Decomposing Trust[*]

Dirk Engelmann[†] Jana Friedrichsen[‡] Roel van Veldhuizen[§] Pauline Vorjohann[¶]
Joachim Winter[‖]

October 30, 2023

Abstract

Trust is an important condition for economic growth and other economic outcomes. Previous studies suggest that the decision to trust is driven by a combination of risk attitudes, distributional preferences, betrayal aversion, and beliefs about the probability of being reciprocated. We compare the results of a binary trust game to the results of a series of control treatments that by design remove the effect of one or more of these components of trust. This allows us to decompose variation in trust behavior into its underlying factors. Our results imply that beliefs are a key driver of trust, and that the additional components only play a role when beliefs about reciprocity are sufficiently optimistic. Our decomposition approach can be applied to other settings where multiple factors that are not mutually independent affect behavior. We discuss its advantages over the more traditional approach of controlling for measures of relevant factors derived from separate tasks in regressions, in particular with respect to measurement error and omitted variable bias.

JEL Classification: C90; D90
Keywords: trust; omitted-variable bias; measurement error

[*] We thank seminar and conference participants at Purdue University, the University of Konstanz, the University of Innsbruck, CERGE-EI, the Free University Berlin, the Spring 2020 online meeting of the Copenhagen Network for Experimental Economics, the 2020 ESA Global Online Around-the-Clock Conference, the SPP 1859 Annual Meeting on "Economic Decision-Making: Perspectives from Economics, History and Psychology", the Arne Ryde Workshop on Measurement, Causes, and Consequences of Social Trust, the Annual Meeting of the Verein für Socialpolitik in Basel, the European Meeting of the Economic Science Association in Bologna, the 7th CRC TRR190 retreat for helpful comments and suggestions. Financial support from the Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged. The pre-registration for this study can be found in the AEA RCT registry at https://doi.org/10.1257/rct.5146-1.0.

[†] Humboldt-Universität zu Berlin; CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Prague; CESifo, Munich. Email: dirk.engelmann@hu-berlin.de

[‡] Christian-Albrechts-Universität Kiel; WZB Berlin Social Science Center; CESifo, Munich.

[§] Lund University. Email: roel.van_veldhuizen@nek.lu.se

[¶] University of Exeter and Global Priorities Institute, Oxford.

[‖] University of Munich

1 Introduction

Trust is a frequently invoked concept in explaining differences in economic development between and within countries. Trust levels are shown to vary widely across the globe, with high levels of trust being particularly pronounced in Northern Europe (Falk et al., 2018). Several studies find evidence of a positive correlation between population levels of trust and economic growth, inflation, and trade volumes (see, e.g., LaPorta et al. (1997) for an early example or see Fehr (2009) for a review). Part of the literature relies on survey measures of general trust,[1] while another part derives a measure of trust from the behavior in versions of the trust game (Berg, Dickhaut, and McCabe, 1995), the stylized setup of which can—if only changed slightly—capture important aspects of many economically relevant interactions. However, despite a large body of research, there still appears to be little consensus on what exactly trust is or what drives the decision to trust.

In this study, we provide a comprehensive approach to decomposing trust into the factors driving it by using a set of binary trust game variations in a laboratory experiment. Previous studies point to four possible factors: risk attitudes, distributional preferences, betrayal aversion, and beliefs about the probability of being reciprocated. We analyze the role of beliefs and control for their influence by implementing a choice-list version of the trust game, in which we elicit the decision to trust conditioning on the number of reciprocators in the session. To assess the relevance of risk preferences, distributional preferences, and betrayal aversion, we compare behavior in the choice list that is equivalent to the trust game to behavior in a series of control treatments that by design remove the role of one or more of the relevant factors influencing trust. We further explain how this decomposition approach can be generally helpful for studying other decisions in which several interdependent factors play a role.

In common language, we trust someone or something if we believe the person will not harm us or something is safe and reliable.[2] In this understanding, trust is the expectation of trustworthiness. Accordingly, Falk et al. (2018) describe trust as a "belief rather than a preference" (p. 1665). Often, however, economists are more concerned with trust as an action, investigating whether a trustor is taking an action that can lead to gains if met with trustworthiness by another agent, but can also lead to losses if the trust is exploited. The action to trust may then be explained by the belief that another party will reciprocate (i.e.,

[1] The most frequently used measure asks survey respondents to answer the question "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" in a binary way by agreeing either with the statement "Most people can be trusted" or with "Can't be too careful."

[2] See for instance the entry *trust* in the Cambridge dictionary, `https://dictionary.cambridge.org/us/dictionary/english/trust`.

be trustworthy).[3]

According to Sapienza, Toldra-Simats, and Zingales (2013), however, trusting behavior in social and economic interactions is not driven by beliefs alone. It is therefore better captured by behavior in the trust game where preferences are allowed to play a role for trust as they would in real interactions. What these preference-related factors that affect trusting behavior are and their relative importance is still subject to debate. While Karlan (2005) emphasizes the role of risk preferences, Bohnet et al. (2008) provide evidence that trust can largely be explained by betrayal aversion. Further, Fehr (2009) argues that distributional preferences affect the decision to trust as well.[4] These aspects help to understand why trust may differ even conditional on beliefs because individuals also differ in their willingness to tolerate risk, their attitudes toward betrayal, and their distributional preferences.[5] Given the importance of the literature on trust and these heterogeneous findings, more work is needed to better understand what trust is, and how the aforementioned factors influence (dis-)trusting behavior. Our design-based approach to study how beliefs, risk preferences, betrayal aversion, and distributional preferences affect the decision to trust works by isolating the influence of individual factors.

In a first step, we present participants with a strategy-method choice list version of the trust game where we elicit their willingness to trust for different possible rates of trustworthiness. This setup eliminates the impact of beliefs about trustworthiness for each individual choice task that conditions on the rate of trustworthiness, and also allows us to investigate the impact of beliefs by comparing how decisions vary across individual choices with different rates of trustworthiness (i.e., reciprocation).[6] We then present participants with several

[3] Indeed, Ashraf, Bohnet, and Piankov (2006) and Eckel and Wilson (2004) elicit beliefs about reciprocation rates in a trust game and find that these are a significant driver of trusting behavior. Furthermore, previous results, e.g., on the positive impact of social connections on trust (Glaeser et al., 2000), are consistent with the idea that beliefs play a crucial role. Also, Sapienza, Toldra-Simats, and Zingales (2013) show that answers to the general question from the World Values Survey whether or not people can be trusted is largely a measure of beliefs about trustworthiness.

[4] We use "distributional preferences" to narrowly refer to concerns about the distribution of (material) payoffs, such as inequality aversion or standard altruism. We use "social preferences" as an umbrella term for general concerns about the outcome and behavior of others, including distributional preferences as well as reciprocity and betrayal aversion.

[5] The extent to which in particular risk preferences relate to trusting behavior remains, however, controversial. Even though a relationship is intuitive given that the decision to trust involves a social risk, evidence is mixed. While Eckel and Wilson (2004) and Houser, Schunk, and Winter (2010) find no evidence for trust being driven primarily by risk preferences, Schechter (2007) and Chetty et al. (2021) find that risk preferences relate significantly to trusting behavior. Garapin, Muller, and Rahali (2015) find that distributional preferences but not risk preferences have predictive power for trust choices. Alós-Ferrer and Farolfi (2019) survey the literature, documenting the partly diverging evidence on the role of both risk and social preferences.

[6] Fairley et al. (2016) use a similar design but their aim is to use the behavior in the treatment where participants condition on the probability of trustworthiness as a measure of risk preferences. This is problematic because social preferences can still matter in this setting as well. For example, Blanco et al. (2014) find that

additional choice lists that by design rule out any potential effects of distributional preferences, betrayal aversion, and risk preferences, respectively. To remove risk preferences, we pay senders based on the expected value of their trust action. To remove betrayal aversion, we let the receiver's decision be determined by a random draw, with the receiver still being affected by the sender's choice. To shut down distributional preferences, we remove the receiver entirely. By comparing the rate of trust or trust-equivalent actions across these sets of tasks, we can decompose trust into its underlying components. We also elicit beliefs about trustworthiness, both using reservation probabilities (see e.g., Karni, 2009) and a simple direct question.

Our main contribution lies in using a comprehensive design-based approach to decompose trust which yields not only new substantive insights on the determinants of trust but also allows us to reconcile seemingly contradictory results in the literature. A key advantage of our non-parametric design-based approach is that it avoids the measurement error critique (Gillen, Snowberg, and Yariv, 2019). Our parallel implementation of standard measures for risk and social preferences further allows us to compare our design-based method to the regression-based approach used commonly in previous work.[7] Our approach generalizes earlier attempts to decompose trust into one or more of its underlying components by comparing trust-game choices to treatments that exclude one explanatory variable.[8]

Our first main finding regarding the determinants of trust is that the belief about the likelihood of trustworthiness is a key driver of the decision to trust. In situations where the likelihood of trustworthiness is so low that the expected payoffs of trust are below the safe payoff of not trusting, very few participants trust, but the frequency of trust is steeply increasing once the likelihood exceeds this point. Hence, selfish interests to be better off due to reciprocated trust are a key determinant.

A second key finding and a proof of concept of our decomposition approach, are our results on the relative importance of the various preference-based factors that influence the decision to trust. We also identify the circumstances under which one factor dominates the other. In particular, we discover non-trivial interactions between the beliefs about trustworthiness and the preference-based determinants of trust. We show that preference-based factors only play a

social preferences even matter when trustors are informed about the actual probability of trustworthiness.

[7] This approach is used for example by Eckel and Wilson (2004), Schechter (2007), Karlan (2005), Chetty et al. (2021), and Houser, Schunk, and Winter (2010).

[8] Cox (2004) decomposes with the help of dictator control experiments the impact of expectations of trustworthiness and total-payoff concerns on part of the sender and trustworthiness and preferences for equality on part of the receiver. Bohnet et al. (2008) assess the relevance of betrayal aversion by replacing the second-mover's choice with a random move. Bohnet et al. (2008) furthermore control for the role of beliefs by asking first movers for the minimum probability of being rewarded that would make them trust. This approach is also followed by Polipciuc (2022).

significant role when beliefs about reciprocity are sufficiently optimistic. Furthermore, we find that while risk-preferences matter most for trusting behavior when beliefs are such that trust just pays off in expectation, distributional preferences and betrayal aversion become more important as beliefs about reciprocation rates increase. Our results demonstrate within a single study that trust is a multi-faceted phenomenon, which is in line with the body of earlier literature finding support for several different factors across studies as discussed above.[9]

Our insights with respect to the interaction between beliefs about trustworthiness and the relative importance of different preference-based factors in determining trust have the potential to reconcile apparently conflicting results of the earlier literature. Differences in experimental design likely influence beliefs about trustworthiness. Further, we find that depending on the beliefs, one or another preference may have a larger impact on trust rates. As we will discuss below, differences in beliefs therefore likely contribute to differences in results among various studies.

On a methodological level, our approach is generally applicable to the decomposition of the determinants of experimental choices if there are several determining factors whose relevance can be controlled by experimental design, in particular if these factors are not mutually independent. It provides a transparent alternative to the regression-based approach that has been commonly used in previous work. The regression-based approach typically involves obtaining separate measures of the determining factors and then using these measures as explanatory variables in a regression. While this approach may yield valuable insights in some situations, it is also susceptible to biases induced by omitted variables, measurement error, and model misspecification.

To illustrate the first problem, consider a case where several factors influence a decision and suppose these factors are correlated with each other. Then, not including some of these correlated factors will lead to an omitted-variable bias, where the impact of an included variable is, e.g., overestimated if it is positively correlated with an excluded variable that influences the outcome variable of interest in the same direction.[10] While such an omitted-variable bias is widely acknowledged in (non-experimental) empirical research, it is not frequently addressed or discussed in (laboratory) experimental research. It is common

[9] We acknowledge that trust decisions may further be affected by ambiguity preferences (Li, Turmunkh, and Wakker, 2019). In the choice lists, our subjects make decisions for a given probability in each single choice, and, hence, attitudes towards ambiguity should not matter. The fact that, as we will see below, choices in the simple trust game do not systematically differ from the corresponding choices in the choice lists for the stated beliefs suggests that attitudes towards ambiguity are not a crucial driver of behavior in the trust game.

[10] In the context of trust, an omitted variable problem arises in particular if beliefs about others' trustworthiness and own social preferences are correlated due to a (false) consensus effect, which in brief states that expectations about other agents' types correlate with one's own type, an issue we return to below.

in the experimental literature to include some easy-to-assess demographic variables (such as gender) and some easy-to-elicit prominent preferences (such as risk aversion) in a regression as controls or explanatory variables, but it is rarely discussed that these factors may be correlated with other factors that are not included.[11] Addressing a similar problem, Andersen et al. (2014) and Antoniou et al. (2015) point out that not controlling for risk attitudes leads to biases when estimating beliefs from choice data.[12]

The standard approach to deal with the omitted variable bias in experimental research is to elicit as many relevant influential factors as possible and include them in the regression analysis. This gives rise to the second problem mentioned above, measurement error. As pointed out by Gillen, Snowberg, and Yariv (2019), if explanatory variables are measured with error, then we are misestimating not only their impact, but also the role of any residual factor. The issue of measurement error is particularly pronounced if the explanatory factor is elicited in an unrelated task because we may not measure, and hence control for, what we intend to measure. For example, risk preferences appear to be a multi-faceted phenomenon.[13] Therefore, if we attempt to estimate someone's risk preferences in a separate task (e.g., a lottery-choice task) and then include the estimated degree of risk aversion as explanatory variable in a trust game, we may not appropriately control for risk preferences. This may be so, for example, because in the trust game, risk is social risk, whereas in a lottery choice, it is not. But measurement error may also arise if subjects do not pay sufficient attention to the respective task, if they misunderstand the instructions, or if they do not care and decide at random.

Our approach assesses the impact of a factor by a comparison of the original decision task with a variant that excludes this factor. Thus, by construction, we assess the impact of a factor exactly in the way it influences the behavior of interest (i.e., we will not assess the "wrong type" of risk preferences if we remove risk from the trust choice). Of course, our approach is not completely immune to measurement error. Behavior in the individual versions of the experimental task will also be influenced by random factors. In our approach,

[11] In the analysis of the role of economic preferences in life outcomes, Falk et al. (2018) argue that it is important to include all relevant economic preferences to avoid omitted variable bias. Ziegler (2021) analyzes the correlation of economic preferences and environmental values and points out that not including social preferences leads to substantial misattribution due to omitted variable bias.

[12] Implicitly, though, the problem of omitted-variable bias is acknowledged when variables are included simply as controls, though it is not frequently discussed explicitly.

[13] Comparing four incentivized risk elicitation tasks, Crosetto and Filippin (2016) show that the mechanics of the tasks may lead to task-specific measurement error but that the actual estimates for the different tasks vary within-subject beyond what can be expected simply from measurement error, suggesting that different tasks may measure slightly different preferences. Pedroni et al. (2017) also find little systematic overlap across various risk elicitation methods and even argue that the elicited risk preference may be constructed on the spot and, thus, depend on the specific cognitive mechanisms that are relevant in a specific elicitation method.

however, if these errors are random noise, they will cancel out, whereas measurement error in the conventional approach typically leads to systematically underestimating the impact of a factor. Measurement error in our approach would only lead to misinterpretation if it is systematically related to the specific design changes that eliminate individual factors, for which there is no good reason.[14]

Even absent measurement error and omitted variables, the regression-based approach relies on parametric assumptions that may not always be justified. For example, a typical strategy is to control linearly for risk preferences and beliefs.[15] However, under expected utility, the effect of risk preferences depends on beliefs in a non-linear way. Our approach allows us to observe and take these non-linearities into account without relying on specific parametric assumptions. In particular, we find that when restricting attention to items of the choice lists reflecting relatively pessimistic beliefs (such that trusting just about pays in expectation), risk preferences are an important determinant of choices. By contrast, restricting attention to optimistic beliefs, social preferences are more important. Identifying such patterns that help make sense of apparently conflicting evidence in the literature about key determinants of trust, appears to be difficult with the regression-based approach.

To provide a comparison between our approach (to assess the impact of a factor by excluding its relevance by design) with the standard approach (to measure a factor in a separate task and include the elicited value in a regression), we also elicit several potentially relevant factors in conventional tasks. Specifically, we elicit risk preferences in lottery tasks and distributional preferences in dictator games as well as beliefs through direct questions. As alternative measures, we additionally elicit risk and distributional preferences through questionnaire items. To address the issue of potentially using inappropriate measures of the relevant factors, we perform the comparison based not only on the estimates derived from these conventional tasks and questionnaire items, but also based on tasks that are derived from the original trust game by eliminating all influential factors except for one. For example, to assess the relevant distributional preferences, we include a choice list with payoffs to two players corresponding exactly to the expected payoffs in the trust game.

Our main insight from the comparison of the decomposition approach with the standard approach is that the standard approach may both overestimate and underestimate the relevance of certain factors, depending on whether measurement error or omitted-variable bias dominates. Due to measurement error, the impact of risk preferences is underestimated in

[14] Econometrically, this is because in our analysis choices serve as the dependent variable and treatment serves as the explanatory variable. In linear regression, which is typically used in this context, measurement error in explanatory variables biases coefficient estimates whereas classical measurement error in the dependent variable increases variance but does not bias coefficient estimates (e.g., Hausman, 2001).

[15] See, e.g., Eckel and Wilson (2004) and Ashraf, Bohnet, and Piankov (2006).

the regression-based approach. In contrast, if we do not control for beliefs, the impact of distributional preferences is overestimated due to omitted-variable bias, since own trustworthiness and the expectation of others' trustworthiness are strongly correlated in line with a consensus effect. Interestingly, the standard approach does not generally fare better if we use measures tailor-made to our task by reducing it to a version that includes no other explanatory variable.

The remainder of the paper is structured as follows. In Section 2, we describe the design of the laboratory experiment and discuss its results in Section 3. In Section 4 we compare our approach to the standard regression-based approach. We conclude in Section 5. Supplementary results and material are contained in the appendix.

2 Experimental Design

Our experiment builds upon a two-player binary version of the trust game (Berg, Dickhaut, and McCabe, 1995) in which a 'sender' and a 'receiver' each start with an endowment of 10 experimental points. The sender can choose whether to 'trust' the receiver by sending half of her endowment (5 points). If she decides to trust, the amount sent is doubled, and the receiver can decide whether to reciprocate by returning half of the amount received (5 points). If she decides to reciprocate, the amount returned is also doubled, so that both players end up with 15 points. If she decides not to reciprocate, the sender and receiver end up with 5 and 20 points respectively.[16]

Our primary interest lies in decomposing trust, that is, the sender's choice in the trust game, into its underlying components. For this purpose, we conduct a laboratory experiment in which participants make choices in fourteen different tasks, split into two parts. Part 1 consists of 8 tasks that are based on variations of the trust game plus two belief elicitation tasks. These tasks jointly allow us to decompose trust into its components. The four tasks in the second part collect auxiliary measures of social and risk preferences. At the end of the experiment, one task from part 1 and one task from part 2 are randomly chosen for payment. Figure 1 presents a summary of all tasks in the experiment.

[16] Note that since the sum of both players' payoffs increases through both trust and reciprocity, the studied trust game is equivalent to a sequential prisoners' dilemma under the assumption that in the sequential prisoner's dilemma the second mover defects if the first mover does. This assumption is consistent with social preference models unless they give a large weight on unconditional altruism or total payoff maximization. It is also strongly supported in experimental studies. For example, Blanco, Engelmann, and Normann (2011) and Clark and Sefton (2001) find 94 and 96 percent of second movers defecting after first mover defection, respectively. In the game by Berg, Dickhaut, and McCabe (1995) points sent by the sender are tripled, while those sent by the receiver are just direct transfers without further increasing the total payoff. The key properties are the same in our game as in the classical trust game, however.

Part 1: trust game choices and belief elicitation

Simple choices and belief elicitation		
1	Trust Game (Receiver)	Binary trust game in the receiver role
2	Beliefs (Simple)	Simple belief elicitation task
	Intermezzo	The game urn and random urn are explained
3	Beliefs (Complex)	Complex belief elicitation task
4	Trust Game (Sender)	Binary trust game in the sender role
Choice lists		
5	Trust Game Urn	Binary trust game strategy method in sender role
6	Other Betrayal Urn	As 5, except receiver's decision determined by third party
7	No Betrayal Urn	As 5, except receiver's decision determined by chance
8	Distribution Urn	As 7, except based on expected value
9	No Receiver Random Urn	As 5, but without receiver, chance determined by random urn
10	No Receiver Game Urn	As 5, but without receiver, chance determined by game urn

Part 2: auxiliary measures

11	Lottery Choice (without losses)	Risk preference elicitation without losses
12	Lottery Choice (with losses)	Risk preference elicitation with a potential loss
13	Dictator Game (ahead)	Dictator Game (earn more than recipient)
14	Dictator Game (behind)	Dictator game (earn less than recipient)

Questionnaire and Payment

Figure 1: Overview of the Experiment

Notes: This figure contains an overview of all the tasks used in the experiment. The first four tasks and intermezzo always occurred first and in the same order. The order of the remaining tasks is randomized, as explained in section 2.4 below.

2.1 Part 1: Choices

Participants start part 1 of the experiment by playing the binary trust game outlined above. Participants play the trust game in both roles, first as the receiver and then as the sender.[17] Deciding first as a receiver requires using the strategy method, which means we ask for a choice conditioning on the sender having trusted. We use neutral framing throughout the experiment, labeling the trust and not-trust actions as “in” and “out” and the reciprocating and not-reciprocating actions as “equal” and “unequal” respectively.

After making their receiver choices but before making their sender choices, we informed each participant that we would collect all the receiver choices made by *other participants* in their session and use them to form a (virtual) “game urn”. Specifically, we told participants that we would add a green ball to the game urn for each of the other participants in the session who reciprocated. For each of the other participants who did not reciprocate, we would add a red ball. Note that this process implies that in a session with n participants, the game urn would contain $n - 1$ balls. We then introduced participants to another urn, the “random urn”, that also contained $n - 1$ either red or green balls, but for which the urn composition was determined randomly by a computer. The explanation of the two urns uses a graphical illustration that is accessible throughout the rest of the experiment (see the translated instructions in appendix D). Participants at this stage also take part in two belief elicitation tasks (one before and one after the urn instructions) that are described in the next section.

Armed with the knowledge of the two urns, participants then make a series of decisions that are organized in choice lists. In each of these choice lists, participants make decisions that closely correspond to the sender choice in the trust game in terms of actions and payoffs. The difference is that in the choice lists participants can condition their choice on the probability that the receiver reciprocates. We represent this probability as the number of green balls present in the game urn or random urn. For example, participants decide whether they would choose “in” or “out” if no one in their session reciprocated (no green balls in the game urn), if one person in their session reciprocated (one green ball), et cetera. Each choice list contains n choices, conditioning on the respective urn containing $0, 1, \ldots, n - 1$ green balls, respectively.

We consider six different choice lists. The *Trust Game Urn* choice list directly corresponds to the sender decision in the binary trust game. In particular, each choice in the choice list corresponds to a binary trust game conditional on a different potential composi-

[17] Apart form the typical practical advantage of role reversal to obtain a data point from each participant on the main choice of interest (first-mover behavior in our study), it is essential to be able to relate first-mover trust to own trustworthiness and to assess the role of a consensus effect.

tion of receiver choices in the current session. If this choice list is selected for payment, only the decision corresponding to the actual distribution of receiver choices in the session would be implemented. In particular, the participant would be randomly matched to one other participant in the experiment, and that participant's receiver decision along with the sender's conditional sender decision would jointly determine both participants' payments, where 'conditional' refers to the decision the sender took in the row where the urn composition matches the actual composition of receiver choices in the experiment.

The primary purpose of this choice list, and of the five other choice lists, is to study sender behavior conditional on beliefs about the probability of being reciprocated. Essentially, a rational first mover in a choice list with a row corresponding to k green balls should make the same decision as in the binary trust game if this sender was sure that k out of the $n-1$ other participants in the session reciprocated. The choice list method allows us to study the role of beliefs by comparing choices across rows, as these vary the probability of reciprocation. Further, within each row, the sender's actual beliefs are irrelevant as the row conditions on one specific number of reciprocators. By presenting participants with variations of the baseline choice list that remove the role of one of the components of trust, we are then able to identify the importance of each of these factors.

We include two choice lists that eliminate the role of betrayal aversion. The *Other Betrayal Urn* is identical to the baseline Trust Game Urn, except that we now independently draw two receivers from the pool of other participants for each sender. One receiver serves as the beneficiary of the receiver payment in the game, as before; the other receiver determines which choice ("equal" or "unequal") will be implemented if the sender chooses "in". This implies that while a low payoff for the sender would be the result of a receiver's betrayal, the beneficiary of that betrayal would be a different participant. While the sender might still feel that her trust is betrayed, the sender may therefore feel less exploited in this case. The *No Betrayal Urn* instead removes all potential impact of betrayal aversion by making the outcome depend on the draw from the random urn, similar to the approach by Bohnet et al. (2008), but in a choice list. Since no other person's action influences the result, the sender cannot feel betrayed.[18]

The next task we include, the *Distribution Urn*, is similar to the No Betrayal Urn but additionally eliminates the role of risk aversion. Specifically, for this task, we no longer draw a ball from the random urn to determine payment. Instead, both the sender and the receiver are paid the expected value of the sender's choice for the given number of green balls in the

[18] By de-coupling the receiver choice that affects the sender and the receiver who benefits from the sender's trust, these treatments also eliminate further social motivations such as getting a utility increase through mutual cooperation or reciprocal kindness as in Dufwenberg and Kirchsteiger (2004).

random urn. Specifically, if the random urn contains no green balls, the sender receives 5 and the receiver 20 if the sender chooses “in”. With each additional green ball in the random urn, the sender earns $10/(n-1)$ more and the receiver earns $5/(n-1)$ less, such that if there are only green balls in the urn, both earn 15. If the sender chooses “out”, both earn 10 for sure. Either way, this task removes all risk from the decision, thereby eliminating the role of risk aversion.

The final two choice lists (the *No Receiver Urns*) eliminate the role of social preferences, including distributional preferences, by removing the payment to another player. In these tasks, no other participant is paid based on the sender’s choice, such that the tasks are just lottery choices where the sender decides between obtaining 10 for sure or obtaining 15 (as opposed to 5) with some probability. The two tasks differ in the type of risk. In the *No Receiver Game Urn* decision, the win probability in the lottery is based on the game urn, such that risk is social because the composition of the urn depends on other participants’ choices. By contrast, in the *No Receiver Random Urn* the win probability in the lottery is based on the random urn and hence there is no influence of other participants.

A few remarks are in order. First, we calibrated the payments of the trust game to ensure a sufficiently high number of participants would choose “in” in the role of the sender (i.e., choose to trust) based on the results of two pilot sessions (N=40). Second, having the receiver make a binary (as opposed to a more continuous) choice greatly simplifies the choice list tasks by virtue of allowing us to condition sender choices on only a single parameter (the number of green balls). Third, letting participants play both roles in the binary trust game under role uncertainty effectively doubles the sample size, and is therefore common practice in the literature. Fourth, using a within-subject design where the same participants go through all tasks allows us to decompose trust at the individual level, increasing statistical power.

2.2 Part 1: Belief Elicitation

To elicit beliefs about the composition of the game urn, participants go through two belief elicitation tasks after making the choice as receiver but before making the choice as sender in the trust game (see Figure 1). In the first of these tasks, participants are asked directly for their belief about the number of other participants in the session who have chosen to reciprocate as receiver. They are incentivized by a linear scoring rule. Specifically, they earn 15 points if they are exactly right and this amount is reduced by 1/2 point times the absolute deviation from the correct number if they are wrong.

The second belief elicitation task uses the reservation probability or “crossover” method (see e.g., Karni, 2009; Mobius et al., 2022). In this task, participants choose, for any possible composition of the random urn, between a lottery that is based on a draw from the random urn and a lottery that is based on a draw from the game urn. If a participant believes that k out of the other $n-1$ participants in the session have chosen to reciprocate, then she should choose a lottery based on the game urn as long as there are fewer than k green balls in the random urn; choose a lottery based on the random urn if there are more than k green balls in the random urn; and be indifferent when there are exactly k green balls in the random urn. Hence, if a participant switches from the game urn to the random urn in the row of the choice list with k green balls in the random urn, we can infer that she expects there to be anything between $k-1$ and k green balls in the game urn.

At this point, it is useful to discuss the role of ambiguity aversion. In particular, if participants are ambiguity averse, the switch point may no longer reflect their (mean) belief about the number of green balls in the game urn. This is because the random urn has an unambiguous winning probability in any row of the choice list, but the participant's belief about the number of green balls (i.e., the number of reciprocators in the session) may be ambiguous. A participant might, for example, prefer to play a lottery based on the random urn if it contains 8 green balls even if she expects that there are 10 green balls in the game urn because her belief about the number of green balls in the game urn should actually be a distribution about the possible numbers of green balls and is hence ambiguous. The belief we elicit through this method hence takes ambiguity aversion into account. We call this the "ambiguity neutral belief equivalent" (in short "belief equivalent").

Note that this feature should not be seen as a weakness of the reservation probability method but rather as a strength because it implies that the decision in the binary trust game is affected by ambiguity aversion in the same way.[19] Specifically, when considering whether for an expected number of k reciprocators a sender considers whether to trust or not, she would not only take into account her risk preferences, but also her ambiguity preferences. In the choice list task that most directly corresponds to the binary trust game (*Trust Game Urn*), risk still matters but ambiguity does not, because in each row senders choose between an option with a known probability and a safe option. Therefore, the belief equivalent is the belief that is relevant in the trust game.

One purpose of eliciting beliefs is to check whether behavior in the choice lists matches behavior in the binary trust game. Hence, we compare whether the choice in the Trust Game Urn task for the number of green balls corresponding to the elicited belief is the same as the choice in the binary trust game. For the stated reasons, the belief equivalent is what we need to do this. For simplicity, however, we will still use the term 'belief' instead of 'belief equivalent' when discussing our main results below, except in cases where we explicitly discuss the role of ambiguity aversion. Note that the belief elicited through the direct question

[19] For the relevance of ambiguity aversion in the trust game, see, e.g., Li, Turmunkh, and Wakker (2019).

does not reflect ambiguity preferences. Hence, we can also directly assess the relevance of ambiguity aversion in our setting by comparing the beliefs from the two belief elicitation methods.

2.3 Part 2: Auxiliary Tasks

In the second part, participants go through four different tasks (all choice lists) that are unrelated to the trust game. The first such task consists of a choice list in which participants make choices between a certain payment increasing from 1 to 9 across rows and a simple lottery yielding 0 and 10 with equal probability. This serves to elicit risk preferences without taking potential loss aversion into account. In the second task, participants choose whether to play a lottery yielding with equal probability a gain of 5 or a loss of c, with c increasing across rows from 0.5 to 5 in steps of 0.5 or take an outside option of 0 for sure. This choice list elicits risk preferences that take potential loss aversion into account. Given that in the trust game participants might take the certain payoff of 10 corresponding to the "out"-choice as a reference point, loss aversion might be relevant in our setting.

The remaining two choice lists elicit social preferences through modified dictator games, similar to Blanco, Engelmann, and Normann (2011). In the third choice list, participants choose between an allocation of 10 for themselves and 0 for another participant and an allocation of x for both of them, with x increasing from 0 to 11 across rows. This elicits altruism or spite towards participants with lower payoffs. In the fourth choice list, participants choose between 5 for themselves and 10 for another participant and increasing equal allocations as in the previous choice list. This task elicits altruism or envy towards participants with higher payoffs.

2.4 Remaining Procedures

In order to control for possible order effects among the tasks, we employ two different sequences of the tasks. Both sequences start with the receiver choice in the binary trust game, followed by the simple belief question, the choice-list belief elicitation, and the sender choice in the binary trust game. Then the two choice lists where no other player is involved follow (9 and 10 in Figure 1), either with the task based on the game urn first or the task based on the random urn first. This is followed by the remaining four choice lists based on the trust game, either moving in the order Trust Game Urn - Other Betrayal Urn - No Betrayal Urn - Distribution Urn or vice versa. In the second part, the lottery tasks are performed first (either the lottery with loss first or the lottery without loss first), followed by the two modified dictator games (either the game involving disadvantageous inequality first or the game involving advantageous inequality first). All participants in a session faced the same sequence and we balanced sequences across sessions.

Participants were paid for one task from the first part and one task from the second part. The paid tasks were the same for all participants in the session. If the chosen task was one where the participant's choice did not affect another participant's payoff, all participants were assigned the same role and were paid for their own active choice in this task. If, instead, the chosen task involved a second participant whose payoff would be affected, participants were randomly assigned to the active or passive role and then randomly matched. If the (non-choice list) trust game was chosen for payment, half of the participants were assigned the sender role and the other half were assigned the receiver role, they were matched in pairs, and their payoffs were determined by their respective choices.

After all participants had completed the first task (receiver choice in the binary trust game), at which stage the random and game urns were explained, they went through the remaining tasks at their own pace. After all participants had completed all tasks, they received feedback on the two tasks they were paid for, being reminded of their own respective choice if a task and role was chosen where their own choice mattered for their payoff. If a task was chosen where the choice of another participant mattered for their payoff, they were informed about this choice. They were also informed about their earnings from each part and their total earnings.

Following the feedback, participants answered a questionnaire asking for self-assessments regarding risk preferences, trust, and social preferences, see Appendix E. Finally, a questionnaire with standard demographic items was filled in.

The payoffs from the two tasks that were chosen for feedback were converted at a rate of 0.75 Euro for each point. In addition, participants received a 3 Euro show-up fee. Average earnings (including the show-up fee) were 16.92 Euro. The sessions took between 51 and 72 minutes, with 28 to 49 minutes for the decision making and the additional time for reading instructions, answering the questionnaire, and receiving payment. The experiments were conducted at the WZB-TU laboratory at the Technical University Berlin between December 11 and December 18, 2019. We conducted 12 sessions, 11 with 20 participants each and one with 16 participants.[20] Out of 236 participants in total, 131 identified as male, 101 identified as female, 1 identified as diverse and 3 did not indicate a gender. The average age of participants was 22.5, and participants most commonly majored in engineering (39%), economics or business (22%), and mathematics or natural sciences (16%).

[20] We had aimed at obtaining 200 participants. Expecting some sessions to be run with fewer than 20 participants due to no-shows, we had recruited for six rather than five sessions per sequence. In the end, we had relatively few no-shows and only one session did not fill. We also conducted two pilot sessions on November 26, 2019 to calibrate the parameters of the trust game (N=40).

The experimental software was programmed and the experiments were run with z-Tree (Fischbacher, 2007). The participants were recruited using ORSEE (Greiner, 2015). Written instructions were provided for the general set-up of the experiment, including illustrations of the game urn and the random urn. Instructions for the individual tasks were provided on the screen. See Appendix D for the translated instructions.

3 Results

3.1 Removing Inconsistent Responses

Out of 236 participants in the experiment, 64 (27%) switched multiple times or switched in the 'wrong' direction on at least one of the choice lists. Following the criteria laid out in the pre-analysis plan, we remove these participants from our main analysis. In addition, we remove another 33 participants who made dominated choices. Specifically, we remove from the sample 10 participants who, in the 'No Receiver Random Urn' choice list, either selected the risky alternative when it had a 0% chance of paying out (i.e., preferred 5 points over 10 points), or selected the safe alternative when the risky alternative had a 100% chance of payment (i.e., preferred 10 points over 15 points). We also remove 23 participants who preferred (10,10) over (15,15) in the last row of any of the other price lists.

These exclusions leave us with a sample of 139 participants for our main analysis.[21] In line with our pre-analysis plan, we replicate all of our analyses using different sample restrictions in Appendix B.[22]

3.2 A First Look

In the baseline trust game, 34 out of the 139 participants (24.4%) chose to trust the other participant and 52 out of 139 participants (37.4%) chose to reciprocate if given the opportunity to do so. When asked how many others they thought would reciprocate, the average answer corresponded to 43.1% on the simple task and 42.8% on the measure based on choosing between game urn and random urn (equivalent to approximately 8 out of 19 participants in a session with a total of 20 participants), both of which are close to the actual number for the full sample (93 out of 236 participants or 39.4%). The two belief measures are also

[21] Given that we have multiple criteria for exclusion and our tasks are quite challenging, we overall remove a relatively high share of participants.

[22] The number of violations based on the pre-registered criteria ranges from 12 in the Distribution Urn to 41 in the complex belief elicitation task, with the remaining urn-based tasks having between 22 and 33 violations. Based on the stricter criteria, the respective numbers are 21 for the Distribution Urn, 64 for the Trust Game Urn and between 32 and 42 for the remaining urn-based tasks.

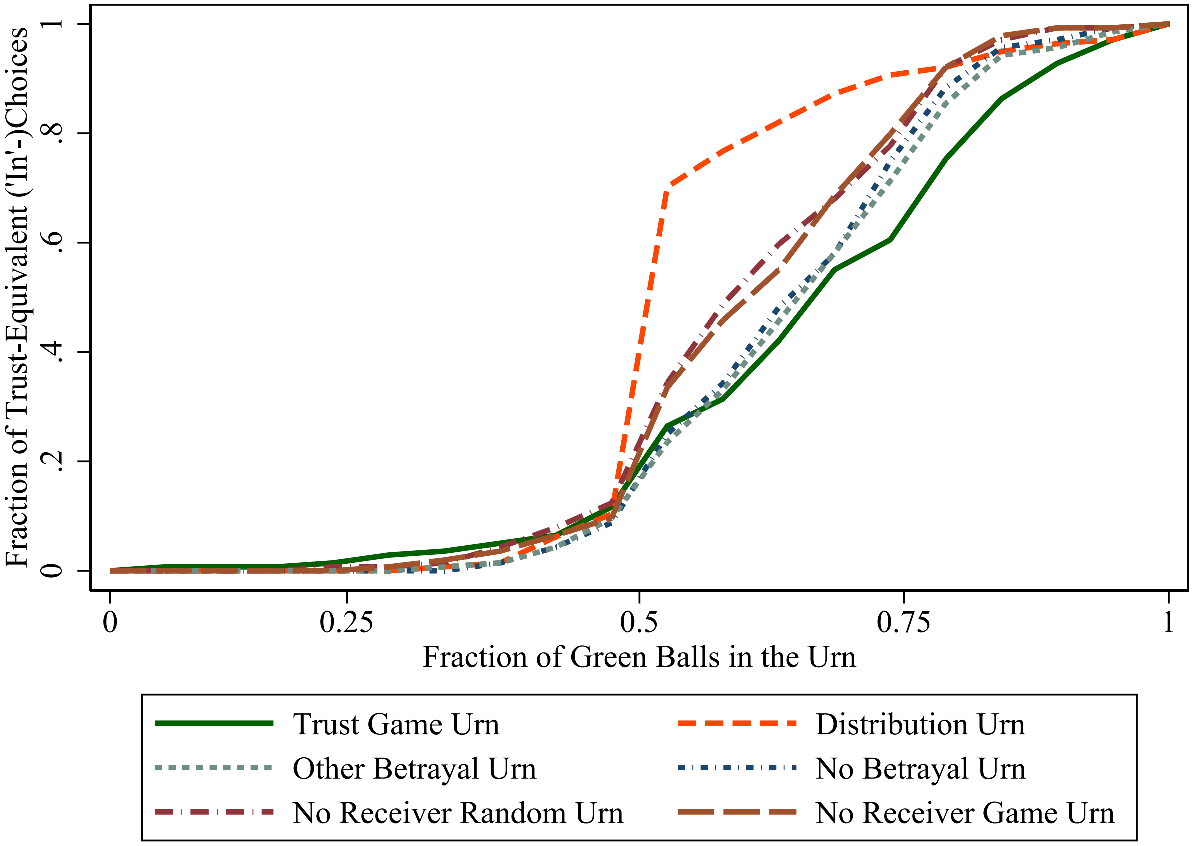


Figure 2: Distributions of Choice List Actions

Notes: These figures plot the fraction of ‘in’- (or trust-equivalent-)choices as a function of the fraction of green balls for all six urn choice lists. The figure uses data from the 139 consistent participants.

highly correlated ($r = 0.51$, $p < 0.001$). The fact that the average belief about the number of reciprocators is nearly identical in the simple task, which should not be affected by ambiguity aversion, and the measure based on the choice between random and game urn, which would be downward biased through ambiguity aversion suggests that ambiguity aversion does not play a major role in our setting.

Figure 2 plots the raw fraction of “in”-choices in the six urns as a function of the fraction of green balls in the urn. Since “in”-choices are equivalent to choosing the ‘trust’ action in the trust game, we will refer to them as ‘trust-equivalent’ choices in the rest of this section. The figure reveals several patterns. First of all, the probability of a positive outcome (i.e., the fraction of green balls) appears to be a key driver of the trust-equivalent action. When the respective urn contains very few green balls (left side of the graph), no participant chooses the trust-equivalent action in any urn. When nearly all balls in the urn are green (right side of the graph), nearly all participants choose the trust-equivalent action. Second, differences

	(1) Trust-equivalent	(2) Trust-equivalent	(3) Trust-equivalent
Other Betrayal Urn	0.011	-0.017*	0.039**
	(0.011)	(0.009)	(0.017)
No Betrayal Urn	0.017	-0.019*	0.054***
	(0.011)	(0.010)	(0.018)
Distribution Urn	0.103***	-0.015	0.220***
	(0.014)	(0.011)	(0.024)
No Receiver Random Urn	0.052***	-0.006	0.109***
	(0.012)	(0.010)	(0.020)
No Receiver Game Urn	0.047***	-0.011	0.104***
	(0.014)	(0.011)	(0.021)
Constant	0.350***	0.033***	0.667***
	(0.014)	(0.010)	(0.023)
Green Balls	All	$\leq 50\%$	$> 50\%$
Observations	16,680	8,340	8,340
Clusters	139	139	139

Table 1: Testing for Differences in Actions across Choice Lists

Notes: Regression estimates (clustered standard errors in parentheses). The dependent variable is whether the participant took the trust-equivalent action (the ‘in’-choice). The independent variables are dummies for the respective urns as described in Figure 1. The number of observations equals the number of participants times the number of urns (6) times the number of rows per choice list (either 20 or 16 depending on the session). *** p<0.01, ** p<0.05, * p<0.1

in behavior emerge between the urns when the trust-equivalent action is appealing (high fraction of green balls). In particular, the tendency to take the trust-equivalent action, i.e. to choose ‘in’, appears to be lowest in the Trust Game Urn, and highest in the Distribution Urn, where participants are paid based on the expected value of the trust action.

Table 1 allows for a first formal look at the importance of the components of trust across the entire space of possible beliefs about reciprocity. The first coefficient reveals that participants are, on average, 1.1 percentage points more willing to trust if the receiver’s decision is made by a third party. The second coefficient tells us that letting the receiver’s decision be determined by a computer increases participants’ tendency to trust by a further 0.6 percentage points. Taken together, these two urns show that removing all betrayal makes participants 1.7 percentage points more willing to trust. Most of this effect (1.1 percentage points) is driven by the fact that the betrayer is the same as the person receiving the money. The remaining effect (0.6 percentage point) is driven by removing all betrayal. However, neither coefficient is significant. In other words, we find no evidence that removing betrayal aversion makes participants significantly more likely to trust. The third coefficient tells us that removing all risk as well as betrayal aversion increases the tendency to trust by 10.3 percentage points. The net effect of removing risk is therefore 8.6pp ($p < 0.0001$, Wald test). The difference between the fourth and second coefficient reveals that the net effect of removing the receiver is 3.5pp ($p < 0.0001$, Wald test). Hence, distributional preferences matter as well, here consistent with an aversion towards disadvantageous inequality. Finally, the comparison between the last two coefficients tells us that removing betrayal only increases participants' willingness to trust by 0.5pp when the receiver obtains no payment; this effect is not significantly different from zero ($p = 0.435$, Wald test).[23]

When averaging across all possible compositions of the urn, though, we also take choices into account where nearly no-one chooses the trust-equivalent action. Indeed, when the fraction of green balls is small, rational participants should only choose the trust-equivalent action if they are risk-seeking or strongly altruistic towards another player who is already better off, because the expected value of the trust-equivalent action is below the sure payoff of the alternative action. Including these choices therefore reduces the average effect of excluding the impact of the different factors.

We therefore also consider the average effect on the choices when the number of green balls exceeds 50% (i.e., when the urn has 10 or more green balls in a 20-person session), see column (3) of Table 1. We see that in the range where the trust equivalent action has a greater expected payment than the alternative, the net effect of removing risk becomes substantially larger (16.6pp), as does the net effect of removing distributional preferences (5.5pp). Furthermore, removing betrayal aversion now also has a significant effect, both when we only exclude betrayal by the beneficiary (3.9pp) and when we remove all betrayal (5.4pp). For completeness, column (2) also contains the estimates for fewer than 50% green balls, where, as expected, the effects are much smaller and not significant at the 5% level.

To understand these results intuitively, it is useful to go back to Figure 2, and focus on the parts of the Figure where the differences between choice lists are largest. The comparison between 'No Betrayal' and 'Distribution Urn' tells us that the impact of risk preferences is largest when the expected payoff is just above the certain payoff (50-70% green balls). By contrast, there is little impact of risk preferences when the expected payoff is much higher than the certain payoff (more than 80% green balls), which is in line with many participants being somewhat risk averse, but no-one being extremely risk averse. Similarly, the comparison

[23] In Appendix B, we redo this analysis using different sample restrictions with similar results. The main difference is that the point estimate for betrayal aversion is significant in less restrictive samples, though further analysis suggests that this estimate is likely to be spurious.

between ‘No Betrayal’ and ‘No Receiver Random Urn’ reveals that distributional preferences appear to matter most in a similar part of the distribution. This suggests that a non-trivial share of participants are somewhat averse towards disadvantageous inequality, but very few are highly inequality averse. Removing betrayal aversion matters most when the fraction of green balls is high (above 0.75). This suggests that betrayal aversion can prevent people from trusting even if trust is very likely to pay off, meaning that it is a strong motive for those affected by it. At the same time, we observe that betrayal aversion has little impact when trust just pays in expectation. This observation suggests that the betrayal-averse participants are also risk averse or inequality averse. Therefore, even though the betrayal motive likely also matters when the fraction of green balls is close to 50%, removing its effect has no impact because the participants still do not trust because of their risk aversion or inequality aversion.

These results suggest that beliefs about the rate of reciprocity (the fraction of green balls) and preferences interact in non-trivial ways. For different beliefs, different aspects of preferences matter most. We take a closer look at the relationship between beliefs and the decomposition of trust in what follows.

3.3 Decomposing Trust using Choice Lists

The previous section was focused on the components driving trust across the full range of possible beliefs about the rate of reciprocity. In this section, we identify the components driving trust for a specific, fixed belief about the rate of trustworthiness. We start by looking at participants’ choices corresponding to their actual belief as elicited in the experiment. We then also look at participants’ choices for an urn composition corresponding to the actual rate of trustworthiness (i.e., for ‘correct’ beliefs). We further decompose trust at several relevant counterfactual rates of trustworthiness. All of the results in this section are similar when we use different sample restrictions, see Appendix B for details.

3.3.1 Based on Elicited Beliefs

To decompose trust using price lists based on the elicited beliefs, we take for each participant the row in each choice list that corresponds to his or her elicited belief. For example, for a participant who believes that 8 other participants reciprocate, we take in each choice list the row where there are 8 green balls in the respective urn. This allows us to decompose trust conditional on how likely each participant considers that his or her trust is reciprocated. Note that since the complex belief elicitation task elicits an interval between k and $k-1$ green balls, we take the average of rows k and $k-1$ for beliefs elicited using this task, where participants who make a different choice in these rows are classified as indifferent.

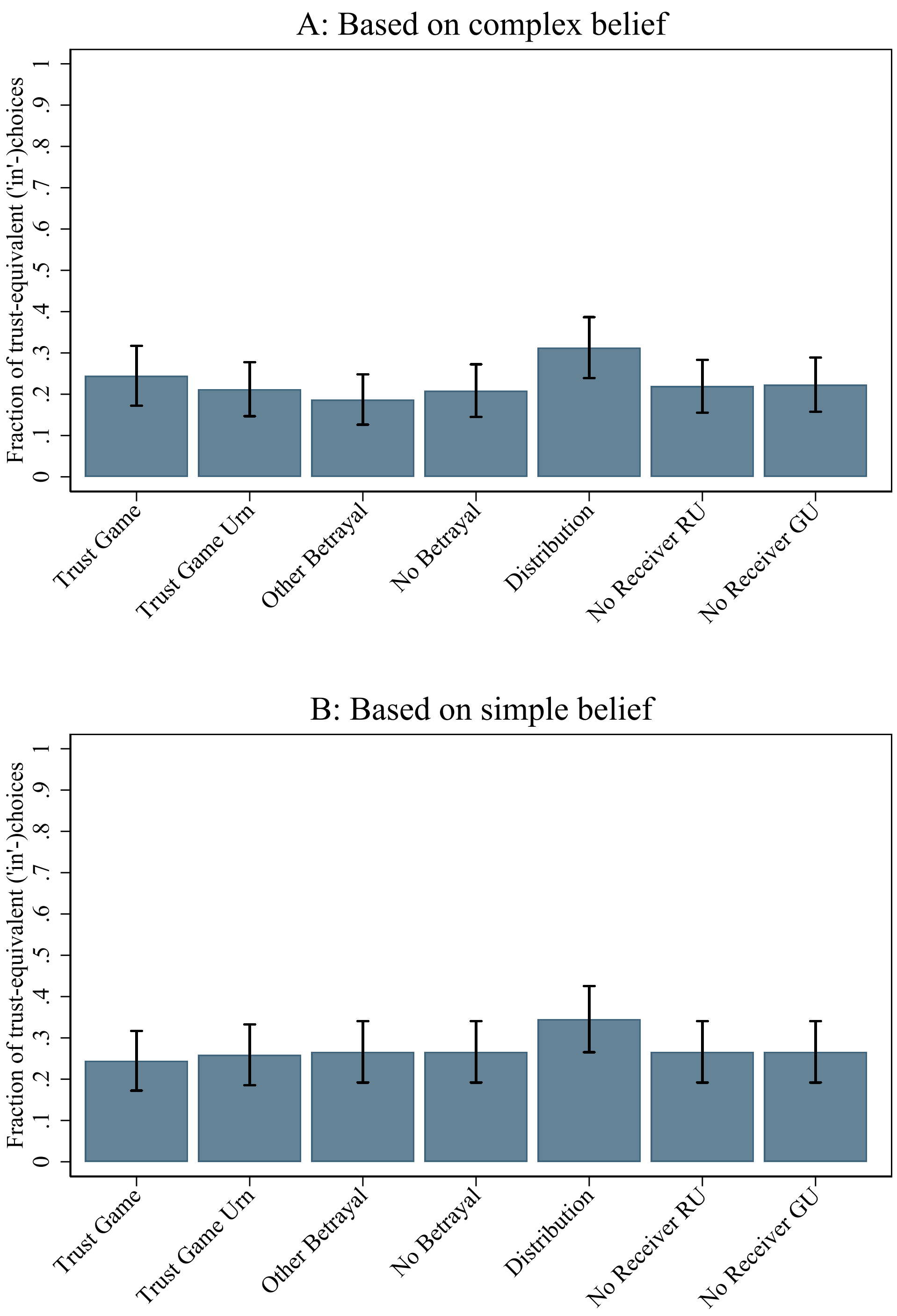
Panel A and B in Figure 3 present the fraction of trust-equivalent choices based on the complex and simple beliefs respectively. In the former case, participants who are indifferent in one task are classified based on the direction they switch to in the other task. In both cases, the comparison between the first two bars shows that average choices are nearly identical regardless of whether we use the actual trust game or the trust choices inferred from the Trust Game Urn (conditional on elicited beliefs). This implies that moving from a binary trust game to a price list does not affect the average tendency to choose the trust action in our experiment.

All other bars are similar as well, with the exception of the 'distribution' bar. Indeed, the comparison between the 'distribution' and the 'no betrayal' bars tells us that removing all risk makes participants 10.4pp more willing to trust based on the complex elicited beliefs (p = 0.001, t-test) and 7.9pp more willing to trust based on the simple elicited beliefs (p = 0.011). Table 2 presents the results of the other relevant pairwise comparisons; none are significant.[24] Note that this analysis keeps the role of beliefs constant; we will examine the role of beliefs in the next section.

Why do only risk preferences play a role in this analysis, when the analysis in the previous section demonstrated that the other components had some role to play as well? Recall that in the previous section, we looked at all 20 (or 16) choices made by participants. By contrast, in this section we look only at one of these choices. More to the point, most differences in Figure 2 only emerged in rows corresponding to optimistic beliefs (with the majority of other participants reciprocating). However, participants' actual beliefs are more pessimistic: 62% of participants (58% based on the simple elicitation) think that the reciprocation rate is less than 50%. This means that the majority of choices we examine here are based on the left side of Figure 2, where none of the components has a significant effect on behavior.

In practice, this implies that in our particular trust game, betrayal aversion and social preferences do not appear to contribute to the decision to trust. This is intuitive because most participants are fairly pessimistic about the likelihood of being reciprocated. As a result, they are likely to avoid the risk of trusting regardless of the role of betrayal aversion and distributional preferences. Only when the risk itself is eliminated (and participants are paid based on expected value) do participants increase their tendency to trust.

[24] Figure A1 in Appendix A illustrates these comparisons.



A: Based on complex belief

B: Based on simple belief

Figure 3: Choices based on Elicited Beliefs

Notes: These figures plot the fraction of ‘in’- (or trust-equivalent-)choices in the binary trust game and in the row in each price list that corresponds to the complex (panel A) and simple (panel B) elicited beliefs, respectively. “No Receiver RU” and “No Receiver GU” refer to No Receiver Random Urn and No Receiver Game Urn choices, respectively. The whiskers are 95%-confidence intervals.

Variable	Comparison	Same	In	Out	P-value
	A: Complex Beliefs				
Binary vs Urn	Trust Urn and Binary Trust	108	13	18	0.363
Other Betrayal	Other Betrayal and Trust Urn	123	6	10	0.162
Betrayal Aversion	No Betrayal and Trust Urn	115	11	13	0.874
Risk Preferences	Distribution and No Betrayal	103	28	8	0.001***
Distr. Preferences (1)	No Receiver Random Urn and No Betrayal	117	12	10	0.624
Distr. Preferences (2)	No Receiver Game Urn and Other Betrayal	121	13	5	0.068*
	B: Simple Beliefs				
Binary vs Urn	Trust Urn and Binary Trust	109	16	14	0.716
Other Betrayal	Other Betrayal and Trust Urn	130	5	4	0.740
Betrayal Aversion	No Betrayal and Trust Urn	130	5	4	0.740
Risk Preferences	Distribution and No Betrayal	120	15	4	0.011**
Distr. Preferences (1)	No Receiver Random Urn and No Betrayal	129	5	5	1.000
Distr. Preferences (2)	No Receiver Game Urn and Other Betrayal	131	4	4	1.000

Table 2: Frequency Table Based on Elicited Beliefs

Notes: This table presents the results of comparisons used to identify the importance of the respective components based on complex elicited beliefs (top panel) and simple elicited beliefs (lower panel) respectively. "Binary vs Urn" reveals the effect of going from a binary trust game to a trust urn. "Other Betrayal" is the effect of having the recipient's choice be determined by a third party. The remaining rows are the estimate effects of betrayal aversion, risk preferences and distributional preferences, respectively. Each comparison uses two choices; 'same' counts participants who made the same decision in both choices. 'In' counts participants who chose 'in' (that is, the trust-equivalent action) in the first listed choice and 'out' in the second choice. 'Out' counts participants who chose 'out' in the first listed choice and 'in' in the second choice. To deal with indifferent participants in the upper panel, 'In' also counts participants who chose 'in' (were indifferent) in the first listed choice but were indifferent (chose 'out') in the second listed choice. Similarly, 'Out' also counts participants who chose 'out' (were indifferent) in the first listed choice but were indifferent (chose 'in') in the second listed choice. The p-values in the last column are based on two-sided t-tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

3.3.2 Based on the True Rate of Reciprocity

Next, we decompose trust at the true rate of reciprocity. In particular, for each participant, we take the average of the choices in the two rows of each price list that most closely correspond to the actual rate of reciprocity in the experiment (39.4%). For sessions with 20 participants, these are rows 7 and 8 respectively.[25] Figure 4 panel A presents the results, where we add the binary trust game as a comparison group. Table 3 presents the frequencies and test results where participants who are indifferent in one task are classified based on the direction they switch to in the other task.

If participants held accurate beliefs, the rate of trust would decrease from 24.4% to 5.8% ($p < 0.001$, t-test). Intuitively, most participants who trust appear to do so because they believe that trust will be reciprocated. If they had instead expected the actual, much lower rate of trustworthiness, they would, according to their decisions in the respective row of the choice list, have decided not to trust. At the actual rate of trustworthiness, none of the components driving the decision to trust play a significant role anymore: almost no participant chooses the trust-equivalent action regardless of which component is able to play a role in their decision. This observation suggests that for rather pessimistic beliefs, the belief alone determines the decision whether or not to trust. If an individual is pessimistic enough, none of the other aspects of the trust decision can make up for the expectation of facing a selfish receiver.

For comparison, Panel B demonstrates the effect of each component if participants believed that only 2 participants in their session would fail to reciprocate (equivalent to them having very optimistic beliefs about reciprocity). The first thing to note is that such a high level of expected trustworthiness greatly increases participants' willingness to choose the trust-equivalent action as compared to choices in the binary trust game, which is intuitive. For a high rate of trustworthiness, the only components apart from beliefs that play a significant role are betrayal aversion and distributional preferences.[26] Note that these effects are visually apparent in Figure 2 for a fraction of approximately 90% green balls; this indeed is the comparison Panel B of Figure 4 captures.

3.3.3 The Role of Beliefs

Figures 3 and 4 further illustrate that beliefs about the probability of being reciprocated play a key role in at least two ways. First, they have a direct effect: participants are much more willing to take the trust-equivalent action if they believe it will be reciprocated. Second,

[25] Participants who switch between rows 7 and 8 are classified as indifferent.

[26] Figure A2 in Appendix A illustrates the treatment effects.

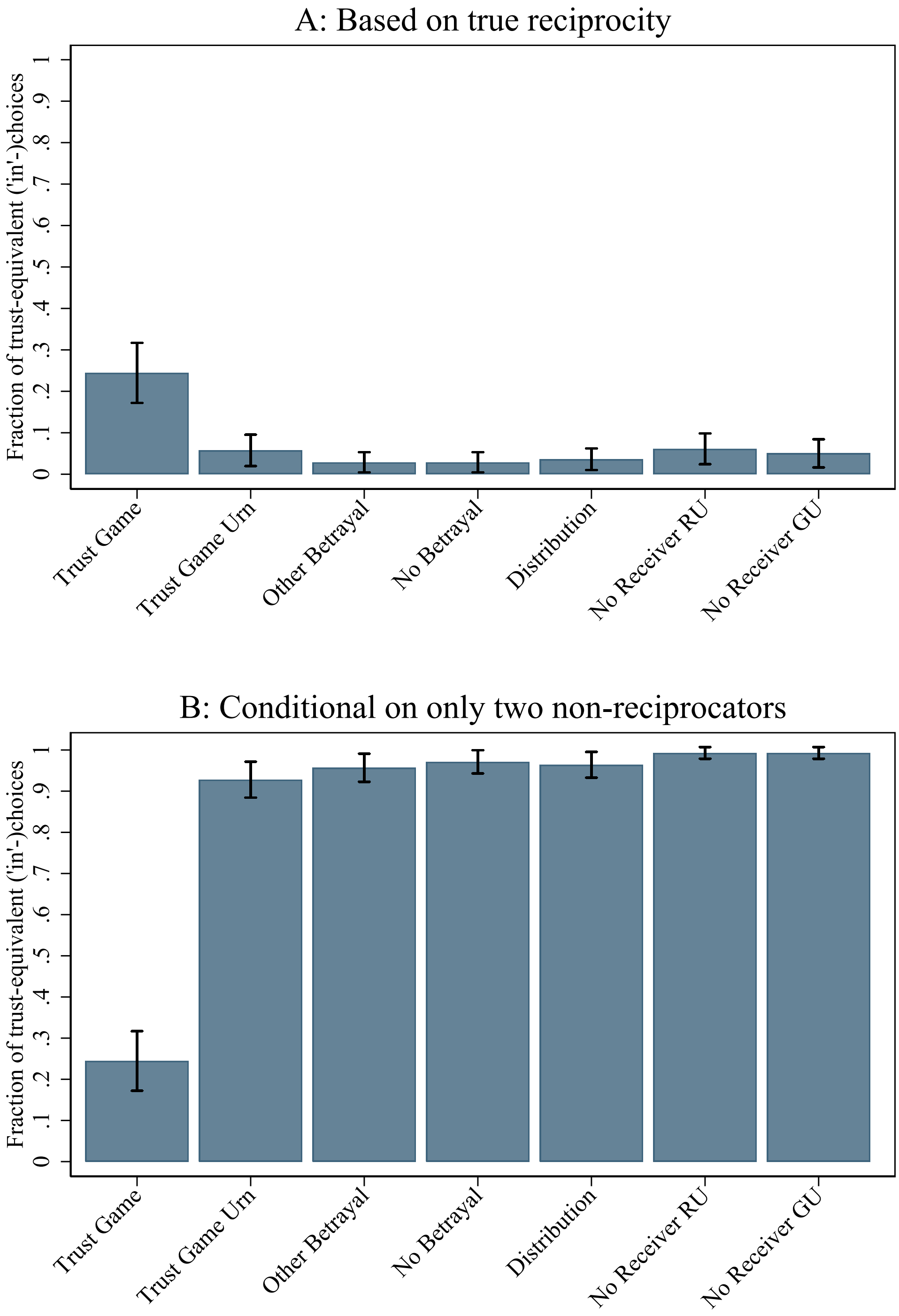


Figure 4: Choices based on True Reciprocity and High Reciprocity Rates

Notes: These figures plot the fraction of 'in'- (or trust-equivalent-)choices in the binary trust game and in the row in each price list that corresponds to the true rate of reciprocity in the experiment (panel A) and to the case where only two other participants fail to reciprocate (panel B), respectively. "No Receiver RU" and "No Receiver GU" refer to No Receiver Random Urn and No Receiver Game Urn choices, respectively. The whiskers are 95%-confidence intervals.

Variable	Comparison	Same	In	Out	P-value
	A: True Reciprocity				
Binary vs Urn	Trust Urn and Binary Trust	101	6	32	0.000***
Other Betrayal	Other Betrayal and Trust Urn	130	3	6	0.103
Betrayal Aversion	No Betrayal and Trust Urn	126	5	8	0.171
Risk Preferences	Distribution and No Betrayal	125	8	6	0.696
Distr. Preferences (1)	No Receiver Random Urn and No Betrayal	127	9	3	0.083*
Distr. Preferences (2)	No Receiver Game Urn and Other Betrayal	127	8	4	0.222
	B: 2 Non-Reciprocators				
Binary vs Urn	Trust Urn and Binary Trust	40	97	2	0.000***
Other Betrayal	Other Betrayal and Trust Urn	129	7	3	0.207
Betrayal Aversion	No Betrayal and Trust Urn	131	7	1	0.033**
Risk Preferences	Distribution and No Betrayal	132	3	4	0.707
Distr. Preferences (1)	No Receiver Random Urn and No Betrayal	136	3	0	0.083*
Distr. Preferences (2)	No Receiver Game Urn and Other Betrayal	134	5	0	0.025**

Table 3: Frequency Table Based on True Reciprocity

Notes: This table presents the results of comparisons used to identify the importance of the respective components based on the true rate of reciprocity (upper panel) and only 2 non-reciprocators (row 18 in sessions with 19 participants). For the definitions of variables and explanations of individual columns, we refer to the notes of Table 2. To deal with indifferent participants in the upper panel, 'In' also counts participants who chose 'in' (were indifferent) in the first listed choice but were indifferent (chose 'out') in the second listed choice. Similarly, 'Out' also counts participants who chose 'out' (were indifferent) in the first listed choice but were indifferent (chose 'in') in the second listed choice. The p-values in the last column are based on two-sided t-tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

they also have an indirect effect as beliefs affect the role played by other components for the decision to choose the trust-equivalent action. Notably, we see that removing the role of risk preferences increases trust in the row corresponding to a participant's elicited belief, but not in the row corresponding to the actual rate of reciprocity in the experiment. Furthermore, betrayal aversion matters if the majority of participants are expected to reciprocate but not at the true rate of reciprocity or for the participants' elicited beliefs.

To further understand the role played by beliefs, it is instructive to look at the row (or equivalently the belief) for which the point estimate for a specific component is the largest. For risk preferences, the largest effect (45.2pp) is observed in the 10th row (or equivalently the 8th row in the session with 16 participants), which is the first row in which the trust-equivalent action has greater expected value than the alternative (compare the No Betrayal Urn to the Distribution Urn when the fraction of green balls just exceeds 0.5 in Figure 2). For distributional preferences, the largest point estimate (14.4pp or 12.7pp depending on whether we use the No Receiver Game Urn or No Receiver Random Urn) is observed in row 11, whereas for betrayal aversion (14.4pp) it is in row 14. This tells us that distributional preferences and betrayal aversion are most likely to drive behavior in cases where participants are fairly optimistic but not too optimistic about the probability of being reciprocated.

4 Decomposing Trust Using Regressions

In this section, we implement the conventional approach of decomposing trust by regressing trust on elicited beliefs and on estimates of preferences that have been derived from additional tasks and questionnaire items. Based on a comparison of the regression results with those from the design-based analysis, we discuss the impact of measurement error and omitted-variable bias on the assessment of the drivers of trust. In this section, we focus on the role of beliefs, risk preferences, and distributional preferences. As we did not obtain separate measures for betrayal aversion, we cannot assess its impact using the regression approach.[27]

4.1 Beliefs about Reciprocity

Table 4 examines the role of beliefs. Columns 1 and 2 show that beliefs about trustworthiness elicited in the questionnaire are already quite predictive of behavior in the binary trust game. In particular, someone who strongly believes that others can be trusted is 31 to 38 percentage points more likely to trust than someone who strongly believes that others cannot

[27] A correlation matrix of the variables used in the regression tables in this section is contained in Appendix A. The regression results in this section are based on the restricted sample described in the previous section. Results for less restrictive samples are reported in Appendix B.

be trusted. When we move to the simple belief elicitation tailored to our experiment, the estimate increases to 62 percentage points, and when we use the complex elicitation belief instead, it increases to 97pp. The treatment comparison estimate (obtained by comparing the first to the last row of the Trust Game Urn) puts the estimate at 100pp.[28] Hence, in this case, the complex elicitation procedure does a better job at predicting choices than any of the simpler measures, and in fact leads to a very similar estimate as our treatment comparison approach. These results are consistent with the idea that the simpler measures (in particular those based on the questionnaire) do not fully capture the beliefs that govern choices in our trust game. This is a form of measurement error that implies that controlling for these variables in a regression understates the true effect of beliefs. We also note that linear regressions of this kind are, by definition, unable to capture the non-linear effect of beliefs we observed in our choice list data, where we saw that changes in beliefs had a much smaller effect on behavior below the 50% threshold than above.

The two approaches also differ in the estimated effect of de-biasing beliefs.[29] Using the regression results to predict the rate of trust if participants had expected the true rate of trustworthiness, we would conclude that correcting beliefs would have almost no effect on observed trust because the true rate of trustworthiness is nearly identical to the average estimated rate of trustworthiness. However, our treatment comparisons reveal that trust would amount to only 5.8% if beliefs corresponded to the actual rate of trustworthiness, much less than the 24.4% in the simple trust game where individuals decide based on the (potentially biased) belief that they actually hold. This discrepancy stems from the fact that the treatment comparison allows us to observe and compare behavior at the individual level, whereas the regression approach relies on predictions based on averages and fitted regression estimates. Specifically, the treatment comparison reveals that participants who are more pessimistic would mostly still not trust if they had correct expectations, whereas those who are more optimistic often decide to trust, but would not do so if they held correct expectations about actual trustworthiness.

4.2 Risk Preferences

Table 5 analyses the impact of risk preferences. All coefficients are scaled to represent a shift from average risk preferences in the population to risk neutrality. Columns 1 and 2 use the elicited measures of risk preferences from the auxiliary lottery choice lists to predict

[28] The last result relates to our exclusion criteria: We exclude any participant who chose an allocation of (10,10) over (15,15). Further, none of the consistent participants choose (5,20) over (10,10).

[29] The comparison of behavior for elicited beliefs and at the actual rate of reciprocity is the test we used as an example in the pre-analysis plan.

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Belief (general)	0.381*** (0.141)				
Belief (today)		0.312** (0.136)			
Belief (simple elicitation)			0.621*** (0.124)		
Belief (complex elicitation)				0.977*** (0.142)	
Treatment Estimate					1.000*** (0.000)
Constant	0.068 (0.066)	0.053 (0.082)	-0.023 (0.043)	-0.147*** (0.042)	0.000 (0.000)
Observations	139	139	139	139	278
Clusters	139	139	139	139	139

Table 4: Identifying the Belief Effect using Regressions

Notes: Regression estimates (cluster-robust standard errors in parentheses). The dependent variable in the first four columns is a binary decision (trust-1). “Belief (general)” asks whether people can be trusted in general (0-fully disagree, 1-fully agree). “Belief (today)” asks whether people can be relied upon today (0-fully disagree, 1-fully agree). The other two belief variables are the elicited beliefs from the experiment, scaled to range from 0 to 1. The fifth column presents the treatment effect of going from zero reciprocity to full reciprocity in the Trust Game Urn task. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Risk Measure (no losses)	0.016*				
	(0.009)				
Risk Measure (with losses)		0.001			
		(0.068)			
Self-Assessment			0.000		
			(0.000)		
No Receiver Random Urn				-0.008	
				(0.026)	
Treatment Estimate					0.104***
					(0.031)
Constant	0.245***	0.245***	0.245***	0.245***	0.209***
	(0.036)	(0.037)	(0.037)	(0.037)	(0.032)
Observations	139	139	139	139	278
Clusters	139	139	139	139	139

Table 5: Identifying the Risk Effect using Regressions

Notes: Regression estimates (cluster-robust standard errors in parentheses). The dependent variable in the first four columns is a binary decision (trust-1). “Risk Measure” is the number of times the participant chose the lottery in the lottery risk measure. “Self-Assessment” is participant’s subjective assessment of their own willingness to take risk. “No Receiver Random Urn” is the number of times a participant chose the lottery option in the risky price list without a receiver. For each of these variables, the coefficient estimate is scaled to represent a shift from average risk preferences in the population to risk neutrality. The fifth column presents the treatment effect of removing all risk by comparing the No Betrayal Urn and Distribution Urn tasks. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

trust. The results imply that removing the role of risk (moving average risk preferences to risk neutrality) would increase trust by 1.6pp (measure without losses) and 0.1pp (measure with losses), respectively. Columns 3 and 4 present the estimates based on the qualitative questionnaire measure and the average choice in the No Receiver Random Urn choice list, which is intended to capture risk preferences in the trust game by reducing the problem to a lottery where all aspects other than risk are eliminated but that has the same possible payoffs as the trust game.[30] Neither variable significantly predicts trust. Finally, column 5 shows our treatment comparison estimate from the previous section based on the complex belief elicitation task.

Several things are worth noting. First, all four regression estimates understate the importance of risk preferences compared to our treatment estimate. Second, contrary to our expectations, the specifically tailored measure from the experiment does worse than the three more general measures. Intuitively, this might be because this measure takes into account risk preferences across the full distribution of beliefs. However, in reality, the only risk preferences that matter are the ones corresponding to someone's actual belief about reciprocity. Put differently, none of our measures appear to capture the risk preferences that govern trust choices in the experiment very well.

4.3 Distributional Preferences

Table 6 presents the regression-based estimates for the impact of distributional preferences on trust decisions. All coefficients are scaled to represent a shift from average distributional preferences in the population to purely selfish preferences. The first three columns represent regression estimates using the modified dictator game, the self-reported questionnaire altruism measure, and distributional preferences elicited using the Distribution Urn respectively.[31] All three estimates are significant, in contrast to the treatment estimate (column 4). Columns 5 and 6 show that the importance of the distributional preference variable is considerably reduced once beliefs are controlled for. This is particularly true for the dictator game effect (compare column 5 to column 1). Interestingly, the measure of distributional preferences obtained from the Distribution Urn task, which should capture the relevant distributional

[30] For the qualitative measure we assume that a score of 5 implies risk neutrality.

[31] Intuitively, the dictator game with disadvantageous inequality should be more informative than the dictator game with advantageous inequality because aversion to disadvantageous inequality would be the distributional motivation to keep first movers from trusting. The version with disadvantageous inequality, however, has very little variation in our data and hence almost no predictive power, because more than 80% of the participants are maximizing their payoff, not willing to pay a positive amount for increasing or decreasing the other player's payoff. We therefore only show the results for the advantageous version here. Including behavior from both dictator games in the regression does not change the result for the advantageous version and finds a coefficient of 0.007 ($p = 0.823$) for the disadvantageous version.

	(1)	(2)	(3)	(4)	(5)	(6)
		Dependent Variable: Trust-Equivalent Action				
Dictator Game	-0.145***				-0.075*	
	(0.040)				(0.042)	
Altruism		-0.259**				-0.161
		(0.113)				(0.106)
Distribution Urn			-0.033**			
			(0.013)			
Treatment estimate				0.011		
				(0.022)		
Belief (complex elicitation)					0.843***	0.946***
					(0.168)	(0.143)
Constant	0.245***	0.245***	0.245***	0.209***	-0.094	-0.135***
	(0.035)	(0.036)	(0.036)	(0.037)	(0.058)	(0.044)
Observations	139	139	139	278	139	139
Clusters	139	139	139	139	139	139

Table 6: Identifying the Social Preference Effect using Regressions

Notes: Regression estimates (cluster-robust standard errors in parentheses). The dependent variable columns 1, 2, 4 and 5 is a binary decision (trust-1). “Dictator Game” is the number of times the participant chose the equal split in the dictator game task (advantageous version). “Altruism” is the degree to which participants thought that the well-being of others is important. “Distribution Urn” is the number of times a participant chose the generous option in the Distribution Urn. The third column presents the treatment effect of removing the receiver by comparing the No Betrayal Urn and No Receiver Game Urn tasks. “Belief (complex)” is the elicited belief. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

preferences best because it reflects the same possible allocations as the trust game, shows a weaker effect of distributional preferences than the measures derived from the dictator game and the self-assessment. Given that we find only a very small effect based on our treatment comparison, this tailor-made measure of distributional preferences arguably does better than the other measures.

Why do the regression estimates imply a larger importance of distributional preferences than our treatment comparison, where we saw the opposite pattern with respect to risk in the previous section? In this case, the main independent variables (in particular the dictator game variable) are positively correlated with elicited beliefs (see Table A1 in the Appendix). Participants exhibit a consensus effect in the sense that those who show pro-social preferences in the dictator game expect others to be more pro-social in the trust game.[32] This means that controlling for only beliefs or distributional preferences individually may lead to an overestimate of their respective effects due to omitted-variable bias. Indeed, in this particular case, after controlling for beliefs, the effect of distributional preferences is reduced in size and no longer significant at the 5% level.

By contrast, our treatment comparison approach exogenously varies the probability of reciprocation along the rows of the price list. This allows us to fix (i.e., control experimentally for) the effect of beliefs by design. In this setting, distributional preferences do not appear to explain choices in the trust game. We are unable to draw similar conclusions using the regression approach without making further assumptions about the underlying correlation between the belief and social preference measure (e.g., that beliefs 'cause' social preferences or vice versa).

4.4 Discussion

In this section, we discuss the implications of our findings for the methodology of decomposing trust decisions into their various potential causes. Overall, the results illustrate that the regression approach may both over- and underestimate the importance of a given component, depending on whether measurement error or omitted-variable bias dominates. In particular, distributional preferences and beliefs are highly positively correlated. This means that controlling for only one of these variables (e.g., distributional preferences) may lead us to spuriously assign too much explanatory power to this variable, whereas behavior is actually driven by the correlated component (e.g., beliefs). In this case, the omitted-variable bias dominates. However, when a component (risk preferences) is not correlated with other

[32] Consistent with this observation, dictator game choices ($r = .59$, $p < 0.0001$) and the altruism variable ($r = .21$, $p = 0.014$) are also positively correlated with reciprocity as a second-mover in the binary trust game.

components, measurement error dominates, leading us to underestimate the importance of risk preferences relative to the treatment comparison estimates.

It is also worth pointing out that the direction of the overall bias for any given regression is difficult to predict without knowing whether and how explanatory variables are correlated with each other. If they are not correlated, measurement error bias will dominate and each individual factor's effect will be underestimated. However, if two variables are positively correlated, controlling for only one of those variables generates an upward bias in this variable's estimated effect. The net effect is then ambiguous.

The obvious way to attempt to address the omitted-variable bias problem in the regression-based approach is to measure as many potentially relevant factors as one can assess and include them all in the regression. However, this is more problematic than it seems. The regression model makes assumptions on the functional form of the impact of the individual factors. If we simply include all potential factors, for instance, an OLS regression would assume they all enter linearly. Our results suggest that this would miss important interaction effects between beliefs and preferences. Including, in addition to all potential factors, a list of potentially relevant interaction effects, still makes restrictive assumptions but also likely leads to over-fitting the data.

An omitted variable problem specifically arises if beliefs about others' behavior are correlated with one's own inclination for this behavior due to a (false) consensus effect.[33] If first movers in a trust game exhibit a consensus effect and we elicit their social preferences but do not take their beliefs into account, we will overestimate the relevance of their pro-sociality for their trust decision.

Incorrect attribution of trusting decisions to preferences if the correlation with beliefs is ignored has been documented by Blanco et al. (2014). They find a strong correlation between first- and second-mover behavior in a binary trust game where participants play both roles. This could be interpreted as a strong correlation between preferences for trustworthiness and trust and hence trust being determined by pro-social preferences. Eliciting beliefs, however, they find that own trustworthiness is strongly correlated with the expectation of trustworthiness in line with a consensus effect and that controlling for beliefs, trust behavior is not correlated with own trustworthiness anymore.

[33] The term was introduced into the psychology literature by Ross, Greene, and House (1977). Dawes (1989) argued that such a correlation does not have to be a bias in a Bayesian sense, because information about one's own type can be considered useful information. Engelmann and Strobel (2000) and Engelmann and Strobel (2012) show that when information about other experimental participants is transparent, participants do not give excessive weight to their own type and hence the consensus effect is not false from a Bayesian perspective, but when information about others is only implicitly provided, a truly false consensus effect can be observed. In our data, there is a strong consensus effect. Trustworthiness as second mover correlates strongly with both simple beliefs ($r = .50$, $p < 0.0001$) and complex beliefs ($r = .61$, $p < 0.0001$).

The results by Blanco et al. (2014), however, also illustrate that the regression approach can be subject to intricate problems of measurement error and that a design-based approach can paint a clearer picture. Consider the finding that when regressing trust on own trustworthiness and expected trustworthiness, the former has no explanatory power. The intuitive interpretation is that any observed correlation between trustor and trustee behavior is entirely driven by a consensus effect and rational trustor behavior given beliefs. In a further treatment, however, Blanco et al. (2014) inform trustors about the distribution of trustee choices before they make their choice. Being informed about the true rate of trustworthiness renders beliefs irrelevant. Regressing trustor choices on the true rate of trustworthiness and their own trustworthiness yields a marginally significant effect of own trustworthiness. This design-based approach therefore reveals that the interaction of beliefs and preferences is more complicated than a linear relationship.[34]

It is also worth noting that the point estimates in the regression approach depend on the specific measures that are used. For example, the effect of distributional preferences varies from 3 to 26 percentage points and the effect of beliefs varies from 31 to 97 percentage points of the respective estimated effects based on treatment comparisons (see Table 6 and 4, respectively). This suggests that regression estimates are likely to be sensitive to the specific measure used as a control variable (as previously highlighted by Gillen, Snowberg, and Yariv, 2019). In addition, in contrast to what we expected and stated in our pre-analysis plan, we do not find that measures specifically tailored to the current experiment consistently outperform more general measures in the regression approach. Therefore, having measures that are specifically tailored to an experimental design is no guarantee for obtaining a better estimate.

Our results on the interaction of beliefs and preferences help us to reconcile earlier, seemingly contradictory results in the literature. When assessing the relative importance of different determinants of trust without controlling for beliefs about trustworthiness, as is common practice in the literature,[35] differences in subject pools, trust game payoff configurations, and methods of eliciting trust behavior across studies are likely to matter because they likely affect subjects' beliefs about trustworthiness. While only a few studies elicit beliefs, we can

[34] There are two likely explanations for this result. First, the measure for trustworthiness is only binary whereas the beliefs are measured on a finer scale so that the beliefs are a more precise measure of the preferences for cooperation than own trustworthiness choices. Second, the consensus effect also precludes observing trustors who due to their strong preference for cooperation would also cooperate if they had a pessimistic belief. These participants simply do not have a pessimistic belief. In the treatment with information about the true rate of trustworthiness, however, they can be exposed to a pessimistic reality and act against the low odds. This is in line with our results that certain preferences matter for certain ranges of beliefs only, something that is easily missed in a regression-based approach.

[35] Eckel and Wilson (2004) and Ashraf, Bohnet, and Piankov (2006) are exceptions.

take the usual approach to assume that average beliefs are roughly correct (as they are in this study).[36] Consistent with our results, Karlan (2005) finds that play in a single trust game where trusting does not pay off in expectation (i.e., trustworthiness rates are relatively low), is uncorrelated with survey questions on social attitudes, but correlates to risk preferences. In contrast, Schechter (2007) find some role of social preferences in addition to risk preferences in a trust game where the trustworthiness rate is relatively high in their study. This is consistent with our result that social preferences only play a role when the trustworthiness rate is relatively high.

Further studies that find a relevant role for both risk and social preferences typically rely on a variant of a choice list or survey question to measure trust. Therefore, they assess trust across a broader range of beliefs, so that according to our results, several aspects matter, including social preferences.[37] Bohnet et al. (2008) elicit minimal acceptable trustworthiness rates and compare these to minimal acceptable winning rates in an equivalent lottery and find that both risk preferences and betrayal aversion matter. By eliciting minimal acceptable trustworthiness rates, they elicit choices in the range of beliefs where betrayal aversion matters by construction, even if their participants do not have sufficiently optimistic beliefs.[38] Relatedly, investigating the cooperation gap observed between the North and the South of Italy, Bigoni et al. (2019) find evidence that preferences for conditional cooperation are similar in subjects from the North and the South. But they show that beliefs are much more optimistic about trustworthiness of people from the North than from the South. While the authors also report suggestive evidence for differences in betrayal aversion between Northerners and Southerners based on Bohnet et al. (2008)'s elicitation method, they interpret their findings in line with our results as beliefs being the key determinant of cooperative behavior.

Furthermore, beliefs may also differ across treatments within a study and hence exaggerate or undermine the effects of treatment variables. For instance, Brülhart and Usunier (2012) describe results from an experiment where trust game transfers do not differ significantly in size when the endowment of the trustee is changed. While the authors interpret their results as evidence against altruism being an important determinant of trustor behavior, our

[36] Evidence that expectations in a trust game are well calibrated is provided by Ashraf, Bohnet, and Piankov (2006).

[37] For example, Fehr (2009) finds that survey measures on trust are correlated both with risk and social preferences.

[38] Specifically, our data suggest that betrayal aversion matters only for fairly optimistic beliefs. Even if betrayal-averse participants do not have such optimistic beliefs, such that their betrayal aversion would not matter for their choice in a direct play of the trust game, they would provide evidence of betrayal aversion in the design by Bohnet et al. (2008) because they would require an even higher reciprocity rate in the trust game in order to trust than to play the equivalent lottery. The design by Bohnet et al. (2008) is therefore extremely efficient in finding an effect of betrayal aversion because it does not require the participants to hold a belief in the range where it actually matters.

results suggest that the finding should be interpreted more cautiously as the analysis does not address the role of beliefs. Beliefs about trustworthiness might change with the trustee's endowment. Specifically, it is plausible to expect higher returns from trustees with a larger endowment. The observed null result in Brülhart and Usunier (2012) would therefore be consistent with two countervailing effects being at work: trustors might be inclined to give more to poorer trustees out of distributional concerns but at the same time be more optimistic about the returned amount to be expected from richer trustees.

5 Conclusion

The purpose of our paper is to decompose trust, studying the way in which beliefs about trustworthiness, distributional preferences, betrayal aversion and risk preferences affect trust behavior. Our main results are that (1) beliefs about trustworthiness are a key driver of trust in our sample, (2) the effect of risk preferences, betrayal aversion and distributional preferences greatly depends on beliefs and (3) on average, risk preferences are more important than distributional preferences, which are more important than betrayal version. Earlier experiments showed mixed evidence on which factors are important. Our results suggest that the relevance of a factor crucially depends on beliefs, a finding that helps organize these mixed results. If participants in some studies were more optimistic than in others, then our results suggest that it is plausible that different factors are important in the different studies. When beliefs are in the area where trust just pays off in expectation, risk preferences are likely to play a major role. When participants are more optimistic, so that trusting appears to be a relatively profitable gamble for a selfish person, social preferences may become more important. The same holds for designs that assess the willingness to trust for a broader range of beliefs, for example by asking for minimum levels of trustworthiness.

In line with our results, Sapienza, Toldra-Simats, and Zingales (2013) already pointed out that beliefs and preferences jointly determine behavior in the trust game with a large impact of beliefs, whereas the trust question used in such surveys as the World Values Survey captures mostly beliefs. However, as our study shows, beliefs and preferences cannot be treated as independent in determining trust. Rather, there are complex interactions between the two with different beliefs activating different preferences. As noted above, these interactions between beliefs and preferences help organize some apparently puzzling results in the literature. Further, our evidence on the non-linear interaction between beliefs and preferences suggests that studies analyzing the determinants of trust need to assess beliefs to allow for reliable conclusions.

As a methodological contribution, being able to cleanly detect the important non-linear interactions between beliefs and preferences is an advantage of our approach that can be applied to settings beyond trust. Using regression-based methods, studying the interactions between beliefs and preferences is likely only possible in a structural approach that requires strong assumptions. Comparing our approach with the regression-based approach further highlights that measurement error typically leads to underestimating the relevance of a factor, while in combination with omitted variable bias the impact of a factor can also be overestimated. Our approach is therefore particularly useful for decomposing complex choices if measures from separate tasks are likely subject to strong measurement error, as is typically the case for risk aversion, partly because risk aversion does not translate well across domains. It is also useful when several factors are strongly correlated as is often the case for a preference and the expectation regarding this preference amongst others; distributional preferences in the trust game are an important example. The same, however, likely holds in other settings where preferences and beliefs are important drivers of choices. Examples include market entry, where managers who are willing to enter a competition likely overestimate how likely others are willing to compete; or collusion, where preferences for conditional cooperation, risk preferences, and beliefs about the choices of other players interact in determining one's choice.

Our approach can also be used to identify the factors driving differences by gender and other demographics. For gender, for example, we could have decomposed a potential gender difference in trust-equivalent actions along the same lines of how we decomposed trust. Previous research suggests that we would find a gender difference in the decision to trust with men being more trusting (see the meta-analysis by Van Den Akker et al., 2020). However, we found no evidence of a difference in trust based on gender or any other demographics we collected (field of study and age). Instead, we refer to work by Van Veldhuizen (2022) as an illustration of our method in the context of gender differences in tournament entry.

References

Alós-Ferrer, Carlos and Federica Farolfi. 2019. "Trust games and beyond." *Frontiers in neuroscience* 13:887.

Andersen, Steffen, John Fountain, Glenn Harrison, and E. Rutström. 2014. "Estimating subjective probabilities." *Journal of Risk and Uncertainty* 48 (3):207–229.

Antoniou, Constantinos, Glenn Harrison, Morten Lau, and Daniel Read. 2015. "Subjective Bayesian beliefs." *Journal of Risk and Uncertainty* 50 (1):35–54.

Ashraf, Nava, Iris Bohnet, and Nikita Piankov. 2006. "Decomposing trust and trustworthiness." *Experimental Economics* 9 (3):193–208.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1):122–142.

Bigoni, Maria, Stefania Bortolotti, Marco Casari, and Diego Gambetta. 2019. "At the root of the North–South cooperation gap in Italy: Preferences or beliefs?" *The Economic Journal* 129 (619):1139–1152.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Normann. 2014. "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis." *Games and Economic Behavior* 87:122–135.

Blanco, Mariana, Dirk Engelmann, and Hans Theo Normann. 2011. "A within-subject analysis of other-regarding preferences." *Games and Economic Behavior* 72 (2):321–338.

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review* 98 (1):294–310.

Brülhart, Marius and Jean-Claude Usunier. 2012. "Does the trust game measure trust?" *Economics Letters* 115 (1):20–23.

Chetty, Rinelle, Andre Hofmeyr, Harold Kincaid, and Brian Monroe. 2021. "The trust game does not (only) measure trust: The risk-trust confound revisited." *Journal of Behavioral and Experimental Economics* 90:101520.

Clark, Kenneth and Martin Sefton. 2001. "The Sequential Prisoner's Dilemma: Evidence on Reciprocation." *The Economic Journal* 111 (468):51–68.

Cox, James C. 2004. "How to identify trust and reciprocity." *Games and Economic behavior* 46 (2):260–281.

Crosetto, Paolo and Antonio Filippin. 2016. "A theoretical and experimental appraisal of four risk elicitation methods." *Experimental Economics* 19 (3):613–641.

Dawes, Robyn M. 1989. "Statistical criteria for establishing a truly false consensus effect." *Journal of Experimental Social Psychology* 25 (1):1–17.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior* 47 (2):268–298.

Eckel, Catherine C and Rick K Wilson. 2004. “Is trust a risky decision?” *Journal of Economic Behavior & Organization* 55 (4):447–465.

Engelmann, Dirk and Martin Strobel. 2000. “The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given.” *Experimental Economics* 3 (3):241–260.

———. 2012. “Deconstruction and reconstruction of an anomaly.” *Games and Economic Behavior* 76 (2):678–689.

Fairley, Kim, Alan Sanfey, Jana Vyrastekova, and Utz Weitzel. 2016. “Trust and risk revisited.” *Journal of Economic Psychology* 57:74–85.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. “Global Evidence on Economic Preferences.” *The Quarterly Journal of Economics* 133 (4):1645–1692.

Fehr, Ernst. 2009. “On The Economics and Biology of Trust.” *Journal of the European Economic Association* 7 (2-3):235–266.

Fischbacher, Urs. 2007. “z-Tree: Zurich toolbox for ready-made economic experiments.” *Experimental Economics* 10 (2):171–178.

Garapin, Alexis, Laurent Muller, and Bilel Rahali. 2015. “Does Trust Mean Giving and not Risking? Experimental Evidence from the Trust Game.” *Revue D’Économie Politique* 125 (5):701–716.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study.” *Journal of Political Economy* 127 (4):1826–1863.

Glaeser, Edward L, David I Laibson, Jose A Scheinkman, and Christine L Soutter. 2000. “Measuring trust.” *The Quarterly Journal of Economics* 115 (3):811–846.

Greiner, Ben. 2015. “Subject pool recruitment procedures: organizing experiments with ORSEE.” *Journal of the Economic Science Association* 1 (1):114–125.

Hausman, Jerry. 2001. “Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left.” *Journal of Economic Perspectives* 15 (4):57–67.

Houser, Daniel, Daniel Schunk, and Joachim Winter. 2010. “Distinguishing trust from risk: An anatomy of the investment game.” *Journal of Economic Behavior & Organization* 74 (1-2):72–81.

Karlan, Dean S. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review* 95 (5):1688–1699.

Karni, Edi. 2009. "A Mechanism for Eliciting Probabilities." *Econometrica* 77 (2):603–606.

LaPorta, Rafael, Florencio Lopez de Silanes, Andrei Shleifer, and Robert W. Vishny. 1997. "Trust in Large Organizations." *American Economic Review Papers and Proceedings* 87 (2):333–338.

Li, Chen, Uyanga Turmunkh, and Peter P Wakker. 2019. "Trust as a decision under ambiguity." *Experimental Economics* 22 (1):51–75.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat. 2022. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science* 68 (11):7793–8514.

Pedroni, Andreas, Renato Frey, Adrian Bruhin, Gilles Dutilh, Ralph Hertwig, and Jörg Rieskamp. 2017. "The risk elicitation puzzle." *Nature Human Behaviour* 1 (11):803–809.

Polipciuc, Maria. 2022. "Group Identity and Betrayal: Decomposing Trust." Discussion Paper. Available at SSRN: https://ssrn.com/abstract=4264157 or http://dx.doi.org/10.2139/ssrn.4264157.

Ross, Lee, David Greene, and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." *Journal of Experimental Social Psychology* 13 (3):279–301.

Sapienza, Paola, Anna Toldra-Simats, and Luigi Zingales. 2013. "Understanding trust." *The Economic Journal* 123 (573):1313–1332.

Schechter, Laura. 2007. "Traditional trust measurement and the risk confound: An experiment in rural Paraguay." *Journal of Economic Behavior & Organization* 62 (2):272–292.

Van Den Akker, Olmo R, Marcel ALM van Assen, Mark Van Vugt, and Jelte M Wicherts. 2020. "Sex differences in trust and trustworthiness: A meta-analysis of the trust game and the gift-exchange game." *Journal of Economic Psychology* :102329.

Van Veldhuizen, Roel. 2022. "Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness?" *Journal of the European Economic Association* 20 (4):1595–1618.

Ziegler, Andreas. 2021. "New Ecological Paradigm meets behavioral economics: On the relationship between environmental values and economic preferences." Tech. rep.

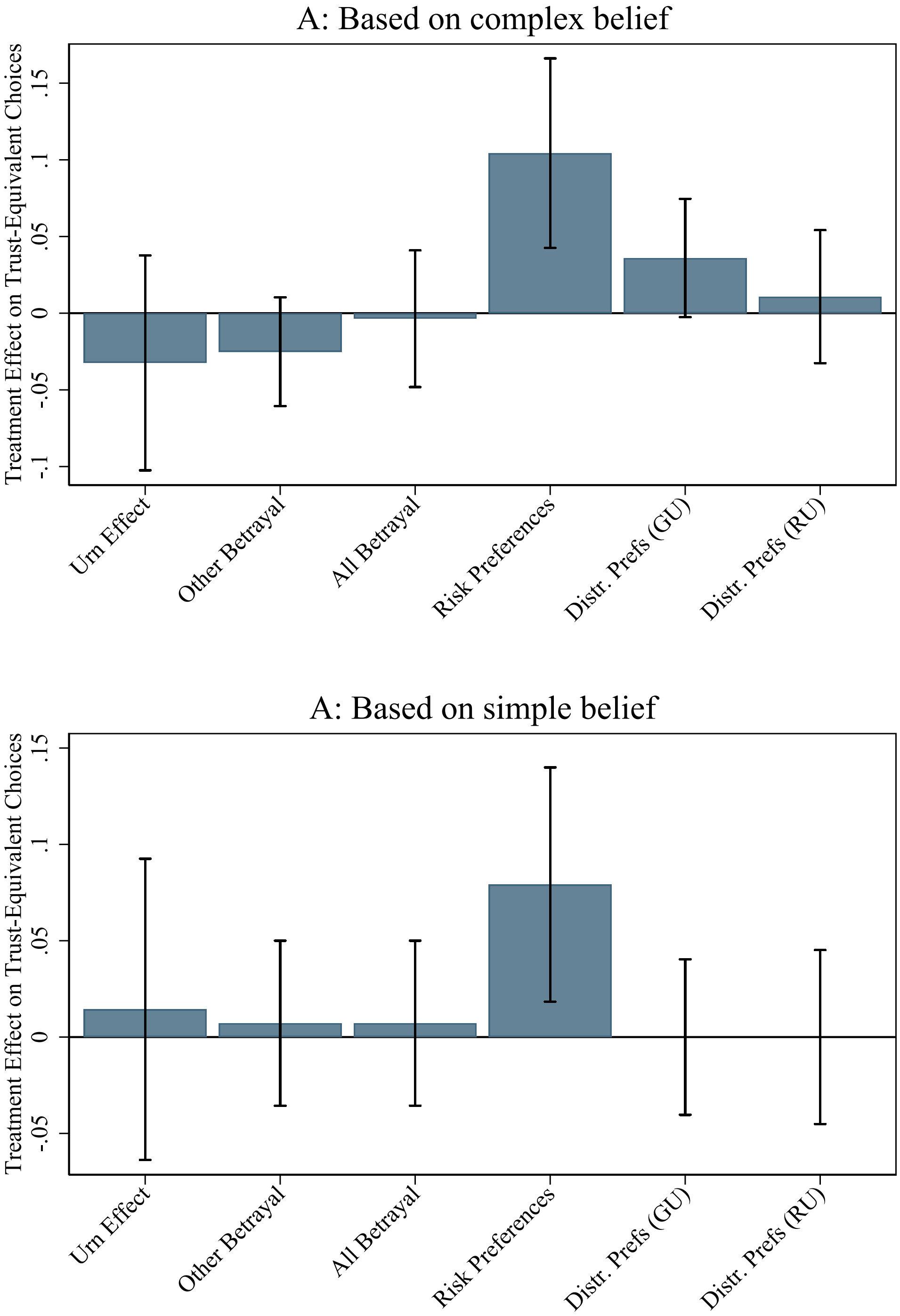
A Additional Tables and Figures

Table A1 presents correlations between the variables meant to capture distinct components and used in our regression tables. The first four variables capture the belief component, variables five to seven capture distributional preferences and the remaining variables are meant to capture risk preferences. Within each component, the relevant variables are mostly positively correlated with each other, as expected. When comparing across components, the main correlations that stand out are the positive correlations between the variables that capture beliefs and those that capture social preferences. Eight out of the twelve coefficients are positive and significant at the 10% level. This explains why not controlling for beliefs may overstate the importance of social preferences, as discussed in section 4.3. The correlations are less consistent for beliefs and risk preferences (2 out of 16 significant) and for risk preferences and social preferences (2 out of 12 significant).

Table A1: Cross-correlation table

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10	(11))
(1) Belief (simple)	1.000										
(2) Belief (complex)	**0.559**	1.000									
	(0.000)										
(3) Belief (general)	**0.158**	**0.149**	1.000								
	(0.062)	(0.080)									
(4) Belief (today)	0.037	0.120	**0.459**	1.000							
	(0.666)	(0.160)	(0.000)								
(5) Altruism	0.072	**0.141**	**0.306**	**0.271**	1.000						
	(0.401)	(0.097)	(0.000)	(0.001)							
(6) Dictator Game	**0.315**	**0.386**	**0.166**	0.102	**0.358**	1.000					
	(0.000)	(0.000)	(0.050)	(0.231)	(0.000)						
(7) Distribution Urn	-0.030	0.029	**0.199**	**0.200**	0.100	**0.201**	1.000				
	(0.730)	(0.734)	(0.019)	(0.018)	(0.241)	(0.017)					
(8) Risk Measure (no losses)	0.140	0.119	0.076	0.007	-0.045	0.084	**0.151**	1.000			
	(0.100)	(0.162)	(0.375)	(0.939)	(0.596)	(0.325)	(0.075)				
(9) Risk Measure (with losses)	-0.125	-0.102	0.007	0.074	-0.138	0.014	0.122	**0.204**	1.000		
	(0.142)	(0.234)	(0.936)	(0.388)	(0.106)	(0.871)	(0.151)	(0.016)			
(10) Risk Self-Assessment	-0.135	-0.010	-0.095	-0.058	-0.088	0.113	0.110	**0.161**	**0.214**	1.000	
	(0.114)	(0.904)	(0.267)	(0.500)	(0.301)	(0.186)	(0.199)	(0.058)	(0.011)		
(11) No Receiver Random Urn	**-0.201**	**-0.210**	-0.013	-0.065	-0.092	0.013	**0.186**	**0.180**	**0.218**	**0.295**	1.000
	(0.018)	(0.013)	(0.881)	(0.445)	(0.283)	(0.882)	(0.028)	(0.034)	(0.010)	(0.000)	

Notes: Correlation coefficients (p-values). For variable definitions, we refer to the notes for Tables 4 to 6. The three boxes contain correlations between variables that are meant to capture the same concept. Bold correlations are significant at the 10% level or better.

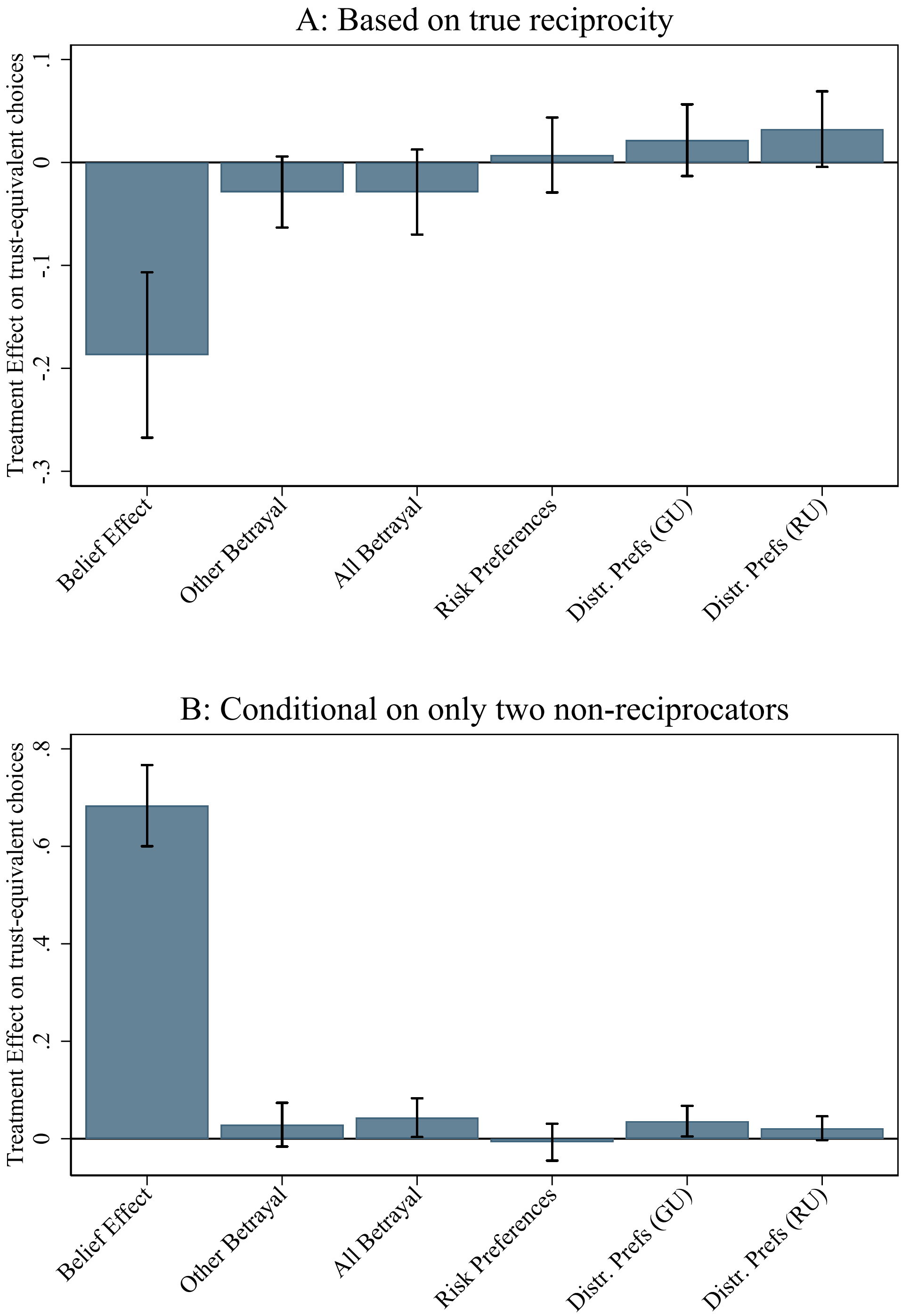


A: Based on complex belief

A: Based on simple belief

Figure A1: Treatment Effects based on Elicited Beliefs

Notes: These figures plot the treatment effect of removing a factor of interest on the fraction of 'in'- (or trust-equivalent-)choices based on the complex (panel A) and simple (panel B) elicited beliefs, respectively. The first bar compares the binary Trust Game decision to the decision in the Trust Game Urn based on the elicited belief. The second bar compares the Other Betrayal Urn to the Trust Game Urn, and the third bar compares the No Betrayal Urn to the Trust Game Urn. The fourth bar compares the Distribution Urn to the No Betrayal Urn. The fifth bar compares the No Receiver Random Urn to the No Betrayal Urn. The sixth bar compares the No Receiver Game Urn to the Other Betrayal Urn. The whiskers are 95%-confidence intervals.


A: Based on true reciprocity

B: Conditional on only two non-reciprocators

Figure A2: Treatment Effects based on True Reciprocity

Notes: These figures plot the treatment effect of removing a factor of interest on the fraction of 'in'- (or trust-equivalent-)choices based on the true rate of reciprocity in the experiment (panel A) and to the case where only two other participants fail to reciprocate (panel B), respectively. Definitions of the respective bars can be found in the notes to Figure A1. The whiskers are 95%-confidence intervals.

B Alternative Samples

In this section we present our results using five alternative sample restrictions. The first sample is the one we used in the main analysis, which removes the 64 participants who either switched multiple times or switched in the wrong direction on at least one of the tasks, as well as a further 33 participants who preferred 5 points over 10 points or preferred (10,10) over (15,15). The second sample includes all observations. The third sample is the one we pre-registered; this sample excludes the 64 participants with irregular switching on one of the tasks, but still includes the 33 participants who preferred 5 points over 10 points or preferred (10,10) over (15,15). The fourth sample is similar to the main sample but adds back 16 participants who switched multiple times in at least one price list but otherwise had no violations, proxying for their true switch point using the difference between their first and last switch point. Finally, the fifth sample considers all participants without violations in the tasks involved in a specific comparison. The fourth and fifth sample correspond to the two robustness checks specified in the pre-analysis plan.

Table A2 reprints the results of column (1) of Table 1 using the five sample selection criteria. The first column reprints the results using the sample used in the main analysis. Compared to the first column, the main difference in the less restrictive samples (columns 2 and 3) is that the first two coefficients are now significant, which implies that removing the effect of betrayal aversion now increases the tendency to trust. Adjusting the choices of multiple switchers has little impact on the parameter estimates (compare columns 1 and 4), and requiring no violations only on the tasks used in this test has no effect, because the table uses the data from all tasks (comparison of columns 1 and 5).

Why is the point estimate for betrayal aversion larger in the less restrictive sample? Figure A3 prints the full distribution for the whole sample (column 2). The figure shows that the point estimate for betrayal aversion (difference between the no betrayal urn and trust urn) is driven by participants who prefer (10,10) over (15,15) in the trust urn but not in the betrayalless urn. Indeed, this is why removing these participants eliminated the effect of betrayal aversion in the main specification. This behavior is difficult to reconcile with betrayal aversion, since in the most optimistic case ((10,10) vs (15,15)) no actual betrayal is possible. Our interpretation is that these participants were not betrayal averse but likely made a mistake in the Trust Game Urn task, which is why we excluded them from the main analysis.

Tables A3 to A6 present robustness tests for the analyses presented in Tables 2 and 3 and accompanying text. In each table, the first column corresponds to the results presented in the main text, the remaining columns use different sample restrictions. Tables A3 and A4 present

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Other Betrayal	0.011	0.036***	0.050***	0.012	0.011
	(0.011)	(0.013)	(0.015)	(0.011)	(0.011)
No Betrayal	0.017	0.048***	0.058***	0.015	0.017
	(0.011)	(0.013)	(0.015)	(0.011)	(0.011)
Distribution	0.103***	0.111***	0.148***	0.090***	0.103***
	(0.014)	(0.016)	(0.017)	(0.014)	(0.014)
No Receiver (RU)	0.052***	0.076***	0.088***	0.051***	0.052***
	(0.012)	(0.014)	(0.014)	(0.012)	(0.012)
No Receiver (GU)	0.047***	0.063***	0.087***	0.045***	0.047***
	(0.014)	(0.014)	(0.016)	(0.013)	(0.014)
Constant	0.350***	0.315***	0.294***	0.352***	0.350***
	(0.014)	(0.014)	(0.016)	(0.013)	(0.014)
Sample	Main	Full Sample	Pre-Reg	MS-adjusted	Local-Rat
Observations	16,680	28,320	20,640	18,840	16,680
Clusters	236	172	139	157	139

Table A2: Testing for Differences in Actions for Different Samples

Notes: Regression estimates (clustered standard errors in parentheses). The first column contains the sample used in the main analysis, the second contains all observations, and the third contains the pre-registered sample. The fourth column adds back multiple switchers who begin and end at the plausible side of the price list in all tasks by proxying for their switch point using the average of the first and last switch point. The fifth column adds back participants with violations in tasks other than the ones considered in the regression (those who are 'locally rational'). For variable definitions see the notes to Table 1. *** p<0.01, ** p<0.05, * p<0.1

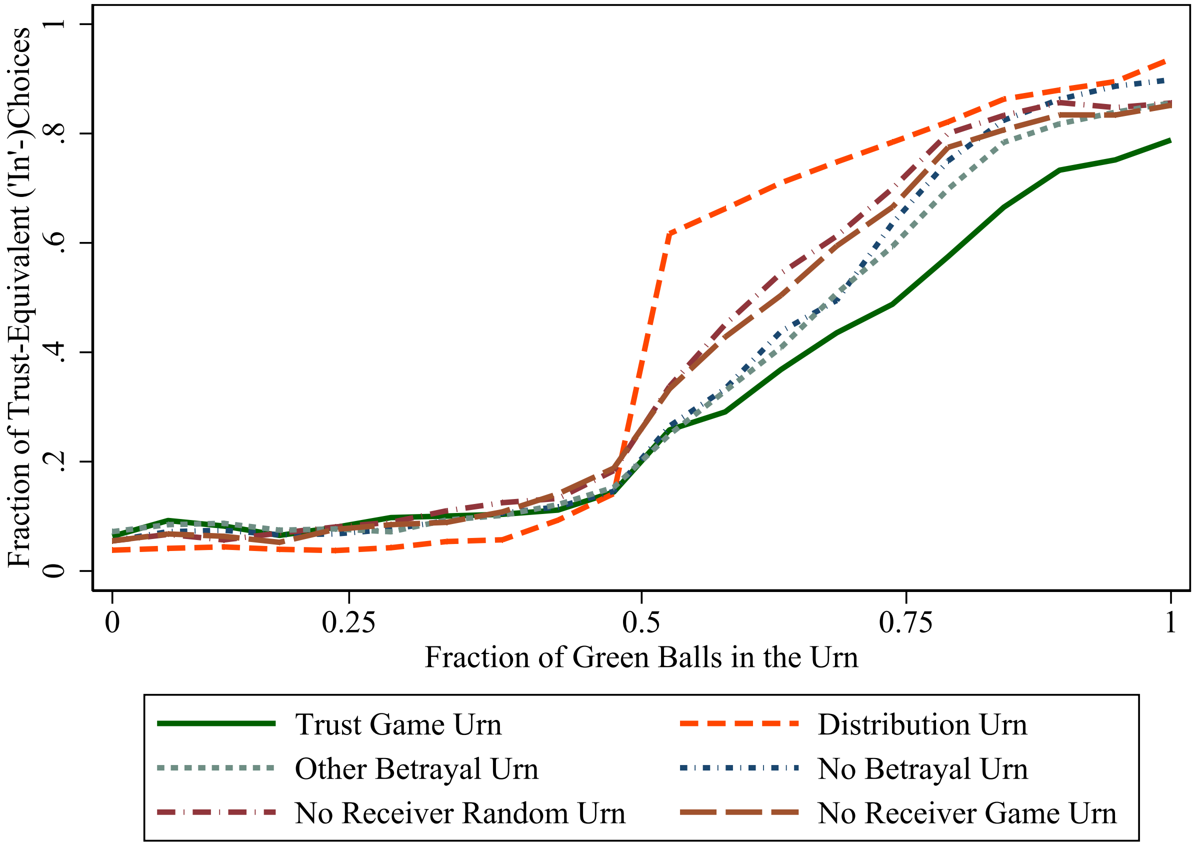


Figure A3: Distributions of Choice List Actions (Full Sample)

Notes: These figures plot the fraction of 'in'- (or trust-equivalent-)choices as a function of the fraction of green balls for all six urn choice lists. The figure uses data from all 236 participants.

	(1)	(2)	(3)	(4)	(5)
	Treatment Estimate				
Binary vs Urn	0.032	0.011	0.032	0.013	0.012
Other Betrayal	-0.025	-0.004	0.009	-0.013	0.000
Betrayal Aversion	-0.004	0.019	0.015	0.013	0.009
Risk Aversion	0.104***	0.053*	0.093***	0.108***	0.067**
Dist. Pref. (RU)	0.011	-0.002	0.006	0.019	0.008
Dist. Pref. (GU)	0.036*	0.036*	0.023	0.051**	0.022
Sample	Main	Full Sample	Pre-Reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	165-195

Table A3: Decomposition based on complex beliefs

Notes: Each entry presents the effect of one of the preference components, analogously to Table 2, Panel A. The first column contains the sample used in the main analysis, the second contains all observations, and the third contains the pre-registered sample. The fourth column adds back multiple switchers who begin and end at the plausible side of the price list in all tasks by proxying for their switch point using the total number of 'in'-choices. The fifth column adds back participants with violations in tasks other than the ones considered in the regression. We proxy for the complex beliefs using the average choice of urn in cases of multiple switching. *** p<0.01, ** p<0.05, * p<0.1

the choices made in the task rows corresponding to the complex and simple elicited beliefs respectively (Table 2). The results are very similar across columns. The main difference is that, when using the full sample, the effect of risk preferences is only significant at the 10% level (where in the complex belief case we proxy for the complex beliefs using the average choice of urn in case of multiple switching).

In Tables A5 and A6 we present results corresponding to Table 3. When we base the analysis on the true rate of reciprocity (Table A5), none of the components have a significant effect on behavior, regardless of the sample used. For the case where only two others fail to reciprocate (2 red balls in the urn, Table A6), the effect of betrayal aversion is larger in the less restrictive samples, and the effect of being betrayed by someone else also becomes significant in some samples. As discussed previously this is driven by participants who prefer (10,10) over (15,15) in the trust game urn but none of the other urns, which we consider a likely mistake.

Tables A7 to A9 replicate the decomposition approach based on regressions. Each column presents the results of the regression approach for a particular sample. Each row represents the main coefficient estimate of a separate regression corresponding to the respective column in the corresponding table in the main text. For example, the third row in Table A7 presents the robustness checks for the third column in Table 4. When it comes to the role of beliefs, A7 demonstrates that noisier samples lead to smaller point estimates across the board, as

	(1)	(2)	(3)	(4)	(5)
			Treatment Estimate		
Binary vs Urn	-0.014	-0.030	-0.006	-0.006	-0.029
Other Betrayal	0.007	0.034	0.035	0.006	0.018
Betrayal Aversion	0.007	0.013	0.023	0.013	0.012
Risk Aversion	0.079**	0.051*	0.081***	0.070**	0.056**
Dist. Pref. (RU)	0.000	0.008	0.000	-0.019	0.016
Dist. Pref. (GU)	0.000	0.008	0.006	-0.006	0.006
Sample	Main	Full Sample	Pre-Reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	165-195

Table A4: Decomposition based on simple beliefs

Notes: Each entry presents the effect of one of the preference components, analogously to Table 2, Panel B. See the notes to Table A3 for further details. *** p<0.01, ** p<0.05, * p<0.1

	(1)	(2)	(3)	(4)	(5)
			Treatment Estimate		
Binary vs Urn	-0.014	-0.030	-0.006	-0.006	-0.029
Other Betrayal	0.007	0.034	0.035	0.006	0.018
Betrayal Aversion	0.007	0.013	0.023	0.013	0.012
Risk Aversion	0.079**	0.051*	0.081***	0.070**	0.056**
Dist. Pref. (RU)	0.000	0.008	0.000	-0.019	0.016
Dist. Pref. (GU)	0.000	0.008	0.006	-0.006	0.006
Sample	Main	Full Sample	Pre-Reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	165-195

Table A5: Decomposition based on true reciprocity

Notes: Each entry presents the effect of one of the preference components, analogously to Table 3, Panel A. See the notes to Table A3 for further details. *** p<0.01, ** p<0.05, * p<0.1

	(1)	(2)	(3)	(4)	(5)
	Treatment Estimate				
Binary vs Urn	-0.683***	-0.517***	-0.570***	-0.675***	-0.684***
Other Betrayal	0.029	0.085***	0.110***	0.045*	0.036
Betrayal Aversion	0.043**	0.131***	0.140***	0.051**	0.060**
Risk Aversion	-0.007	0.017	0.023	-0.019	-0.015
Dist. Pref. (RU)	0.022*	-0.004	0.023	0.025**	0.032*
Dist. Pref. (GU)	0.036*	0.013	0.047**	0.032**	0.028*
Sample	Main	Full Sample	Pre-Reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	165-195

Table A6: Decomposition based on only two non-reciprocators

Notes: Each entry presents the effect of one of the preference components, analogously to Table 3, Panel B. See the notes to Table A3 for further details. *** p<0.01, ** p<0.05, * p<0.1

expected with measurement error in the x-variables. However, the overall pattern is similar throughout: the treatment estimate is always largest, the complex belief elicitation coefficient is similar in size, and all other coefficients are smaller.

Table A8 presents the corresponding results for risk preferences. Interestingly, the predictive power of the loss-based price-list measure of risk preference is largest in samples where violators of expected utility (including multiple switchers) are included (column 2 and 5 in particular). However, this is not true for the qualitative self-assessment measure or the price list measure without losses. Taken together, this suggests that the larger coefficients estimated in these samples may reflect spurious correlations driven by people who violate Expected Utility in similar ways in several price lists. In line with this, multiple switchers in the risk measure task with losses are both more likely to trust and score higher on the risk measure, which could explain the significant coefficients observed in columns 2 and 5. The fact that the treatment estimate for risk preferences is slightly lower in columns 2 and 5 is due to the increased point estimate for betrayal aversion in those samples.

Table A9 presents the results for social preferences. Here the main patterns are similar to the main analysis: measures of social preferences predict trust in the trust game, their point estimates are reduced when controlling for beliefs, and the treatment estimate is not significantly different from zero. The main difference is that in some samples controlling for beliefs no longer renders the social preference coefficients insignificant (the final two rows in the table). Apart from the larger sample size, this may also be due to the belief variable becoming less predictive in larger samples due to increased noise. This may then increase the scope for social preference variables to spuriously pick up an effect that is actually driven by beliefs, but not filtered out in the regression due to measurement error in the belief variable.

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Belief (general)	0.381***	0.251**	0.346***	0.310**	0.304**
	(0.141)	(0.112)	(0.118)	(0.135)	(0.127)
Belief (today)	0.312**	0.203**	0.344***	0.265**	0.241**
	(0.136)	(0.096)	(0.110)	(0.130)	(0.116)
Belief (simple elicitation)	0.621***	0.489***	0.563***	0.626***	0.582***
	(0.124)	(0.102)	(0.112)	(0.122)	(0.115)
Belief (complex elicitation)	0.974***	0.670***	0.776***	0.952***	0.930***
	(0.141)	(0.112)	(0.132)	(0.138)	(0.131)
Treatment estimate	1.000***	0.725***	0.843***	1.000***	0.988***
	(0.000)	(0.034)	(0.028)	(0.000)	(0.008)
Sample	Main	Full	Pre-reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	171

Table A7: Identifying the Belief Effect using Regressions: Other Samples

Notes: Each entry presents the coefficient estimate from a separate regression of the binary trust decision (trust-1). The first column contains the sample used in the main analysis, the second contains all observations, and the third contains the pre-registered sample. The fourth column adds back multiple switchers who begin and end at the plausible side of the price list in all tasks by proxying for their switch point using the total number of 'in'-choices. The fifth column adds back participants with violations in tasks other than the urn with betrayal. Each row contains the main coefficient estimate from a separate regression analogous to the five regressions presented in Table 4. *** p<0.01, ** p<0.05, * p<0.1

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Risk Measure (no losses)	0.016*	0.022***	0.023**	0.018**	0.026***
	(0.009)	(0.006)	(0.010)	(0.009)	(0.008)
Risk Measure (with losses)	0.001	0.102**	0.081	0.017	0.101*
	(0.068)	(0.051)	(0.060)	(0.062)	(0.057)
Self-Assessment	0.000	0.001	0.002	0.000	0.001
	(0.000)	(0.001)	(0.002)	(0.000)	(0.001)
Random Urn	-0.008	0.037**	0.027	-0.002	0.020
	(0.026)	(0.014)	(0.020)	(0.024)	(0.016)
Treatment estimate	0.104***	0.053*	0.093***	0.108**	0.067**
	(0.032)	(0.028)	(0.029)	(0.036)	(0.028)
Sample	Main	Full	Pre-reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	195

Table A8: Identifying the Risk Effect using Regressions: Other Samples

Notes: Each entry presents the coefficient estimate from a separate regression of the binary trust decision (trust-1). The first column contains the sample used in the main analysis, the second contains all observations, and the third contains the pre-registered sample. The fourth column adds back multiple switchers who begin and end at the plausible side of the price list in all tasks by proxying for their switch point using the total number of 'in'-choices. The fifth column adds back participants with violations in tasks other than the urn without betrayal and the distribution urn. Each row contains the main coefficient estimate from a separate regression analogous to the five regressions presented in Table 4. *** p<0.01, ** p<0.05, * p<0.1

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Trust-Equivalent Action				
Dictator Game	-0.145**	-0.111***	-0.094***	-0.157***	-0.118***
	(0.040)	(0.027)	(0.032)	(0.037)	(0.031)
Altruism	-0.259**	-0.302***	-0.227**	-0.295***	-0.274***
	(0.113)	(0.081)	(0.092)	(0.101)	(0.088)
Distribution Urn	-0.033**	-0.024**	-0.027***	-0.038***	-0.033**
	(0.013)	(0.011)	(0.010)	(0.014)	(0.013)
Treatment estimate	0.011	-0.002	0.006	0.019	0.000
	(0.022)	(0.022	(0.019)	(0.023)	(0.021)
Dictator Game	-0.075*	-0.082***	-0.046	-0.097**	-0.077**
(Controlling for Beliefs)	(0.042)	(0.027)	(0.032)	(0.038)	(0.031)
Altruism	-0.162	-0.249***	-0.173*	-0.215**	-0.212***
(Controlling for Beliefs)	(0.106)	(0.079)	(0.088)	(0.096)	(0.084)
Sample	Main	Full	Pre-reg	MS-adjusted	Local-Rat
Observations	139	236	172	157	185

Table A9: Identifying the Distributional Preferences Effect using Regressions: Other Samples

Notes: Each entry presents the coefficient estimate from a separate regression of the binary trust decision (trust-1). The first column contains the sample used in the main analysis, the second contains all observations, and the third contains the pre-registered sample. The fourth column adds back multiple switchers who begin and end at the plausible side of the price list in all tasks by proxying for their switch point using the total number of 'in'-choices. The fifth column adds back participants with violations in tasks other than the urn without betrayal and the random urn. Each row contains the main coefficient estimate from a separate regression analogous to the five regressions presented in Table 6. *** p<0.01, ** p<0.05, * p<0.1

C Pre-Analyis Plan

In this section, we reproduce the pre-analysis plan (as registered on the AEA registry at `https://www.socialscienceregistry.org/trials/5146`). After each section, we also add a few remarks explaining where the respective results can be found in the paper, and how our analysis differs from the pre-analysis plan (if at all).

C.1 Main Analysis

Our dependent variable (DV) is trust behavior, i.e. a dummy variable capturing whether a participant chooses the trusting or the non-trusting option in a trust game. The aim of our study is to decompose variation in trust behavior into variation in four underlying factors:

1. Risk attitudes 2. Social preferences 3. Betrayal aversion 4. Beliefs about the probability of being reciprocated We compare two main approaches of decomposing trust:

1. In the first approach, we decompose trust by removing the four factors or combinations thereof from the trust game in a series of within-subject control treatments and comparing trust behavior across these trust game variants.

2. In the second approach, we decompose trust by measuring each factor in a separate task and using these measures as independent variables in a regression explaining trust behavior.

We conduct three variants of the second approach:

The first (2.a), following traditional regression approaches, uses choices in unrelated tasks to measure preferences, e.g. with respect to risk. The second approach (2.b) follows the traditional regression approach but with standard measures of self-declared risk preferences, trust, reciprocity and altruism instead of choices from incentivized tasks. The third approach (2.c) instead uses tasks that are derived from the trust game to isolate the risk and social preference factors inherent in the trust game.

Authors' Notes: We present the results of the first approach in section 3.3, and the results of the second approach in section 4.

C.2 Main Hypotheses

Our key hypothesis regarding approaches 1 and 2 is that in approach 1, we can avoid the mis-attribution of explanatory power to one the four underlying factors that may emerge in approach 2 due to omitted variable bias.

Our key hypothesis regarding the comparison of approaches 2.a, 2.b, and 2.c is that due to higher measurement error of the relevant preferences in the unrelated tasks and self-declared preferences, approach 2.c yields stronger explanatory power of the individual factors than 2.a and 2.b. We are agnostic with respect to the comparison of the incentivized unrelated tasks and the self-declared preferences, but will compare their explanatory power for exploration.

C.3 Details

A key hypothesis underlying our systematic approach is that the four relevant explanatory factors of trust behavior are correlated. If they are, omitting one of the factors will lead to over-attribution to an included factor, if the omitted factor is positively correlated with the included factor and both affect trust choices in the same direction (e.g., if pessimistic expectations are correlated with risk aversion). By contrast, it will lead to under-attribution to the included factor if the factors are negatively correlated but affect trust choices in the same direction (e.g., if risk aversion and pro-social preferences are negatively correlated).

We will therefore test whether the factors as elicited in the second approach are correlated and the degree to which this will cause an omitted variable bias. We will then test whether our first approach allows us to reduce or eliminate this omitted-variable bias.

A potential related issue in the second approach is that the factors used to explain trust in the second approach are likely to be measured with error. If this is true, then their coefficients will underestimate the true effect of the underlying factor in expectation. To estimate the size of the resulting bias, we will compare the coefficients on the underlying factor in approach 2 to the results of approach 1.

We aim to apply our two approaches to two types of research questions. First, a direct analysis of trust behavior. Second, to the explanation of possible demographic differences in trust. We will therefore investigate whether there are relevant differences in trust by gender and field of study and if we identify such a difference, analyze misattribution due to omitted-variable bias in the regression-based approach.

Authors' Notes: We explore correlations between the underlying factors in Table A1, and compare the results of the two approaches in section 4, where we also touch upon both omitted variable bias and measurement error. We found no significant differences in trust behavior by gender or field of study, and therefore did not use our two approaches to analyze the factors driving such differences.

C.4 Exclusion Criteria

In our main analysis, we exclude all participants who make at least one inconsistent choice, i.e. participants who have multiple switches in a choice list and those who are switching in the direction that is inconsistent with reasonable utility maximization. As robustness checks, we

1. include for the analysis of any specific test all participants for whom inconsistent choices occurred only in tasks that are unrelated to this test
2. include multiple switchers if they begin and end on the plausible side of the respective choice list by proxying their intended switch point by the mid-point between their first and last switching point.

Authors' Notes: For the main analysis we decided to also remove 33 additional participants who made dominated choices, as we explain in the beginning of the results section. We include the results from the pre-registered sample and the two robustness checks in section B in Appendix B, where we show that the results are very similar in all cases.

C.5 Example: Beliefs about Probability of Being Reciprocated

As an example of approach 1 and 2, consider the effect of beliefs about being reciprocated. This uses the following variables collected in the experiment:

- Trust: equal to 1 if the participant 'trusts' as first-mover in the trust game.
- Belief: expected number of participants who reciprocate in the session (reservation probability elicitation method).
- BeliefSimple: expected number of participants who reciprocate in the session (simple elicitation method).
- TrustPL: price lists that asks people whether they would trust conditional on the number of participants in their session who would reciprocate.
  - TrustPLBelief: the row on the price list that corresponds to the participant's belief. E.g., if the participant thinks there are 11 participants who reciprocate in her session, this would be the row in which 11 participants reciprocate in the price list (i.e., there are 7 green balls in the 'game urn').
  - TrustPLTrue: the row on the price list that corresponds to the true (expected) number of participants who reciprocate in a session (based on the behavior of participants in all sessions). E.g., if 8.5/19 participants reciprocate on average across all sessions, then we would take the average of row 8 and row 9.

The tests conducted for each approach would then be as follows:

- Approach 1: we test if TrustPLBelief=TrustPLTrue using a t-test.

  - Alternatively, test Trust=TrustPLTrue

- Approach 2: we regress Trust on Belief.

In approach 1, a positive effect (i.e., TrustPLBelief¿TrustPLTrue) would indicate that removing the effect of subjective beliefs (i.e., de-biasing beliefs) makes people less likely to trust (e.g., if people were too optimistic about the trustworthiness of others). Finding a significant positive 'belief' coefficient in approach 2 would similarly tell us that lowering the belief would make people less likely to trust. We can also use approach 2 to look at the effect of 'de-biasing beliefs' by multiplying the estimated coefficients of beliefs by the difference between the average elicited belief and the true (expected) number of reciprocating participants in a session.

Authors' Comments: We include the analysis of approach 1 in section 3.3.3, and the analysis of approach 2 in section 4, where we also discuss the effect of de-biasing beliefs.

C.6 Other Factors

For the other factors, the two approaches work in a similar way. Approach 2 always regresses trust on the measure of the factor of interest. Approach 1 always compares the difference between the action taken in a particular row in particular price list to the action taken in another row in another price list.

D Instructions (translated from German)

Welcome to this experiment! You can earn money and the amount of money you get in the end depends on the decisions you and other participants make, as well as on random draws.

During the experiment you are not allowed to use electronic devices or communicate with other participants. Please use only the programs and functions intended for the experiment. Please do not talk to the other participants. If you have a question, please raise your hand. We will then come to you and answer your question silently. Please do not ask your questions out loud under any circumstances. If the question is relevant to all participants, we will repeat it out loud. If you violate these rules, we will have to exclude you from the experiment and the payment.

Please read these instructions carefully now. The instructions are identical for all participants.

The experiment consists of several tasks. The tasks are divided into two groups. Some of the tasks in the first group are similar to each other, but they are not identical. You will receive detailed instructions for each task on the screen. Read these instructions very carefully. You will perform each task only once. For each task, you will be shown the instructions first and only after a certain amount of time will you be able to make a decision.

Your payoffs for the tasks will be measured in points.

The tasks of the first group relate to the following decision situation. There are two people, A and B. Person A decides between "IN" and "OUT". Person B decides between "EQUAL" and "UNEQUAL". If A chooses OUT, person A and person B each get 10 points, regardless of B's decision. If A chooses IN, B's decision matters. If B chooses EQUAL, then A and B both get 15 points each. If B chooses UNEQUAL, A gets 5 points and B gets 20 points. Hence, A can guarantee himself a payoff of 10 points by choosing OUT. Whether A does better or worse by choosing IN than by choosing OUT depends on which decision B takes. In the rest of the experiment, we will refer to this decision situation as the GAME.

The course of the experiment is as follows: In the first part, you make decisions in ten different tasks of the first group, all related to the GAME. In the second part, you make decisions in four tasks of the second group that do not relate to the GAME. Your decisions in the first part have no effect on the second part. Finally, there is a short questionnaire.

Your final payoff comprises the payoffs for one task from the first part and one task from the second part. Which task this is, is determined randomly in each case. Some tasks involve two people. You could take either role. Your role is determined randomly in these cases. In some of these tasks, only one person's decision determines the payoff for both people. If you take the other role, your payoff is independent of your decisions in that part. This will become clear from the descriptions of each task.

You will not get information about the result for any of the tasks until you have gone through all of them. At the end, you will learn the results of the two tasks that were selected for you to be paid out.

The exchange rate for the points you can earn in the course of the experiment is

$$1 \text{ point} = 0.75 \text{ Euro.}$$

You will also receive a fixed amount of 3 euros for participating.

If there is anything that you have not understood, please indicate this by a show of hands. We will then answer your questions one by one.

ON-SCREEN INSTRUCTIONS FOR EACH TASK

TASK 1

In this task, you decide in the GAME as person B. As a reminder, if person A chooses OUT, your decision is irrelevant and you and A both get 10 points each. If person A chooses IN, your decision matters. If you choose EQUAL, you and A both will receive 15 points each. If you choose UNEQUAL, you will receive 20 points and A will receive 5 points.

If this task is chosen as relevant for your payment, you will be paired with another person. Your payoff then depends on your decision as B and the other person's decision as A. You receive the payoff for B, and the other person receives the payoff for A.

Your decision in task 1 may also affect other people's payoffs, or possibly your payoff if any of the other tasks in the first group are drawn as payoffs for other people. In terms of payoff possibilities, the other tasks are very similar to the GAME.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 2

There are 20 participants in the lab. Now it is your task to estimate how many of the other 19 participants chose EQUAL in Task 1.

If you estimate the exact number correctly, you will receive 15 points. If your estimate differs from the actual number, you will lose half a point for each incorrectly estimated person.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

INTERIM SCREEN

For the remaining tasks we need two virtual urns. Balls will be drawn from these urns in later tasks, which can determine your payoffs.

The first urn (GAME URN) is composed of the choices made by the other participants in TASK 1. For each participant who chose EQUAL in this task, a green ball goes into the urn, and for each participant who chose UNEQUAL, a red ball goes into the urn. Since there are 20 participants in the experiment, but your own decision in task 1 is not taken into account, there are exactly 19 balls in the GAME URN.

The second urn (RANDOM URN) is composed purely at random. Like the GAME URN, it is filled with 19 balls and all balls are either green or red. However, the number of green and red balls is determined randomly. Any possible number of green balls (and correspondingly red balls) between 0 and 19 is equally likely.

[Participants received an additional sheet illustrating the GAME and RANDOM URNS graphically. This illustration is reproduced at the end of these instructions.]

TASK 3

In this task you decide between the GAME URN and the RANDOM URN. A ball is drawn from the urn you choose. If a red ball is drawn, you get 5 points, if a green ball is drawn, you get 15 points. So if there is exactly one green ball in the urn you choose, you have a 1/19 chance of getting 15 points. If there are exactly two green balls in the urn, this chance increases to 2/19, and so on. If all the balls in the chosen urn are green, you will surely get 15 points.

You do not know the composition of the GAME URN, nor of the RANDOM URN. However, in this task you can decide for each possible composition of the RANDOM URN from which urn the ball should be drawn. Hence, you choose an urn in case there is no green ball, one green ball, two green balls, etc. in the RANDOM URN.

If this task is drawn as payoff-relevant, your choice will be implemented for the actual composition of the GAME URN and one ball will be drawn from the chosen urn. Therefore, for a given number of green balls in the RANDOM URN, you should choose the GAME URN if you believe there are more than that number of green balls in the GAME URN and the RANDOM URN if you believe there are less than that number of green balls in the GAME URN. Obviously, you should choose the GAME URN if there are 0 green balls in the RANDOM URN and the RANDOM URN if there are 19 green balls in the RANDOM URN. You should switch from the GAME URN to the RANDOM URN as soon as the number of green balls in the RANDOM URN exceeds your estimate of the number of green balls in the GAME URN.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 4

In this task, you decide in the GAME as person A. Remember, if you choose OUT, person B's decision is irrelevant and you and B both get 10 points each. If you choose IN, person B's decision matters. If B chooses EQUAL, you and B both get 15 points each. If B chooses UNEQUAL, you will receive 5 points and B will receive 20 points.

If this task is chosen as relevant for your payoff, you will be paired with another person. Your payoff then depends on your decision as A and the decision of the other person as B from TASK 1. You will receive the payoff for A, and the other person will receive the payoff for B.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 5

In this task, your decision affects only your own payoff. Your payoff depends on your decision and, if necessary, a draw from the RANDOM URN.

You decide between IN and OUT. If you choose OUT, you will receive 10 points, regardless of the draw from the RANDOM URN. If you choose IN, your payoff depends on the draw from the RANDOM URN. If a green ball is drawn, you get 15 points, if a red ball is drawn, you get 5 points.

You do not know the composition of the RANDOM URN but you can choose between IN and OUT for each possible number of green balls in the RANDOM URN. If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the RANDOM URN and one ball will be drawn from the RANDOM URN.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 6

In this task, your decision affects only your own payoff. Your payoff depends on your decision and, if necessary, a draw from the GAME URN.

You decide between IN and OUT. If you choose OUT, you get 10 points, regardless of the draw from the GAME URN. If you choose IN, your payoff depends on the draw from the GAME URN. If a green ball is drawn, you will receive 15 points, if a red ball is drawn, you will receive 5 points.

You do not know the composition of the GAME URN, but you can choose between IN and OUT for any number of green balls in the GAME URI. If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the GAME URN and one ball will be drawn from the GAME URN.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 7

In this task, your decision affects your own payoff and that of another participant. Both payoffs depend on your decision and, if applicable, the composition of the RANDOM URN.

You decide between IN and OUT. If you choose OUT, you and the other participant will each receive 10 points, regardless of the composition of the RANDOM URN. If you choose IN, your payoff depends on the composition of the RANDOM URN. If there are only red balls in the RANDOM URN, you will receive 5 points and the other participant will receive 20 points. For each green ball in the RANDOM URN, your payoff increases by 10/19 points and the other participant's payoff decreases by 5/19 points. It is rounded to tenths of points in each case. So if there are only green balls in the RANDOM URN, you get 5+19*(10/19) = 15 points and the other participant gets 20-19*(5/19) = 15 points. You do not know the composition of the RANDOM URN, but for each possible number of green balls in the RANDOM URN, you can choose between IN and OUT. Thus, for example, if the number of green balls in the RANDOM URN is 0, you choose whether you and the other participant each get 10 points or you get 5 points and the other participant gets 20 points. If, on the other hand, all 19 balls in the RANDOM URN are green, you choose whether you and the other participant each get 10 points, or you and the other participant each get 15 points. For each decision, you will find the resulting payoffs of IN and OUT below.

If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the RANDOM URN. No ball is drawn, only the composition of the RANDOM URN is relevant. For another participant, the result will also be implemented according to your choice and the composition of the RANDOM URN. The payoff of this participant is then independent of their own decisions in the first part of the experiment. Similarly, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff is independent of your decisions in the first part of the experiment.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 8

In this task, your decision affects your own payoff and that of another participant. Both payoffs depend on your decision and, if necessary, on a draw from the RANDOM URN.

You decide between IN and OUT. If you choose OUT, you and the other participant will each receive 10 points, regardless of the draw from the RANDOM URN. If you choose IN, your payoff depends on the draw from the RANDOM URN. If a green ball is drawn, you and the other participant will each receive 15 points; if a red ball is drawn, you will receive 5 points and the other participant will receive 20 points.

You do not know the composition of the RANDOM URN, but you can choose between IN and OUT for each possible number of green balls in the RANDOM URN.

If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the RANDOM URN and a ball will be drawn from the RANDOM URN. For another participant, the result is also implemented according to your choice and the draw from the RANDOM URN. The payoff for this participant is then independent of their decisions in the first part of the experiment. Similarly, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff is independent of your own decisions in the first part of the experiment.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 9

In this task, your decision affects your own payoff and that of another participant. Both payoffs depend on your decision and, if necessary, on a draw from the GAME URN.

You decide between IN and OUT. If you choose OUT, you and the other participant will each receive 10 points, regardless of the draw from the GAME URN. If you choose IN, your payoff depends on the draw from the GAME URN. If a green ball is drawn, you and the other participant will each receive 15 points; if a red ball is drawn, you will receive 5 points and the other participant will receive 20 points. Note: the other participant is NOT the one whose decision in TASK 1 (deciding as person B in the GAME) determines the color of the ball. Hence, whether the ball drawn is green or red is independent of the other participant's decision in TASK 1.

You do not know the composition of the GAME URN, but you can choose between IN and OUT for any possible number of green balls in the GAME URN. If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the GAME URN and one ball will be drawn from the GAME URN. For another participant, the result will also be implemented according to your choice and the draw from the GAME URN. The payoff for this participant is then independent of their own decisions in the first part of the experiment. Likewise, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff is independent of your decisions in the first part of the experiment.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 10

In this task, your decision affects your own payoff and that of another participant. Both payoffs depend on your decision and, if applicable, the other participant's decision and on a draw from the GAME URN.

You decide between IN and OUT. If you choose OUT, you and the other participant will each receive 10 points, regardless of the draw from the GAME URN. If you choose IN, your payoff depends on the draw from the GAME URN. If a green ball is drawn, you and the other participant will each receive 15 points; if a red ball is drawn, you will receive 5 points and the other participant will receive 20 points. Note: the other participant is the one whose decision in TASK 1 (decision as person B in the GAME) determines the color of the ball drawn. Thus, the ball drawn is green if the other participant chose EQUAL in TASK 1 and the ball is red if the other participant chose UNEQUAL in TASK 1.

You do not know the composition of the GAME URN, but you can choose between IN and OUT for any possible number of green balls in the GAME URN. If this task is drawn as relevant for payment, your choice will be implemented for the actual composition of the GAME URN and one ball will be drawn from the GAME URN. For another participant, the result will also be implemented according to your choice and the draw from the GAME URN. The payoff for this participant therefore depends on their own decision in TASK 1 and on your decision in TASK 10. Similarly, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff depends on your own decision in TASK 1 and on the decision of the other participant in TASK 10.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

INTERIM SCREEN

Now the second part of the experiment begins. Your payoff in this part of the experiment is independent of the two urns from the first part of the experiment and your decisions in the first part of the experiment. One of the tasks from this part will be chosen as relevant for your payment at the end of the experiment. However, you may instead be selected as an affected second participant whose payoff depends on the decision of another participant. In this case, your payoff for the second part is independent of your decisions.

TASK 11

In this task, your decision affects only your own payoff. Your payoff depends on your decision and, if necessary, a random draw.

In this task, you make a sequence of decisions to play a lottery or not. In each case, the lottery yields a payoff of 0 points with a probability of 50% and a payoff of 10 points with a probability of 50%. If you decide not to play the lottery, you will receive a safe payoff. This safe payoff varies between the individual decisions. In the first decision it is 1 point, in the last decision it is 9 points. For each decision you can find the safe payoff below.

If this task is drawn as relevant for payment, one of the possible safe payoffs is randomly selected as relevant. Each of the possible safe payoffs has the same probability of being selected. Your decision for the case of this safe payoff will then be implemented. If you choose the safe payoff, you will receive this payoff. If you choose the lottery, it will be played and you will receive 0 or 10 points, each with the same probability.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 12

In this task, your decision affects only your own payoff. Your payoff depends on your decision and, if necessary, a random draw.

In this task, you make a sequence of decisions to play a lottery or not. In each case, the lottery yields a negative payoff with a probability of 50% and a payoff of 5 points with a probability of 50%. If you decide not to play the lottery, you will not receive a payoff. The negative payoff varies between the individual decisions. In the first decision it is -0.5 points, in the last decision it is -5 points. For each decision you can find the negative payoff below.

If this task is drawn as relevant for payment, first one of the possible negative payoffs is randomly selected as relevant. Each of the possible negative payoffs has the same probability of being selected. Your decision for the case of this negative payoff will then be implemented. If you do not choose the lottery, you will not receive a payoff. If you choose the lottery, it will be played and you will receive 5 points or the negative payoff, each with the same probability. If you choose the lottery and it results in a negative payoff, it will be offset against your payoff from the first part of the experiment.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 13

In this task, your decision concerns your own payoff and that of another participant. Your payoff and that of the other participant depend only on your decision.

In this task, you decide in a sequence of decisions between two distributions of payoffs to yourself and to the other participant. In each decision one available option is a distribution where you get 10 points and the other participant gets 0 points. In the other distribution, you and the other participant each get the same payoff of X points. Between the individual decisions, the amount X of this payoff varies. In the first decision it is 0 points, in the last decision it is 11 points. For each decision you can find below the amount of this payoff.

If this task is drawn as relevant for payment, first one of the possible X is randomly chosen as relevant. Each possible X has the same probability of being selected. Your decision for the case of this X will then be implemented. If you choose the first distribution, you will receive 10 points and the other participant will receive 0 points. If you choose the second distribution, you and the other participant will each receive X points. Keep in mind that each of your choices is only relevant to the X in question and you cannot influence the choice of the X. For another participant, the outcome is likewise implemented according to your decision for the randomly selected X. The payoff for this participant is then independent of their own decisions in the second part of the experiment. Similarly, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff is independent of your own decisions in the second part of the experiment.

Please make your decision now! If anything is unclear, raise your hand and we will come to you.

TASK 14

In this task, your decision involves your own payoff and that of another participant. Your payoff and that of the other participant depend only on your decision.

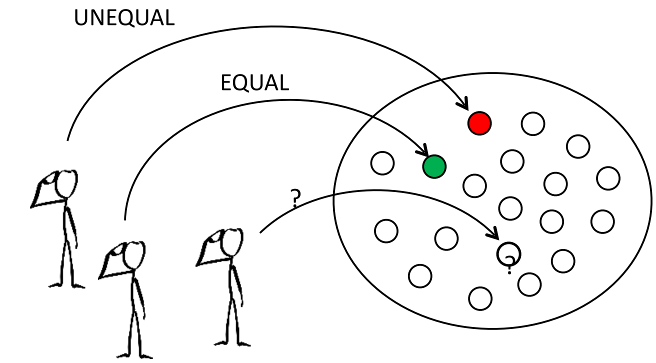
In this task, you decide in a sequence of decisions between two distributions of payoffs to yourself and to the other participant. In each decision one available option is a distribution where you get 5 points and the other participant gets 10 points. In the other distribution, you and the other participant each get the same payoff of X points. Between the individual decisions the amount X of this payoff varies. In the first decision it is 0 points, in the last decision 11 points. For each decision you can find below the amount of this payoff.

If this task is drawn as relevant for payment, first one of the possible X is randomly selected as relevant. Each possible X has the same probability of being selected. Your decision for the case of this X will then be implemented. If you choose the first distribution, you will receive 5 points and the other participant will receive 10 points. If you choose the second distribution, you and the other participant will each receive X points. Keep in mind that each of your choices is only relevant to the X in question and you cannot influence the choice of the X. For another participant, the outcome is likewise implemented according to your decision for the randomly selected X. The payoff for this participant is then independent of their own decisions in the second part of the experiment. Similarly, it may happen that another participant is selected and their decision is implemented and you are assigned the role of the second participant. In this case, your payoff is independent of your own decisions in the second part of the experiment.

Please make your decision now! If something is unclear, raise your hand and we will come to you.

Illustration of GAME and RANDOM URNS

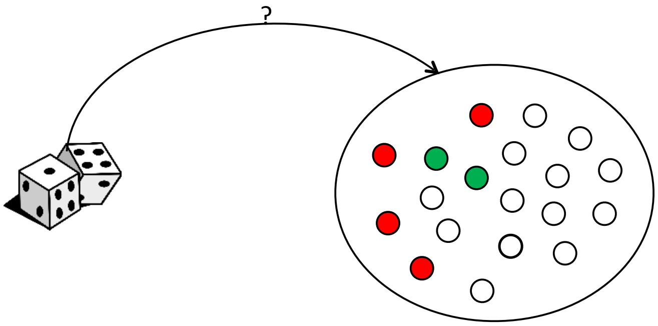
GAME URN



The behavior of the 19 other participants in TASK 1 determines the number of red and green balls in the GAME URN.

For every participant who chooses EQUAL one green ball is added to the GAME URN; for every participant who chooses UNEQUAL, a red ball is added to the GAME URN.

RANDOM URN



A random draw determines the number of red and green balls in the RANDOM URN.

All combinations of green and red balls, from 0 green balls and 19 red balls to 19 green balls and 0 red balls, are equally likely.

E Self-Assessment Questions (translated from German)

After obtaining feedback on the results of the experiment, the participants answered the following questions.

1. How do you rate yourself? Are you in general a person willing to take risks or do you try to avoid risks? Please answer using the following scale where value 0 means “not at all willing to take risks” and value 10 means “highly willing to take risks”. With values in between you can provide a gradual assessment.

2. What is your opinion on the following three statements? The points from left to write means “agree completely”, “tend to agree”, “tend to disagree”, “disagree completely”.

   (a) In general, one can trust people.

   (b) Today, you cannot rely on anyone anymore.

   (c) If you are dealing with strangers, it is better to be careful before you trust them.

3. What ist your self-assessment with respect to the following questions? The points from left to write means “agree completely”, “tend to agree”, “tend to disagree”, “disagree completely”.

   (a) When someone does me a favor, I am willing to return it.

   (b) When I feel treated unfairly, I take revenge on the first opportunity, even if this has some costs.

   (c) I am willing to contribute to good causus.

   (d) The well-being of other people is important for me.