2 are constant:

quency of belief statements on a grid with stepsize 5%: in 79.5% of our observations in Condition I are the (b_1, z)-pairs consistent with the hypothesis that the participants simply add their value of z to a belief about a_2 that lies on the grid $\{0.2, 0.25, 0.3, 0.35, ...\}$.[16] In Figure 2, we provide graphical evidence of this regularity. In the figure, the red line represents the "target", which is given by the mean behavior of trustees in Condition I plus a trustor's specific value of z. This line's slope is equal to one, and the set of points on the line correspond to the set of ex-post optimal belief statements, conditional on the subjects' information z. The black line is the Tobit regression line generated from the depicted data. As the figure shows, a prominent feature of the data is that the large majority of (b_1, z)-pairs lie on straight lines. The fact that the regression line has a slope that is statistically indistinguishable from unity (0.841, std. dev. 0.116) is also consistent with the overall picture that many subjects appear to add the instrument to an underlying belief, as predicted. Finally, a comparison of the second and third columns of Table 2 shows that the exclusion of the inadmissible data does not have much of an effect on the sample statistics.

In sum, the data analyses in this subsection show that introducing the instrument has no undesirable side effects on the distributions of the behavioral variables and that any differences conform to the theoretical predictions of the instrument's effects. We therefore conclude that we can proceed to the IV analysis of data from Condition I and compare its results to the benchmark data from Condition NI.

3.2 Regression analysis: The causal effects of beliefs

In this subsection we present the regression results to assess the causality of beliefs for actions. As discussed in Section 2 we write trustor i's transfer share a_{1i}^* as a linear function of her stated belief

the standard deviation of simulated beliefs in Condition NI is 0.181, very close to the standard deviation of stated beliefs in Condition I (0.178).

[16] For comparison, in Condition NI, 90.5% of stated beliefs about the transfer share are multiples of 5%. Stated beliefs about the "adjusted transfer share" $\widetilde{a}_2$ in Condition I are multiples of 5% in only 39.3% of all cases, indicating that many subjects form a well-defined underlying belief on the grid and then add z. For further comparison, Costa-Gomes and Weizsäcker (2008) also find in their 3x3 games that subjects state beliefs that are multiples of 5 percentage points around 90% of the time.

Figure 2: Trustors' stated beliefs upon observing transfer shifts z.

b_{1i} and a vector of control variables $\mathbf{x}_i$ of self-reported demographic information, socio-economic indicators, cognitive skills, and measures of trust:[17]

$$a^*_{1i} = \beta_0 + \beta_1 b_{1i} + \boldsymbol{\beta}_2 \mathbf{x}_i + u_i \tag{2}$$

Since a^*_{1i} is censored at 0.2 and 0.8 we regard it as a latent variable that underlies the observed value a_{1i},[18]

[17] See Appendix C for the exact wording of the questionnaire.

[18] In Condition NI, 31 and 14 observations out of a total of 95 are at the lower and upper limits, respectively. In Condition I, 35 and 23 out of a total of 117 observations are at the lower and upper limits, respectively.

$$a_{1i} = \begin{cases} 0.2 \text{ if } a^*_{1i} < 0.2 \\ a^*_{1i} \text{ if } 0.2 \leq a^*_{1i} \leq 0.8 \\ 0.8 \text{ if } a^*_{1i} > 0.8 \end{cases}$$

For the instrumentation in Condition I we also write trustor i's stated belief b_{1i} as a linear function of the transfer shift z_i and control variables.[19] We first use a two-limit censored Tobit model to estimate the relation between the latent trustors' transfer shares a_{1i} and their stated beliefs (as in expression 2) in the NI data, both with and without control variables. This is the analysis that one would carry out in order to establish causality in the absence of endogeneity problems. The results are in Table 4 (standard errors in parentheses; detailed control variables estimates in Table 1A of Appendix B). The Tobit estimates show a strong correlation of trustors' stated beliefs and their transfer shares, with a slope coefficient of 1.317.[20] Taking into account the effect of data censoring at transfer shares of 0.2 and 0.8, the result of column (1) translates (via post-regression analysis) into an average marginal effect of 0.722. That is, on average for all observations, including those at the boundary, an increase in the belief of 10 percentage points translates into an increase of 7.2 percentage points in the transfer share. In the regression with controls, the estimated effect is even larger, with a slope of 1.638 that translates into an average

[19] The dependent variable is doubly censored at 0 and 1 but these two belief values appear in less than 5% of the observations.

[20] In this and subsequent tables, the goodness of fit is measured by $\widehat{R}^2$, denoting the correlation between predicted and the observed values of the dependent variable. The difference in the number of observations between columns (1) and (2) is due to non-response values in the personal questionnaire.

marginal effect of 0.89.[21]

	Transfer Share in Condition NI	
	(1)	(2)
	Tobit	Tobit
Belief statement	1.317 (0.288)	1.638 (0.308)
Constant	-0.090 (0.112)	0.507 (0.439)
Personal controls	no	yes
# of Obs.	95	92
$\widehat{R}^2$	0.214	0.402

Table 4: Regressions estimates of trustors' transfer shares in Condition NI.

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 5 has the IV Tobit results from Condition I, showing that

[21] The estimates in Appendix B show that age has a statistically significantly negative effect on the transfer share, while the subject's father's level of education, living with a partner, and having a loan as the main source of income have significantly positive coefficients. However, none of these four effects extends to the data from Condition I. Three observations were dropped in the regression with controls due to subjects not answering some of the questionnaire's questions.

the answer is affirmative.

	Transfer Share in Condition I			
	(1)	(2)	(3)	(4)
	Tobit	Tobit	IV Tobit	IV Tobit
Belief statement	0.936 (0.230)	1.008 (0.231)	1.004 (0.389)	0.959 (0.408)
Constant	0.098 (0.086)	0.323 (.454)	0.075 (0.135)	0.343 (0.476)
Personal controls	no	yes	no	yes
# of Obs.	117	116	117	116
$\widehat{R}^2$	0.149	0.254	0.149	0.167

Table 5: Regressions estimates of trustors' transfer shares in Condition I.

As indicated in column (3) of Table 5, the IV coefficient without control variables is estimated at 1.004. This is insignificantly smaller than the non-instrumented coefficient in Condition NI.[22] However, the most important aspect of the IV results is that the coefficient is substantial and significantly different from zero. The average marginal effect is 0.5.[23] To our knowledge this is the first evidence (with the important qualifier about the structural work discussed in the Introduction) that the correlation between first-order beliefs and actions in an experimental game is indeed causal.

The results also show that within Condition I, there is no discernible difference in the results of Tobit versus IV Tobit. This is another indication that there cannot be a strong omitted-variable problem. One may worry about the observation that the Tobit coefficients differ between Condi-

[22] To obtain a statistical test for the comparison, we use the estimated standard deviations of both slope coefficient estimates. The estimates are independent and asymptotically normal. Under the null hypothesis of equal slope coefficients, the standard deviation of the difference between the slope estimates is thus estimated as $\sqrt{0.389^2 + 0.288^2} = 0.484$. The estimated slope difference of $1.317 - 1.004 = 0.313$ is within one estimated standard deviation around zero and has a t-value of $\frac{0.313}{0.484} = 0.647$. Comparing the coefficients from regressions with controls, the analogously standard deviation is $\sqrt{0.408^2 + 0.308^2} = 0.511$ and the slope difference has a t-value of $\frac{1.638-0.959}{0.511} = 1.329$. The IV Tobit coefficient is also insignificantly larger than the Tobit coefficient from the data with instrument, but this in itself is not an interesting observation because the belief statements in the condition with instrument have been affected, as they should, by the introduction of the instrument. (Cf. page 6.)

[23] The estimates in Appendix B show that none of the control variables has a statistically significant effect on the transfer share in the IV regression.

tions NI and I. The difference is insignificant, however, in a regression that includes all main and interaction effects of conditions and belief statements.[24] A further indication that omitted variables do not play a large role is that the regressions with personal control variables yield essentially unchanged results.

In Table 6 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 Condition NI (0.807 versus 1.317, 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 instrument on actions.

	Transfer share in Condition I	
	(1)	(2)
	Tobit	Tobit
Shift	0.807 (0.337)	0.763 (0.344)
Constant	0.416 (0.040)	0.739 (0.479)
Personal controls	no	yes
# of Obs.	117	116
$\widehat{R}^2$	0.059	0.138

Table 6: Regressions of trustors' transfer shares on shift in Condition I.

4 Conclusion

The paper makes two contributions. First, adding to the previous structural approaches discussed in the Introduction, it establishes that there is a causal link between beliefs and actions in an investment/trust game. The finding confirms the implicit supposition of such a link in many previous analyses of stated beliefs, both in surveys and in experiments. The question of causality

[24] To the extent that there is a difference between the two treatments, it could be generated by reciprocity: under Condition NI, trustors may want to be kind to their opponents if they expect them to be kind as well. In Condition I, part of the belief is driven by the computer draw, so a reciprocal agent may respond less to this belief.

between beliefs and actions is potentially relevant for many applied policy issues, and we point out that our "positive" evidence may not generalize to contexts outside of the clean laboratory environment.

Second, the paper discusses a new methodology—artificially created instruments in the laboratory—that can be employed to examine other questions. It has always been the hallmark of experimental economics to manipulate directly the explanatory variables of interest, allowing causal insight. Indeed, this is the main reason for why experiments have become 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 even by an experimenter. Yet in such contexts, 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. Non-linear specifications may follow in subsequent research, as may the combination of exogenous randomization with structural-model estimations. Similar procedures to ours may also be applied in studies where the explanatory variable of interest is of a different nature, but is likewise endogenous to the choice process: for example, information about past outcomes, responses to attitudinal questions, happiness reports or even neurological data. To our knowledge, Gill and Prowse (2011) is the only existing paper that employs an instrumental variable created in a laboratory—their method allows inference about the effect of past successes in tournament games. Another related analysis is in List and Millimet (2008) who use an artificially created instrument in the field to avoid selection effects in subsequent experimental interaction of participants.

An unusual feature of our study is that we explicitly question the link between expectations and 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 or a general decision-theoretic view on the topic, instead we simply take belief statements as our data. But the statistical establishment of a causal link between exüectations and actions is at least pragmatic. Indeed the empirical link may be the only thing that matters for a policy maker who runs a campaign to change expectations, in order to accomplish a behavioral

change.

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.

Bellemare, C., S. Kröger and A. van Soest (2008), "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities", Econometrica 76, 815-839.

Bellemare, C., S. Kröger and A. van Soest (2011), "Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool", Journal of Economic Behavior & Organization 78, 39-365.

Bellemare, C., A. Sebald and M. Strobel (2011), "Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models", Journal of Applied Econometrics 26, 437-453.

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", American 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. (2011), "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya", American Economic Journal: Applied Economics 3, 1-34.

Gill, D., and V. Prowse (2011), "Gender differences and dynamics in competition: The role of luck", mimeo, Cornell University.

Guiso, L., P. Sapienza and L. Zingales (2006), "Does Culture Affect Economic Outcomes?", Journal of Economic Perspectives 20, 2, 23-48.

Guiso, L., P. Sapienza and L. Zingales (2009), "Cultural Biases in Economic Exchange?", forthcoming, Quarterly Journal of Economics.

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. (2010), "The (perceived) returns to education and the demand for schooling", Quarterly Journal of Economics 125, 515-548.

List, J.A., and D.L. Millimet (2008), "The market: Catalyst for Rationality and filter of irrationality", B.E. Journal of Economic Analysis & Policy (Frontiers), Vol. 8, Article 47.

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's utility is penalized by a

term γ_i if she does not comply with the norm but plays $a_i = 0$ instead. 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 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 $a_2 = 0$ 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 cooperation: $\Pr(a_j = 1|s_i) = \sum_{\widetilde{s_j} \in \{0,1\}} \Pr(s_j = \widetilde{s_j}|s_i)(1 - \Pr(\gamma_j < \widehat{\gamma}_j(\widetilde{s_j})))$. 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. To find the solution, we aggregate over the possible range of punishment parameters and 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

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

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 as s exceeds r and u exceeds t, both by a considerable margin.[25]

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, 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

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

[25] The equilibrium is in (essentially) pure strategies, as a player with a given type has a strict best response, except for the zero-probability event that her realized value γ_i makes her indifferent, i.e. $\gamma_i = \widehat{\gamma}_i(s_i)$.

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 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: 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.[26] 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:

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

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.

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:

[26] 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.

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

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: Regression tables

The following tables replicate Tables 4 through 6 but contain the full sets of coefficient estimates. The data on personal characteristics are self reported at the end of the experimental session (see instructions in Appendix C).

	Transfer share, Condition NI	
	(1)	(2)
	Tobit	Tobit
Belief statement	1.317 (0.288)	1.638 (0.308)
Age		-0.036 (0.011)
Sex		0.009 (0.071)
Monthly budget/1000		0.098 (0.1335)
Works for money		0.047 (0.091)
Lives (alone)		0.151 (0.113)
Lives (with partner)		0.359 (0.157)
Lives (with children)		0.121 (0.175)
Lives (with others)		0.113 (0.092)
Math Course at UG level		0.002 (0.072)
# Years of mother's education		-0.023 (0.014)
# Years of father's education		0.033 (0.014)
Trusts People		-0.054 (0.058)
Expected prob. lost item returned		0.001 (0.001)
Good detector trustworthy people		-0.009 (0.040)
# Correct math questions		-0.048 (0.037)
Income source (work)		0.058 (0.140)
Income source (scolarship)		0.129 (0.140)
Income source (loan)		0.375 (0.110)
Income source (savings)		0.198 (0.124)
Income source (other)		0.114 (0.112)
Constant	-0.090 (0.112)	0.507 (0.439)
# of Obs.	95	92
$\widehat{R}^2$	0.214	0.402

Table A1: Transfer shares in Condition NI. Note: "Lives_i" indicates co-habitation in participant's household, with baseline category "with parents"; "Income source_i" indicates main source of income with baseline category"parents"; "Trusts people", "Expect prob. lost item returned" and "Good detector" correspond to survey questions about trust and trustworthiness (standard deviations in parentheses).

	Transfer share, Condition I			
	(1)	(2)	(3)	(4)
	Tobit	Tobit	IV Tobit	IV Tobit
Belief statement	0.936 (0.230)	1.01 (0.231)	1.004 (0.389)	0.959 (0.408)
Age		-0.000 (0.011)		-0.000 (0.012)
Sex		0.002 (0.083)		0.004 (0.084)
Monthly budget/1000		-0.007 (0.154)		0.006 (0.155)
Works for money		-0.018 (0.087)		-0.020 (0.089)
Lives (alone)		-0.148 (0.130)		-0.147 (0.130)
Lives (with partner)		-0.293 (0.164)		-0.285 (0.175)
Lives (with children)		0.254 (0.523)		0.256 (0.523)
Lives (with others)		0.03 (0.089)		0.029 (0.089)
Math Course at UG level		0.062 (0.079)		0.061 (0.079)
# Years of mother's education		-0.010 (0.013)		-0.010 (0.013)
# Years of father's education		0.002 (0.012)		0.002 (0.012)
Trusts People		-0.033 (0.057)		-0.030 (0.059)
Expected prob. lost item returned		-0.000 (0.001)		-0.000 (0.002)
Good detector trustworthy people		-0.025 (0.050)		-0.026 (0.050)
# Correct math questions		0.014 (0.034)		0.014 (0.034)
Income source (work)		0.001 (0.165)		0.003 (0.165)
Income source (scolarship)		-0.075 (0.110)		-0.078 (0.112)
Income source (loan)		-0.058 (0.111)		-0.060 (0.113)
Income source (savings)		-0.183 (0.148)		-0.176 (0.156)
Income source (other)		-0.155 (0.124)		-0.158 (0.126)
Constant	0.098 (0.086)	0.323 (0.455)	0.075 (0.135)	0.343 (0.478)
# of Obs.	117	116	117	116
$\widehat{R}^2$	0.149	0.254	0.149	0.167

Table A2: Transfer shares in Condition I. Note: See Table A1.

	Transfer share, Condition I	
	(1)	(2)
	Tobit	Tobit
Shift (z)	0.807 (0.337)	0.763 (0.344)
Age		-0.001 (0.013)
Sex		0.055 (0.089)
Monthly budget/1000		-0.001 (0.167)
Works for money		-0.043 (0.093)
Lives (alone)		-0.096 (0.139)
Lives (with partner)		-0.165 (0.174)
Lives (with children)		0.278 (0.568)
Lives (with others)		0.029 (0.096)
Math Course at UG level		0.067 (0.086)
# Years of mother's education		-0.011 (0.014)
# Years of father's education		-0.003 (0.013)
Trusts People		-0.007 (0.061)
Expected prob. lost item returned		0.000 (0.002)
Good detector trustworthy people		-0.039 (0.054)
# Correct math questions		0.004 (0.037)
Income source (work)		-0.033 (0.182)
Income source (scolarship)		-0.134 (0.118)
Income source (loan)		-0.090 (0.120)
Income source (savings)		-0.097 (0.156)
Income source (other)		-0.191 (0.134)
Constant	0.416 (0.040)	0.740 (0.480)
# of Obs.	117	116
$\widehat{R}^2$	0.059	0.138

Table A3: Transfer shares in Condition I. Note: See Table A1.

7 Appendix C: Instructions[27]

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

[27] The instructions are those in Treatment CTG/I (see footnote 5). 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.

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 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.

7.1 Survey

Please provide the information requested below, but do not write your name. (Please respond truthfully, to support us in our research. You can be assured that all information will be stored in a 100% anonymous way, ensuring your privacy.

Age:___ Sex:_____ Nationality: _____

Undergraduate___ Graduate___ Year of study _____.

Main Subject of Study _____

Your average monthly budget, including all expenses for food and lodging: _____

Do you currently work for money? _____

Please indicate your main source of income: _____

In your household, do you live (check all that apply): _____with parents _____alone _____with partner _____with children _____none of the aforementioned, but sharing an apartment with someone else.

Did you take a mathematics course as an undergraduate? _____yes _____no

Indicate the duration of schooling that your mother received, including any higher education, by checking the number of years that comes closest: _____4 ___8 _____12 _____16 _____20

Indicate your father's years of schooling: _____4 _____8 _____12 _____16 _____20

PLEASE ASSESS THE FOLLOWING STATEMENTS. PLEASE TICK ONE OPTION FOR EACH STATEMENT:

Generally speaking, would you say that most people can be trusted or that one cannot be too careful in dealing with people?

___ Most people can be trusted.

___ Can't be too careful.

___ Don't know.

Suppose that in the local city centre you loose your wallet with £500 inside. A random person that you do not know finds it. S/he does not know you, but s/he is aware that the money belongs to you and knows your name and address. S/he can keep the money without incurring any punishment. What do you think is the probability that s/he will return the money to you? Report a number between 0 and 100, where 0 means that the money will not be returned for sure, and 100 means that it will be returned for sure. _____

How good are you in detecting people who are trustworthy?

___ Not good at all.

___ Not very good.

___ Good.

___ Very good.

___ I don't know.

THE FOLLOWING ARE SOME NUMERICAL PROBLEMS. PLEASE ANSWER THEM AS WELL AS YOU CAN.

First problem: What is 15% of 1,000? _____

Second problem: A car rental agency charges $35 a day plus $0.14 per mile for its rental cars. If these charges include tax, what is the total cost of travelling 300 miles over 3 days in a car rented from this agency?

_____$42 _____$105 _____$125 _____$147 _____$300

Third problem: Which of the following is larger than 3/5?

_____19/35 _____13/20 _____4/7 _____7/13 _____None of the above

Fourth problem: If it takes 5 people 5 months to save a total of $5,000, how many months would it take 100 people to save a total of $100,000? _____

Fifth problem: A TV and a radio cost $110 in total. The TV costs $100 more than the radio. How much does the radio cost? _____

Sixth problem: In a lake, there is a patch of lily pads. Each day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____