Trust, Values and False Consensus[*]

Jeff Butler, Paola Giuliano and Luigi Guiso[1]

Abstract

Individuals' beliefs about the trustworthiness of a generic member of the population are both heterogeneous across individuals and persistent across generations. We investigate one mechanism yielding these dual patterns: false consensus. In the context of a trust game experiment, we show that the relationship between behavior and beliefs is consistent with individuals extrapolating their trust beliefs from their own trustworthiness and that this tendency continues even after substantial learning opportunities. We go on to provide evidence suggesting that one's own trustworthiness can be traced back to the values parents transmit to their children during their upbringing.

JEL Classification: A1, A12, D1, Z1

Keywords: Trust, trustworthiness, culture, false consensus.

---

[*] Submitted October 2012, major revision July 2013, accepted March 2014.

[1] We thank Guido Menzio and two anonymous referees for comments that substantially improved the paper. We are also grateful to seminar participants at the Bank of Spain, the CCPR at UCLA, the Einaudi Institute for Economics and Finance, the Kaler Meeting at UCLA, the FEEM conference on the Economics of Culture, Institutions and Crime, the EALE/SOLE joint conference in London, the 9th IZA/SOLE Transatlantic Meeting of Labor Economists, the seventh International Meeting on Behavioral and Experimental Economics, the Higher School of Economics in Moscow, the London School of Economics, the University of California Davis, the NBER Political Economy Meeting, University of Mannheim, Universidad Pompeu Fabra, University of San Diego, University of Siena, Stanford University and Toulouse University for helpful comments. Luigi Guiso thanks EIEF for financial support and LUISS University for making the LUISS lab available. Paola Giuliano thanks the UCLA-CIBER grant for financial support and the Russell Sage Foundation for its wonderful hospitality.

1. Introduction

Many decisions rely upon our beliefs about the trustworthiness of another person about whom we know little or nothing about. Call these beliefs "trust beliefs." The large body of literature studying trust beliefs has documented a remarkable persistence of the distribution of these beliefs across generations using evidence from different datasets and a wide variety of countries.[2] At the same time, trust beliefs have also been shown to be quite heterogeneous across individuals.[3] In this paper we provide evidence suggesting that false consensus, the tendency of individuals to extrapolate from their own type the behavior of others (Ross, Green and House, 1977), can explain these dual patterns.

Persistent heterogeneity in trust beliefs, even in the same community, has been explained in the literature in various ways. According to one view, individuals' beliefs are initially acquired through cultural transmission and then slowly updated through experience from one generation to the next. This line of argument has been pursued by Guiso, Sapienza and Zingales (2008b) who build an overlapping-generations model in which children absorb their trust priors from their parents and then, after experiencing the real world, transmit their (updated) beliefs to their own children. Dohmen et. al (2012) provide evidence consistent with this view. Heterogeneity is the result of family specific shocks. Within a generation, correlation between current beliefs and received priors is diluted as people age and learn. Yet this dilution needs not to be complete and a high degree of persistence may still obtain.

[2] See Algan and Cahuc (2010), Butler, Giuliano and Guiso (2012a), Dohmen et al. (2012), Guiso, Sapienza and Zingales (2008a).

[3] Butler et al. (2012a,b) and Dohmen et al. (2012)

On the other hand, a slightly different explanation is that parents instill values, such as trustworthiness, rather than beliefs. Cultural transmission of values of cooperation and trustworthiness is the focus of Bisin and Verdier (2000), Bisin, Topa, and Verdier (2004) and Tabellini (2008). They show how norms of behavior are optimally passed down from parents to children and persist from generation to generation. Heterogeneity in parents' preferences and experiences may then result in heterogeneity in instilled trustworthiness. Even if parents do not teach beliefs directly, individuals may extrapolate from their own type when forming beliefs about others' trustworthiness. As Thomas Schelling once wrote "you can sit in your armchair and try to predict how people behave by asking yourself how you would behave if you had your wits about you. You get free of charge a lot of vicarious empirical behavior" (1966, p. 150).

In this paper we show that false consensus is one mechanism that could help to explain how heterogeneity in values could translate into heterogeneity in beliefs. In line with the view voiced by Schelling in the quote above, we view false consensus as a source of individuals' initial trust beliefs, or trust priors. Specifically, in the absence of a history of information about the reliability of a pool of people with whom one may potentially interact, individuals may form their trust priors by asking themselves how they would behave in similar circumstances. Since different individuals may behave differently, this process of mental simulation can lead to heterogeneous trust priors. If the effect of false consensus on trust priors and, consequently, on subsequent trust beliefs vanishes only slowly with learning about the population, then miscalibrated trust beliefs will also persist. In our context false consensus implies that highly trustworthy individuals will tend to think that others are like them and form overly optimistic trust beliefs, while highly untrustworthy people will extrapolate from their own type and form excessively pessimistic beliefs. Both highly trustworthy and highly untrustworthy individuals

will tend to systematically form more extreme trust beliefs than are warranted by their experiences. A long history of research on false consensus has indeed shown it to be a persistent phenomenon (Krueger and Clement, 1994) that need not be drowned out by monetary incentives for accurate predictions (e.g. Massey and Thaler, 2006).

To show the relevance of false consensus for trust beliefs we conduct two experiments. The first experiment implements a repeated version of the standard trust game in the laboratory (Berg, Dickhaut and McCabe, 1995).[4] The experiment allows us to obtain a measure of participants' own (initial) trustworthiness and also to elicit participants' trust beliefs after each round of game play. We first document a strikingly high correlation between participants' trust beliefs and their own trustworthiness, suggesting that beliefs are formed by extrapolating from one's own type. Moreover, we show that this correlation remains strong and significant even after several rounds of game play. Finally, we investigate one plausible source of initial trustworthiness, and hence trust priors: parentally instilled values. By complementing the experimental data with a survey-based measure of the values our participants' parents emphasized during their upbringings, we provide evidence suggesting that initial trustworthiness varies substantially with the emphasis participants' parents placed on good behavior.

Because particular features of the canonical trust game make it poorly suited to studying the relationship between earnings and trust beliefs, we conduct a second experiment featuring a slightly modified trust game in order to investigate the economic consequences of false consensus.[5] We show that it is indeed the case that the most (least) trustworthy participants tend to form overly-optimistic (overly-pessimistic) trust beliefs and, consequently, trust more (less)

[4] See the literature review for a detailed description of the trust game.

[5] We provide details of this argument when we discuss the design of Experiment 2.

than they should. Participants with miscalibrated beliefs earn substantially less in the process: mistaken trust beliefs can cost participants up to 18% of their experimental earnings, on average, compared to the earnings of participants with properly-calibrated beliefs.

The remainder of the paper proceeds as follows. In Section 2, we review the literature on the trust game. In Section 3, we describe Experiment 1, which is designed to document the impact of false consensus on trust beliefs as well as the persistence of this impact, and present results. We also discuss alternative interpretations of our main findings and provide further evidence consistent with our interpretation. In Section 4, we motivate and provide details on the design of Experiment 2 and present results suggesting that the economic costs of false consensus can be substantial. In section 5, we summarize our findings and present concluding remarks.

2. Closely Related Literature

The trust game is a two-player sequential moves game of perfect information in which the first-mover ("Sender") is endowed with some fixed amount of money, $X > 0$, and chooses how much of this endowment to send, $0 \leq S \leq X$, to the second-mover ("Receiver"). Any money sent is increased by the experimenter according to a commonly known function, $f(S)$, before being allocated to the Receiver. The Receiver then chooses to return any amount, $0 \leq R \leq f(S)$ to the Sender, ending the game. With purely own-money-maximizing players, the Subgame Perfect Nash Equilibrium (SPNE) of this one-shot game is simple and straightforward: Receivers never return any money and, consequently, Senders never send any money.

The experimental trust game literature begins with Camerer and Weigelt (1988), who use a repeated version of the game to study reputation. The trust game appears in what has come to be its canonical form---e.g., one-shot, with $f(S) = 3s$--in Berg, Dickhaut and McCabe (1995).

Subsequent experimental trust game research is voluminous and spans many disciplines.[6] In contrast to behavior predicted in the unique SPNE, a recent meta-analysis of over 160 trust game experiments conducted across several disciplines and cultures documents three prevalent and robust patterns: i) most receivers return a strictly positive amount that allows senders to make a small positive expected profit from sending $S > 0$; ii) the proportion receivers return is increasing in the proportion of the sender's endowment sent; iii) most senders send $S > 0$, with a strong tendency to send about half of their endowment (Johnson and Mislin, 2011). Accordingly, the extent to which a sender's or a receiver's behavior differs from the purely self-regarding SPNE prediction above---i.e., senders sending $S > 0$ or receivers returning $R > 0$---has come to be used as a measure of trust or trustworthiness, respectively.[7]

Our paper contributes to two different strands of the trust literature. A series of studies has suggested that beliefs about others' behavior is important for own behavior. Buchan, Croson and Solnick (2008) show that trust beliefs and trust behavior are positively correlated---senders who expect more back send more. Costa-Gomes, Huck and Weizsäcker (2010) use an

[6] For a more exhaustive overview see Camerer (2003) and the references therein.

[7] The interpretation of sender behavior as a measure of trust is controversial, however. Early on, Glaeser et al., (2000), documented that sender behavior is not correlated with alternative survey-based measures of trust, such as those appearing in the World Values Survey. Cox (2004) decomposed the trust game into strategically simpler constituent games and showed that senders' behavior depends, at least partially, on factors besides trust (e.g., altruism). Partially reconciling this conundrum, trust game research has recently begun to focus on trust beliefs---beliefs about receivers' trustworthiness---as a less problematic measure of an individual's underlying propensity to trust. Integrating a trust game experiment into a large-scale survey, Fehr et al. (2003) show that trust beliefs are correlated with a standard survey measure of generalized trust; Sapienza, Toldra-Simats and Zingales (forthcoming) document a similar finding by incorporating a measure of generalized trust typically used in surveys into a trust game experiment.

instrumental variables approach to show that most of this correlation is causal: senders send more because they expect more back. Bellemare and Kröger (2007), using trust game data from samples of Dutch students and adults, find a positive relationship between senders' beliefs about other senders' behavior and senders' own behavior. They interpret this correlation as evidence of the importance of social norms; the pattern is also consistent with the presence of false consensus. Taken together, all of these studies suggest that beliefs about others' behavior are important for own behavior. Many of the studies suggest that trust beliefs, in particular, are important for trust behavior. Our study takes one step back and asks how trust beliefs themselves are determined, suggesting and providing evidence consistent with one plausible mechanism: false consensus.

We also contribute to the literature on the evolution of preferences. Güth, Kliemt, and Peleg (1998) present a theoretical model of the coevolution of preferences and information in a simplified, binary-choice, trust game, showing that the distribution of trustworthy types in the long run is sensitive to whether information gathering by senders is modeled as a choice. Bohnet, Frey and Huck (2001) posit and experimentally investigate an alternative model of preference evolution in a repeated binary-choice trust game setting in which types are dichotomous, fixed and common knowledge. They examine how evolutionary forces interact with institutions to determine the long-run distribution of trustworthy types, showing that when contract enforcement institutions are weak trustworthy types eventually predominate. When institutions are strong, both types exist in the long run. In the intermediate case, untrustworthy types earn more on average, implying that they become predominant in the long run. In Bohnet and Huck (2004), the authors study how information on participants' past play (reputation) interacts with the institutional setting---fixed-pairs, random matching or an intermediate case---in determining

behavior in a binary-choice trust game. They find that an institution that allows direct reputation building---fixed-pairs matching---increases trusting and trustworthy behavior, as predicted by standard theory, but surprisingly, in contrast to theoretical predictions, these beneficial effects persist even when the institution is exogenously changed to be less favorable to reputation-building. We contribute to this literature by showing that trust beliefs are quite persistent, possibly even being transmitted across generations indirectly along with the values parents teach to their children through a well-established psychological mechanism: false consensus.

3. False Consensus, Values and Persistence

3.1. Experiment 1: Design and Procedures

Participants were recruited from a pre-existing list of students who had previously expressed a general willingness to take part in experiments at LUISS Guido Carli University in Rome, Italy. All laboratory sessions were conducted at CESARE, the lab facility at LUISS. The experiment was programmed and implemented using the software z-Tree (Fischbacher, 2007). In total, 124 students participated in Experiment 1.

After showing up to the lab at pre-scheduled session times, participants were seated at individual desks in the lab. Seating was randomly assigned and each desk was equipped with its own computer. Participants were separated from one another by opaque dividers, effectively creating individual private cubicles. Once all participants were seated, instructions were read aloud and participants' questions, if any, were answered by the experimenters.

After instructions were read and questions were answered, subjects proceeded to the game-playing phase. This phase consisted of up to twelve rounds of the trust game, as described below. Participants were not informed how many rounds of game-play there would be, but rather only instructed that there would be several rounds. This was meant to minimize end-game effects

possible when the number of rounds is known. Because sessions were scheduled to last (up to) two hours, and because most participants had never participated in any experiment before (CESARE is a relatively new facility) the number of rounds per session varied widely. Sessions consisted of anywhere from 3 to 12 rounds, with the majority consisting of 12 rounds.

Before each round, each participant was randomly and anonymously matched with a co-player, and, within each pairing, roles were randomly assigned. That is to say, neither pairings nor roles persisted across rounds. These design features allow for learning about the population's traits and preferences but not about any specific person's traits/preferences. They also serve to ameliorate many repeated-game effects that are possible when partners are uniquely identifiable or persist over rounds, such as reputation building or directly punishing/rewarding specific partners for past behavior, as such effects are not the focus of this experiment.

Within each randomly-formed pairing, participants played a standard trust game[8] where the sender is endowed with 10.50 euros and the receiver is given no endowment. The sender chooses to send some, all or none of his or her endowment to the receiver. Any amount sent is tripled by the experimenter before being allocated to the receiver. The receiver then chooses to return some, all or none of this tripled amount back to the sender, ending the game. Sending a positive amount entailed a small fee---0.50 euros.

Feasible actions for the sender in our implementation were to send any whole-euro amount: 0,1,2,…,10. Receivers' decisions were collected using the strategy method. Before receivers discovered how much their sender sent, they specified how much they would return for any amount of money they could receive. One critique of the strategy method is that it is "cold" and does not elicit the same reaction as if participants are faced with an actual decision. To partially address this critique, and make receivers' decisions feel as real as possible, receivers

[8] See the general description of the trust game in the literature review section above.

were faced with a series of ten separate screens. Each screen asked only one question: "if you receive m euros, how much will you return?" For each separate screen, m was replaced with exactly one value, $m \in \{3, \dots, 30\} = \{3 \times 1, \dots, 3 \times 10\}$. The order of possible amounts, m, was randomized in order to avoid inducing any artificial consistency in receivers' strategies. This random order was the same for all receivers within each round, and was re-randomized between rounds. Obviously, no information about receivers' decisions was shared with senders in any way before the end of each round.

At the end of each round, each sender and receiver pair was informed of the outcome of their interaction---i.e., how much the sender sent, and, if this was a positive amount, how much the receiver returned as determined by the relevant element of the receiver's strategy vector. No other element of the receiver's strategy vector was revealed, nor was any information about the outcome in any other participant pair.

To collect beliefs, within each round every participant---regardless of the role they had been assigned---was asked to estimate the amounts receivers would return, on average, for each possible amount receivers could receive. Specifically, participants answered ten questions: "How much will receivers return, on average, if they receive m euros?", $m \in \{3, \dots, 30\}$. Participants who were currently receivers were told to exclude their own actions from this estimate and that they would be remunerated on this basis. That is to say, they were asked to estimate how much other receivers would return. This serves to rule out any mechanical---real or imagined---connection between participants' own actions and their estimates.

Incentives to report beliefs truthfully were given by paying subjects according to a quadratic scoring rule.[9] Beliefs were elicited either before or after participants submitted their actions, with this order being randomly determined for each participant before each round.

When all rounds were completed, one round was selected at random and participants were paid in accordance with their actions and the accuracy of their estimates in that round. This procedure is meant to eliminate wealth effects from accumulated earnings over rounds and it is standard in the literature. All of these design elements were commonly known by all participants.

3.2. Our Measures of Trust Beliefs and Trustworthiness

Because we elicited trust beliefs and receivers' behavior for all possible send amounts, we can investigate the relationship between trustworthiness and trust beliefs at various levels of aggregation. We follow much of the literature in assuming that participants' return proportions in the role of receiver are a measure of trustworthiness. At all levels of aggregation, we will make

[9] It is well-known that this rule gives (risk-neutral) individuals incentives that are compatible with reporting truthfully the mean of their subjective distribution of beliefs. Specifically, for each of the ten belief questions participants earned an amount of money given by the function below, where $\widehat{r_m}$ is a participant's stated estimate of receivers' average return amount conditional on receiving m euros, r_m is receivers' actual average return amount conditional on receiving m euros and, as above, m ∈ {3, ...,30}:

$$Earnings = 1 - \left(\frac{\widehat{r_m} - r_m}{m}\right)^2$$

For example, if a participant's estimate of receivers' average return amount, conditional on receiving 9 euros, was 6 euros--i.e. $\widehat{r_9} = 6$,---and receivers' strategy vectors entailed returning 2 euros, on average, conditional on receiving 9 euros, then that participant's estimate would earn the participant (in euros)

$$1 - \left(\frac{6-2}{9}\right)^2 = 1 - \frac{16}{81} \approx 0.80$$

A perfect estimate would obviously pay 1 euro, so that subjects could earn up to 10 euros in total each round from their ten estimates.

use of a measure of initial trustworthiness largely untainted by learning. Specifically, we assign to participants their relevant trustworthiness measure from the first time they played as a receiver, provided this occurred in one of the first two rounds.[10] Since roles are randomly re-assigned each round, this measure is defined for a large majority of participants, but not all of them (92 of 124).

For ease of exposition, we will make extensive use of summary, unidimensional, measures of trustworthiness and trust beliefs. To construct a summary measure of trust beliefs for each participant, we take the average over all ten elements of his or her vector of return proportion beliefs. For example, suppose a participant's vector of beliefs about the amounts receivers will return on average is (1,2,…,10)---i.e., he or she believes that receivers will return an average of 1 if they receive 3×1=3, an average of 2 if they receive 3×2=6, etc. We would divide the first element by 3, the second element by 6 and the n^{th} element by 3n to get a vector of return proportion beliefs $(\frac{1}{3}, \frac{2}{6}, \ldots, \frac{10}{30})$. Finally, we would take the average over all ten elements of this vector to get a single number---here, (1/3), or 0.33---as a unidimensional measure of this participant's trust belief. We construct a summary measure of initial trustworthiness in an analogous fashion: we first convert participants' initial return amount vectors into initial return proportion vectors, dividing the n^{th} element by 3n; averaging over all ten elements of this initial return proportion vector yields a scalar measure of initial trustworthiness.

[10] The choice of the first two rounds balances two concerns: i) contamination by learning which suggests only including those who were receivers in the first round---and leaving the measure undefined for half of the participants; ii) concerns about sample size which suggest extending the definition to include as many rounds as possible. In the end, we believe our definition is reasonable.

Next, since false consensus also predicts a positive relationship between one's own actions and beliefs about others' actions more generally, we examine the relationship between receivers' own strategies and beliefs about others' strategies at a more disaggregated level. We do so by examining the relationship between own initial return proportions and beliefs about others' return proportions for each possible amount a receiver could receive, separately.

Receivers' behavior can be an expression of both reciprocity and baseline trustworthiness. In this context, a natural question to ask is whether false consensus applies to beliefs about others' reciprocity, beliefs about others' baseline trustworthiness or both. To answer this question, we model the receiver's return proportion function in a linear format, $\mathrm{r(s)} = \mathrm{ms} + \mathrm{b}$, where s denotes the amount the sender chooses to send and $\mathrm{r(s)}$ denotes the receiver's return proportion conditional on s. The slope term, m, can be thought of as a rough measure of reciprocity, while the intercept term, b, may serve as a measure of baseline trustworthiness.

We therefore estimate the best linear fit of each participant's (initial) trustworthiness vector using ordinary least squares. In particular, we estimate the equation $r_i(s) = m_i s + b_i + \varepsilon_i$ for each participant, i, by regressing return proportions on send amounts. Since we used the strategy method, this estimate is based on ten observations for each receiver: one for each send amount $\mathrm{s} = 1, \ldots, 10$. By running these regressions, we are able to obtain individual-specific slope and intercept estimates, $\widehat{m_i}$ and $\widehat{b_i}$, which we interpret as measures of a participant's reciprocity and baseline trustworthiness, respectively.

Using the same procedure, for each period we also estimate the best linear fit of each participant's trust beliefs vector using ordinary least squares. In particular, we estimate the equation $\tau_i(s) = m_i^e s + b_i^e + u_i$ for each participant, i, by regressing trust beliefs (τ_i) on send amounts. Because we collected beliefs for each send amount, separately, each regression

incorporates ten observations for each participant in each round: one for each send amount $s = 1, \ldots, 10$. By running these regressions, for each period we obtain individual-specific slope and intercept estimates, $\widehat{m_i^e}$ and $\widehat{b_i^e}$, which we interpret as within-period beliefs about others' reciprocity and baseline trustworthiness, respectively. To investigate whether there is false consensus in reciprocity or trustworthiness, we express expected reciprocity, $\widehat{m_i^e}$, as a function of own reciprocity, $\widehat{m_i}$, and expected baseline trustworthiness, $\widehat{b_i^e}$, as a function of own baseline trustworthiness, $\widehat{b_i}$.

3.3. Parentally-Instilled Values

Finally, all participants filled out a brief survey which they received by e-mail. The survey was sent either several days before or several days after the participant's specific laboratory session in order to mitigate the concern that taking the survey could systematically affect behavior in the lab through, e.g., priming. One part of the survey asked respondents to report, on a scale from 0 to 10, how much emphasis their parents placed on a number of principles and behavioral rules during their upbringing (frugality, prudence, loyalty, etc.).[11] We use answers from a subset of these questions to construct a measure of the strength of received cultural values and norms of trustworthiness for each participant.[12]

[11] A wide array of questions was asked, some completely irrelevant to trust and trustworthiness, in order to mitigate experimenter/demand effects in the survey answers and in the experiment.

[12] We acknowledge that such self-reported retrospective questions are likely to be noisy measures of the values our participants' parents actually emphasized. For example, individuals may selectively remember some lessons and not others, biasing their recollection of what their parents taught them. Unfortunately, our data do not allow us to address this criticism directly since we do not survey our participants' parents. However, it is reasonable to assume such self-reports convey some information about the values our participants believe their parents transmitted to them, which should lend some credence to our interpretation of them as received cultural values.

3.4. Results

3.4.1. Heterogeneity in Trust Beliefs and Trustworthiness

We start with our simplest measures of trust and trustworthiness. Figure 1 shows the distribution of our unidimensional measure of trust beliefs in the first round of the trust game, before any learning about the trustworthiness of the pool of participants had yet been possible (panel A) and of our summary measure of own initial trustworthiness (panel B). Since trust beliefs and trustworthiness are measured in terms of the proportion of the amount received that participants expect receivers will send back, and by the average proportion that receivers are willing to send back, respectively, these variables take values between 0 and 1. Because these measures are continuous variables we report kernel density estimates.

The top panel of Figure 1 documents considerable heterogeneity in initial trust beliefs, or, trust priors. Since beliefs in the experiment refer to a common pool of people, heterogeneity in trust beliefs cannot be automatically ascribed to variation in the pools of people whose trustworthiness is being estimated, which is a common critique of standard survey-based measures of trust beliefs such as those appearing in the World Values Survey.[13] Furthermore, since beliefs are measured independently of behavior, the heterogeneity in Figure 1, panel A, cannot reflect differences in attitudes toward risk.[14] In the sample the average level of trust priors

[13] It is true that Figure 1, panel A, reports trust priors for all sessions pooled, so some people might still question the source of heterogeneity. However, plotting the trust belief densities for each session separately (not reported, but available upon request) also yields quite a lot of heterogeneity.

[14] Unless the elicitation procedure is biased by risk preferences as well. We cannot rule this out completely, as how to do so is a still-unsettled debate within experimental economics. We use a very standard quadratic scoring rule. There is experimental evidence suggesting that this mechanism, in practice if not in theory, elicits beliefs reasonably accurately regardless of risk preferences (see, e.g., Huck and Weiszäcker, 2002).

is 0.27 and the sample standard deviation is 0.16.[15] The figure (bottom panel) also documents substantial heterogeneity in own initial trustworthiness, whose sample mean and standard deviation are 0.32 and 0.16, respectively. Thus, while dispersion in trust priors and trustworthiness is similar, beliefs about others' trustworthiness are initially more pessimistic than warranted on average. Table 1 shows summary statistics.

[Figure 1]

[Table 1]

3.4.2. False Consensus and Persistence

Considering initial trustworthiness and the evolution of trust beliefs over rounds, we next investigate whether heterogeneity in trustworthiness is reflected in heterogeneous trust beliefs and to what extent learning about the population dampens this relationship. We begin, again, with our summary measures of trust beliefs and initial trustworthiness. In Table 2, panel A, we report regressions of trust beliefs on trustworthiness across several rounds. To isolate, as best as possible, trustworthiness as an individual trait, we use initial trustworthiness as a regressor. To reduce sampling variation due to small sample size we aggregate observations over blocks of three rounds. Since this results in possibly multiple observations per participant within each three-round block, we cluster standard errors by individual.

As the first column shows, in early rounds initial trustworthiness is strongly positively correlated with trust beliefs, lending support to the idea that individuals form beliefs about others' trustworthiness by extrapolating from their own types.[16] Quite remarkably, own

[15] Since every dollar sent is tripled, 0.33 would imply senders believe that receivers will return as much as is sent.

[16] We acknowledge that the other direction of causation is also possible, i.e., that individuals form beliefs about others' behavior and then seek to conform their own behavior to others' behavior. While our data do not allow us to

trustworthiness explains about 60% of the initial heterogeneity in beliefs. As the second column shows, this tendency does not vanish when the game is repeated and people are thus given the opportunity to learn about the pool of participants. The correlation weakens, and the effect is somewhat smaller, in later rounds but both remain sizable and significant. Thus, initial trustworthiness still affects trust beliefs even after the game has been played several times, always drawing from an invariant pool of individuals, which we take as evidence that false consensus persists. However, the decline in the strength of the link also suggests that given enough opportunities to learn about a stable pool of people, the tendency to attribute to others one's own trustworthiness may vanish.[17]

One concern with the regressions using our summary measures of trust beliefs and trustworthiness is how consistent the apparent false consensus effect is over send amounts. It could be that false consensus is particularly pronounced for only a few send amounts and non-existent for all others so that the correlation between our summary measures overstates the prevalence and generality of false consensus. To address this concern, in Table 2, Panel B, we regress trust beliefs on own initial trustworthiness for each send amount, separately, using the same three-round blocks as above.[18] We find that, although false consensus tends to be more pronounced for higher send amounts, it is nonetheless both statistically significant and of substantial magnitude across all possible amounts a sender could send. Moreover, the bias in

---

fully rule out this possibility, in the Appendix we provide evidence suggesting that behavior and beliefs in our experiment are more consistent with false consensus than with a preference-for-conformity story.

[17] An interesting question which is beyond the scope of the current study is whether the false consensus effect reappears any time an individual faces a new pool of people or the pool she is interacting with changes.

[18] To account for possibly multiple observations per participant in each three-round block, we cluster standard errors by individual.

false consensus towards high send amounts fades over rounds so that in later rounds the coefficients on initial trustworthiness are roughly equal across all send amounts, taking on a value of about 0.50.[19]

Next, we consider the view noted above that receivers' trust game behavior can be described in terms of two phenomena: a baseline measure of trustworthiness which is modified by reciprocity. This view lends itself naturally to a two-parameter description of receivers' return proportion vector, one common example of which obtains by linearizing. Consequently, in Table 2, Panel C, we report simple regressions of expected reciprocity on own reciprocity and expected baseline trustworthiness on own baseline trustworthiness, separately.[20] We find evidence consistent with false consensus for both baseline trustworthiness and reciprocity. The correlations between one's own behavior and expectations of others' behavior are in fact quite similar within each three-round block of observations for both baseline trustworthiness and reciprocity. As with our analysis above, we again find that while each of the correlations declines over rounds with learning, they remain both substantial in magnitude and highly statistically significant into later rounds. Even in the last block of rounds considered, for instance, one's own initial reciprocity (baseline trustworthiness) accounts for about 30% of the variation in beliefs about others' reciprocity (baseline trustworthiness).

[19] One story consistent with this pattern over rounds is that high send amounts are less common, so that, initially, trust beliefs for higher amounts are based on less true information and more introspection. Consistent with this story, conditional on a strictly positive send amount, send amounts weakly less than 6 euros account for 67% of the data in both earlier rounds (1 to 6) and later rounds (7 to 12).

[20] We again cluster standard errors by individual to account for the possibility of multiple observations per participant in each three-round block.

[Table 2]

Finally, having computed both a two-parameter measure and a summary, unidimensional, measure of initial trustworthiness, the question arises: which component of trustworthiness does the summary measure capture? The summary measure is often used for convenience or ease of exposition, so that having some indication of what it tells us may prove useful even beyond the current study. Fortunately, the answer is clear: our summary measure of initial trustworthiness is highly significantly correlated with baseline initial trustworthiness ($\rho = 0.83, p < 0.01$), while its correlation with our initial reciprocity measure is essentially zero ($\rho = -0.03, p > 0.80$).[21] The same is true for trust beliefs.[22] This suggests that if our main interest is in trustworthiness and beliefs about others' trustworthiness, and not in reciprocity per se, little is lost by using our relatively simple summary measures in lieu of the two-parameter formulation.

3.4.3. Trust Beliefs, Trustworthiness and Values

The evidence presented so far is consistent with the idea that trust priors are driven, through false consensus, by individuals' own levels of trustworthiness. An intriguing further conjecture is that trustworthiness is determined by the values parents transmit, so that to the

[21] This need not be the case. As the clearest counter-example, assume all participants' return proportion functions are upward sloping rays from the origin. The intercept term (baseline trustworthiness) would always be zero so that baseline trustworthiness would be uncorrelated with the summary trustworthiness measure; at the same time, the slope term (reciprocity) would be positively correlated with the summary trustworthiness measure.

[22] The correlation between the summary measure of trust beliefs and the intercept term for linearized trust beliefs is at least 0.75 in 11 of the twelve rounds, taking on a low of 0.67 in round 9, and is always highly statistically significant ($p < 0.01$ in every round); on the other hand, the correlation between the summary measure of trust beliefs and the slope term for linearized trust beliefs is only significant in one of the twelve rounds (round 4: $\rho = 0.23; p = 0.02$) and is typically small in magnitude, ranging from -0.13 to 0.13 in all other rounds.

extent values are stable across generations so too may the distribution of trust beliefs persist. To shed light on this conjecture we take advantage of our participants' recollections of the moral values emphasized by their parents during their upbringings. For our purposes, we use parents' emphasis on two values: the first is how much emphasis an individual's parents placed on always behaving like a good citizen; the second is the emphasis parents placed on loyalty to groups or organizations. We average the responses to these two questions and divide the result by 10 to get a measure of received cultural values on a scale from 0 to 1---a scale comparable with beliefs. For ease of exposition, we refer to the resulting index as simply "good values."

Before proceeding, let us note that there are several factors stacked against finding any relationship between good values and beliefs or behavior. The first is a statistical artifact: the measure of received cultural values we have is, at best, a noisy measure of transmitted cultural values, being retrospective and self-reported. Secondly, participants' own values are likely to reflect not just the values their parents transmitted, but also, to varying degrees, values acquired through other channels of socialization which we do not measure. Finally, the identifying variation in receivers' behavior is attenuated by the nature of the pecuniary incentives involved which pull in only one direction---toward uniformly returning nothing. With all of these factors working against us, we would view finding a significant positive correlation between good values and behavior or beliefs, regardless of the level of significance, surprising.

We begin by investigating the link between values and trustworthiness. In Table 3 we regress, at various levels of aggregation, initial trustworthiness on good values. We start with our summary, unidimensional, measure of initial trustworthiness (Panel A). At this most aggregated

level, we find a positive relationship between initial trustworthiness and good values.[23] While the formal statistical significance of this relationship is not high, the magnitude is substantial---increasing good values from 0 to 1 is associated with an increase in trustworthiness of more than 50% of the unconditional sample mean. In Panel B, we find that the estimated effect of good values on initial trustworthiness is consistent across send amounts: the estimated coefficients on good values are always positive and never far from the summary-measures point estimate of 0.17. Focusing next on the two-parameter decomposition of trustworthiness (Panel C), we find the effect of good values is concentrated on baseline trustworthiness: the estimated coefficient is identical to the coefficient estimated using our summary measures above;[24] while, at the same time, we find literally zero relationship between good values and initial reciprocity.

[Table 3]

Turning from behavior to beliefs, we investigate next the link between culturally received values and trust beliefs directly. In Table 4, we regress trust beliefs on good values, considering how the relationship evolves over rounds. Starting with our summary measure of trust beliefs (Panel A), we find that good values do indeed predict trust beliefs. The estimated positive relationship between values and beliefs is statistically significant in all but the last three-round block as well as sizable in magnitude. The coefficients from the first three three-round blocks---

[23] One might worry that this correlation simply reflects priming participants to think about morality by the mere fact of answering the survey. If so, one would expect the correlation to be particularly strong for participants who took the survey before their experimental session. We check for this by splitting the sample into those who took the survey *before* their session and those who took the survey after their session. The correlation between good values and initial trustworthiness is positive in both subsamples, but is larger in the subsample of those who took the survey *after* the experiment suggesting priming is not the primary driver of the correlation.

[24] The significance of this effect just misses being significant: ($p = 0.104$).

0.12, 0.13 and 0.12---imply that increasing good values from 0 to 1 is associated with an increase in trust beliefs equal to 38%, 48% and 55% of the unconditional sample mean of trust beliefs in the relevant three-round block. Considering this relationship over each send amount separately (Panel B), we find that the magnitude of the correlation between good values and trust beliefs is largely uniform in the earliest block of rounds, always hovering around 0.12, but that the significance of the relationship is higher for higher send amounts. As with the summary measure, the magnitude of the relationship between good values and trust beliefs remains roughly the same over the subsequent two three-round blocks, send amount by send amount, before vanishing in the last block of rounds. In Panel C, we consider the two-parameter decomposition of trust beliefs, regressing beliefs about reciprocity and baseline trustworthiness on good values, separately. Similar to our findings above, the data suggest that good values do not affect beliefs about reciprocity. On the other hand, the estimated impact of good values on expected baseline trustworthiness is virtually identical to the estimated relationship between good values and our summary trust beliefs measure.

[Table 4]

All together, the data suggest a consistent and persistent relationship between culturally received good values, trustworthiness and, ultimately, trust beliefs. Interestingly, the patterns over rounds and across send amounts suggest that the effect of values on trust beliefs in particular is strongest when there is the least information---for high send amounts and for earlier rounds---which is consistent with a false consensus explanation in which introspection substitutes for observation.

Summing up, the evidence from Experiment 1 suggests three things. First, when no information is available about a group, individuals may form beliefs about the trustworthiness of

others by extrapolating from their own types, which are quite heterogeneous. Second, this tendency is highly persistent, though attenuated through learning. And third, one source of heterogeneity in own initial trustworthiness, and hence trust priors and subsequent trust beliefs, may be heterogeneity in the cultural values transmitted by parents. This last pattern, if verified and replicated in further studies, implies that measures of culturally transmitted values could prove to be valuable instruments for trust beliefs---an implication which may prove useful in empirical studies on the effects of trust beliefs.

3.4.4. Trust Beliefs, Trustworthiness and Values

We briefly describe here some potential alternative interpretations of our main findings and provide evidence consistent with our interpretation of the results. Full details are provided in the Appendix.

The first concern is related to the possibility that trust beliefs do not measure expected trustworthiness, but rather expected rationality. For example, suppose rational individuals understand that returning nothing is a dominant strategy in the role of receiver. However, suppose rational individuals also understand that slightly less rational individuals, who may fail to understand that returning nothing is a dominant strategy and may, therefore, mistakenly return some positive amount, are also playing the game. For such fully rational individuals, beliefs about expected return proportion may then reflect not just beliefs about others' trustworthiness, but also beliefs about the proportion of less rational players in the population.

Ideally one would like to re-run our experiment on a population which would a priori be expected to be as rational, strategically sophisticated and purely money-maximizing as possible and where these features of the population are common knowledge among the participants. Exactly this type of exercise has been conducted by Sapienza, Toldra-Simats and Zingales

(forthcoming). There, the entire incoming class of MBA students at the University of Chicago's Booth School of Business were enlisted to play a trust game for substantial stakes. As part of the experiment, senders' beliefs about receivers' return proportions were elicited. Reassuringly, senders' beliefs about receivers' behavior in even this exceedingly rational population looked quite similar to our participants' beliefs (see Appendix, Table A1).

We also address this concern, although more indirectly, using our own data. We make use of the following conjecture. If beliefs about return proportions reflect the expected reasoning ability of the population, one may expect two patterns: i) receivers with lower reasoning ability may be less likely to recognize that always returning nothing is a dominant strategy so that we might expect a negative relationship between return proportions and cognitive ability; and ii) senders with more reasoning ability may be more likely to anticipate that some receivers will misunderstand the game and return more than nothing, so that higher-ability senders may report higher return proportion beliefs. We test for both of these conjectures using participants' scores on a standardized math test required of Italian high school students as a proxy for cognitive or reasoning ability. We regress both return proportions and beliefs about others' return proportions on this proxy for reasoning ability (see Appendix, Tables A2 and A3). We find no relationship between cognitive/reasoning ability and trust beliefs or trustworthy behavior.

The second concern is related to the relationship between trust beliefs and behavior. The correlations we document between one's own trustworthiness and trust beliefs are consistent with a false consensus story in which own type/behavior affects beliefs. However, they are also consistent with a story featuring the opposite direction of causation: it could very well be that individuals first form beliefs about others' behavior and then try to conform their own behavior to these beliefs. Unfortunately, our data do not allow us to prove which direction of causation is

at work. This would require showing that exogenous variation in trustworthiness causes trust beliefs to change, and we have no such exogenous variation. In Section 1.1 of the Appendix, however, we provide evidence suggesting that our data are inconsistent with a reverse causation story. In particular we study one specific aspect of reverse causality: that the correlation is driven by a desire to conform to the empirical norm. We model such norm-compliant preferences in a simple reduced form way by assuming that overall utility is a weighted average of utility from money outcomes and disutility from deviating from the norm. We show that one strong implication of this model is that receivers should never return strictly more than they believe others will return if their behavior is driven by conformity to the norm. A false consensus story, on the contrary, does not have such an implication. We test this implication of the norm-compliance story and find that it is not consistent with our data: a non-trivial fraction of receivers (between 20 and 30 percent) returns strictly more than they report believing other receivers will return.[25]

An additional concern that one might have is how robust our results are to putting more weight on actually chosen send amounts. For example, in Experiment 1 where participants play only one role in each round, a sender's stated beliefs about amounts he or she chooses not to send are, in a sense, less consequential as they relate to outcomes which will never occur. One might worry that giving too much weight to such "less consequential" beliefs in our summary trust beliefs measures could call into question their legitimacy as an indicator of some underlying

[25] In Figure A1 of the appendix, we show scatter plots for several send amounts of receivers' return amount versus receivers' beliefs about other receivers' return amounts. If conformity is the primary driver of the behavior/beliefs correlation, we would expect very few points below the 45 degree line. In stark contrast to this prediction, evident from the scatter plots is the existence of a substantial proportion of observations in which return proportions are strictly greater than beliefs about other proportions (Appendix, Table A4).

individual propensity to trust. To address this concern, for Experiment 1 we restrict our attention to participants playing the role of sender and construct an alternative summary measure of their trust beliefs for each round, separately. This alternative measure places 80% of the weight on the trust belief related to the sender's chosen send amount and spreads the remaining 20% of the weight evenly over the remaining return proportion beliefs. The details on how this measure is constructed are provided in the Appendix (section 1.2). We show that when we re-calculate our main results using this alternative trust beliefs measure, we find estimates that are strikingly similar both qualitatively and quantitatively, suggesting that our main findings of false consensus and its persistence are consistent with this re-weighting (see Appendix, Tables A5, A6 and A7).

4. The Economic Cost of False Consensus

If false consensus leads to miscalibrated trust beliefs, as is consistent with our data so far, the question of consequences naturally arises. That is to say, one would like to get some sense of how financially costly miscalibrated trust beliefs might be. Doing so requires examining how trust beliefs translate into trust behavior, as behavior is what determines earnings. Intuitively, one would expect both the extensive margin and the intensive margin of trust---how much to trust, conditional on whether one trusts at all---to play an important role in this calculation: the gains from exhibiting a small amount of trust, relative to no trust at all, should initially be large as this opens up a wide array of potentially beneficial exchanges (cf., Arrow, 1972); beyond some point, however, the additional economic benefit from a marginal increase in trust probably diminishes, and might even turn negative, as individuals become overly exposed to the risk of being cheated. Putting these insights together, one might expect a concave or even hump-shaped relationship between trust behavior and earnings (cf., Butler, Giuliano and Guiso, 2012a). If trust behavior, in

turn, varies positively with trust beliefs and individuals have self-interested money-maximizing preferences and, moreover, are interacting with a fixed population, one would expect earnings to be maximized when beliefs are correct: miscalibrated trust beliefs should only reduce economic performance.

Unfortunately, the canonical trust game is an environment hostile to the study of the relationship between trust beliefs and the intensive trust margin: linear trust production functions too often imply corner solutions and, consequently, trust behavior that is constant over wide ranges of trust beliefs. As an illustration, consider a risk neutral money-maximizing sender in the canonical trust game with trust production function: $f(s) = 3s$. For now, abstract from expected reciprocity so that the sender expects return proportions to be constant across send amounts. For all trust beliefs, $(\frac{E(r)}{3s})$, less than 1/3, the sender optimally sends zero: every dollar sent entails an expected loss. For any trust belief greater than 1/3, the sender optimally always sends his or her entire endowment. Only for a trust belief of exactly 1/3 is an interior send amount optimal, but then the sender is indifferent between all possible send amounts. Consequently, smooth variation in trust behavior in this game is unlikely to be due to variation in trust beliefs alone; rather, variation in other factors such as distributional social preferences, risk attitudes or expected reciprocity should be the primary drivers.[26]

In light of this limitation of the canonical trust game implemented in Experiment 1, we conduct a second experiment to study the relationship between trust beliefs and earnings, and

[26] If we could be certain that all senders were sufficiently risk averse, we could expect trust behavior to vary smoothly with trust beliefs. Over the small stakes implemented in the lab, however, an influential argument due to Rabin (2000) implies that our participants should be risk neutral. Indeed, one way to justify implementing the concave trust production function described below is that we are inducing risk averse preferences (Smith, 1976).

hence the economic costs of false consensus. In this second experiment we slightly modify the canonical trust game in order to provide an environment more favorable to the study of both the intensive and the extensive margins of trust. In particular, we follow much of the finance literature and assume that returns to trust/investment are concave in the level of trust/investment. Specifically, we implement a concave trust production function calibrated to provide internal optimal send amounts over a wide range of plausible trust beliefs and, more importantly, to generate trust behavior that varies smoothly with trust beliefs.

4.1. Experiment 2: Design and Procedures

Participants for Experiment 2 were recruited from the same pre-existing list of potential experimental student participants at LUISS in Rome, Italy. All sessions were conducted on-line. This experiment was conducted on four separate days, each day constituting a session. In total, 122 students participated in the on-line experiment. We excluded from the list of invitees anybody who had taken part in the laboratory experiment, so that no individual took part in both the in-lab and the on-line experiment.

This on-line experiment implemented one round of the trust game in the same manner as above with four exceptions. The first exception is that the function used to transform money sent into money received was no longer linear but, rather, quadratic which will facilitate our investigation of the intensive margin of trust by providing both an internal optimal send amount and smooth variation in optimal trust behavior over a wide array of trust beliefs and preferences.[27] This function was presented to participants in table format (Table 5).[28] Secondly, a

[27] For example, assume a risk neutral sender facing a trust production function given by f(s) who believes the receiver will return, in expectation, a proportion of the amount he or she receives which is constant over s. Denote this fixed expected return proportion by $\gamma = \frac{R(s)}{f(s)}$ where R(s) is the expected return amount conditional on investing

full strategy method was used: participants submitted their decisions in both possible roles before learning which role they would be assigned. Thirdly, participants did not know their beliefs would be elicited until after they submitted their decisions. This weakens concerns that belief elicitation itself could affect decisions. Finally, only one stated belief was randomly chosen to count towards each participant's earnings---a feature meant to allay concerns about hedging across belief estimates that would be possible if, as in Experiment 1, all reported beliefs were remunerated with certainty.

[Table 5]

To elicit beliefs in Experiment 2 we use a randomized quadratic scoring rule (rQSR) rather than the deterministic quadratic scoring rule (dQSR) used in Experiment 1. The former has been proven to be incentive compatible assuming only expected utility, while the latter requires both expected utility and risk neutrality to provide proper incentives (Schlag and van der Weele, 2012).[29] Given the payment structure of Experiment 2 in which one decision or reported belief

---

s. The sender's expected earnings from sending an amount $s > 0$ are $y = 10 - s + \gamma f(s)$. The first order conditions for an interior optimal amount to invest satisfy $f'(s) = (1/\gamma)$. We implement $f(s) \approx 8s^{1/2}$, which implies $s^* \approx 16\gamma^2$, whereas the canonical trust production function $f(s) = 3s$ never yields unique internal optima under these assumptions.

[28] The table represents the function $f(s) = 8.05s^{1/2}$ with some minor modifications for appearance.

[29] Recall that in the dQSR in Experiment 1, for each amount a receiver could receive, $m \in \{3,\ldots,30\}$, if the true average amount receivers returned was r_m, and a participant's stated belief about this value was, $\widehat{r_m}$, the participant earned an amount of money given by the formula: $Earnings = 1 - \left(\frac{\widehat{r_m} - r_m}{m}\right)^2$. Maximizing the expected value of this function is equivalent to minimizing a mean squared error, $\left(\frac{\widehat{r_m} - r_m}{m}\right)^2$, which is accomplished by reporting the mean of one's subjective belief distribution. Assuming risk neutrality, the dQSR is therefore incentive compatible and elicits the mean of the participant's beliefs distribution.

can determine a substantial proportion of a participant's entire earnings, we were more concerned about the plausibility of the risk neutrality assumption here than in our previous experiment and, so, chose a beliefs elicitation mechanism which does not rely upon it.[30]

At the end of the experiment, after all decisions and beliefs were collected, 10 percent of participant pairs were randomly chosen to be paid according to their decisions and estimates. Since the on-line experiment required much less of participants' time, this kept hourly earnings comparable to earnings in the laboratory experiment.[31]

If the participant seeks to maximize something other than the expected value of this function because, e.g., s/he is risk averse or risk seeking, the dQSR rule may not be incentive compatible. The rQSR (Schlag and van der Weele, 2012) remedies this problem by use of a clever trick. Start with the same function as in the dQSR above: $y = 1 - \left(\frac{\widehat{r_m} - r_m}{m}\right)^2$. Notice that *y* is a number between 0 and 1 that is increasing in the accuracy of the reported belief, $\widehat{r_m}$. Instead of paying *y* euros directly, however, the rQSR uses *y* to determine the probability with which a participant earns a fixed prize (in our experiment, 5 euros) in the following way: the experimenter draws a number, z, at random from the set [0,1]; if $z \leq y$, the participant earns 5 euros; otherwise the participant earns nothing. Any expected utility maximizer will optimally seek to maximize the expected value of *y*, as this maximizes the probability of winning 5 euros. As we have already seen, maximizing the expected value of y is accomplished by reporting the mean of one's subjective probability distribution. Consequently, the rQSR is theoretically incentive compatible even if a participant is not risk neutral, assuming expected utility. Moreover, the assumption of expected utility is sufficient, but may not be necessary.

[30] If, as is commonly argued, the small stakes of both experiments make risk neutrality a valid assumption, then nothing is lost by using rQSR instead of dQSR---they are both incentive compatible; if, on the other hand, risk neutrality is not a valid assumption, then using the rQSR in Experiment 2 may serve as a check on how robust the patterns we find in Experiment 1 are to the incentive compatibility of the beliefs elicitation mechanism.

[31] For evidence that paying only a randomly selected subset of participants in the trust game does not substantially affect behavior or beliefs relative to a case where all participants are paid see, e.g., Butler, Giuliano and Guiso (2012b).

4.2. Measuring Trust Belief Errors

From the data we construct two measures of the errors in participants' trust beliefs, or, more succinctly, their trust belief errors. We first construct an individual-specific measure of trust belief errors for each send amount, separately, as follows. For each participant, i, and for each s $\in \{1, \ldots, 10\}$, we compute the average return amount of all participants in a session except participant *i*. This gives us, for each participant, *i*, a 10-element session-specific vector of the average return amounts of all participants besides *i*.[32] To put this in percentage terms, we divide the s^{th} element of this vector by $f(s)$ where $f(\cdot)$ is the trust production function and s $\in \{1, \ldots, 10\}$. Call this vector an average others' trustworthiness vector. We compute a vector of trust belief errors for each participant, i, and each send amount, *s*, as the element-by-element difference between *i*'s vector of stated trust beliefs and *i*'s average others' trustworthiness vector. Since both trust beliefs and others' trustworthiness are computed in terms of the percentage of the amount received, each element of a participant's trust belief error vector can take values from -1 to 1.

The second measure of trust belief errors we construct is a summary, unidimensional, measure obtained by computing the average over all ten elements of each participant's vector of trust belief errors. As each element of a participant's vector of trust belief errors can take values from -1 to 1, so can this summary measure. Negative values indicate that a participant's trust beliefs are generally too pessimistic, while positive values indicate unwarranted optimism about others' trustworthiness.

[32] It is important to exclude participant i him/herself from this calculation to minimize any actual or expected mechanical correlation between own trustworthiness and others' trustworthiness. This is a concern here, but not in Experiment 1, because we use the full strategy method. Participants are instructed that we will remunerate the accuracy of their beliefs with respect to this measure of others' average return amounts.

4.3. Measuring Economic Performance

As our primary focus will be on the economic consequences of trust belief errors, we also compute for each participant a measure of economic performance. One way of doing this would be to simply use the earnings for each participant that we calculated in conducting the experiment. While this has the advantage of implementing exactly the situation described to participants, it has the drawback of throwing away a substantial amount of information--only half of the participants would be assigned the role of sender, and hence we could only examine the earnings/trust belief errors relationship for half of our participants. To balance the concern with data loss against a desire to implement a situation as close as possible to the situation described to participants, we choose a different route: in essence, we compute the earnings each participant would have earned if they had been assigned the role of sender. This preserves the integrity of the experiment---each participant was asked to decide as if they had been assigned the sender role---while minimizing data loss. Specifically, for each participant i, we construct a measure of performance by randomly choosing one other participant, $j \neq i$, from the same experimental session and computing *i*'s earnings using *i*'s sender strategy and *j*'s receiver strategy.[33] [34]

4.4. Results

[33] Earnings for each participant i are given by: $Y_i = 10.5 - s_i + r_j\big(f(s_i)\big) - 0.5I(s_i)$, s_i where denotes how much *i* sends to *j*, $r_j\big(f(s_i)\big)$ denotes how much *j* returns to *i* conditional on receiving $f(s_i)$ and $I(s_i)$ is an indicator function equal to 1 if *i* sends a strictly positive amount.

[34] One may wonder whether restricting matches by session is appropriate here, in this on-line environment. One reason we feel that it is appropriate is because participants signed up for specific sessions and therefore may have expected that matchings would be restricted by sessions. Moreover, restricting matching to be within-session maintains as close a parallel as possible to the most straightforward option of using participants' actual experimental earnings as our measure of performance.

Before analyzing earnings, we document two features of the data necessary to finding a substantial cost of false consensus: i) a relationship between behavior and beliefs consistent with a substantial false consensus effect; ii) substantial variation in the sign and magnitude of trust belief errors. We begin by providing evidence on the relationship between trust beliefs and own trustworthiness in Experiment 2 (Table 6). We first use our summary measures, regressing trust beliefs on own trustworthiness (Panel A). We find a strong, positive and significant relationship between our unidimensional measures of beliefs and behavior. Considering each send amount separately (Panel B), we document that own trustworthiness is a highly significant predictor of trust beliefs irrespective of which send amount we consider.[35] Next, we regress trust belief errors on own trustworthiness for each send amount separately (Panel C) and find, as expected, a strong positive relationship between trustworthiness and trust belief errors. The sign and magnitude of the constant terms indicate that very untrustworthy participants tend to underestimate population trustworthiness (negative trust belief errors) but that underestimation may fade and switch to overestimation (positive trust belief errors) for highly trustworthy participants.

[Table 6]

[35] One potential concern common to most experimental research relates to stake size. It could be that participants rely on heuristics such as extrapolating from their own types only when stake sizes are small. Although we cannot directly address that concern here since we did not vary the payoffs for correct beliefs in this experiment, we have a related paper which uses the same "quadratic trust game" in which we vary payoffs for correct beliefs across sessions (Butler, Giuliano and Guiso, 2012b). There, in some treatments exactly correct beliefs pay 5 euros---as they do here---while, in other treatments, exactly correct beliefs earn the participant four times as much---20 euros. We find that the correlation between own trustworthiness and trust beliefs increases when the payment for correct beliefs increases.

The variation in terms of sign and magnitude of trust belief errors is evident in Figure 2, which presents a scatter plot of trust belief errors by own trustworthiness using our summary measures of each. Each dot corresponds to one participant. We overlay the scatter plot with the best linear fit of the data. There are two main points worth noting. First of all, there is a substantial proportion of both overly pessimistic and overly optimistic participants in our data---i.e., those whose trust belief errors are negative and positive, respectively. The second point to notice is that trust belief errors go from being mostly negative to being mostly positive over the range of observed trustworthiness levels. These two features together provide the necessary conditions for false consensus to have an impact on senders' economic performance.

[Figure 2]

Having established the preconditions for false consensus to matter for earnings, we now turn to our main result from Experiment 2: miscalibrated trust beliefs are associated with large losses in earnings. As a first pass, in Figure 3 we plot the relationship between senders' earnings and our summary measure of trust belief errors and overlay the plot with the best quadratic fit of the data. Each point in the plot represents one participant. As we hypothesized above, we find a hump-shaped relationship between trust belief errors and senders' earnings: those who hold overly pessimistic trust beliefs, as well as those who hold overly optimistic trust beliefs, earn less than senders whose beliefs are approximately correct---i.e., those whose trust belief errors are approximately zero.

[Figure 3]

The statistical significance of the humped shape is confirmed by the regressions presented in Table 7, where we estimate senders' earnings as a quadratic function of trust belief errors. In our most basic specification (column 1), the coefficient on trust belief errors squared is

both negative and significant. The coefficient estimates imply earnings achieve their maximum when trust beliefs errors are close to zero. Both of these features are robust to controlling for session fixed effects (column 2) and to clustering standard errors by session (column 3) to allow for arbitrary within-session correlation of behavior.[36] The estimates suggest that senders earn from about 11 euros (column 3) to about 11.47 euros (column 1) on average when belief errors are zero, which constitutes a 5 to 9 percent increase over the safe return (10.50 euros) from sending nothing. All three estimates imply earnings of around 7 euros for the most pessimistic observed trust belief error (-0.36) and approximately 9 euros for the most optimistic trust belief error in the data (0.35), implying an earnings shortfall of 19% to 34% compared to exactly correct trust beliefs.

[Table 7]

To get another measure of the magnitude of the earnings consequences of false consensus that is less dependent on functional form assumptions, we next divide participants into three categories---"under-estimators," "over-estimators," and "accurate-estimators"---according to whether their trust belief errors are positive, negative or approximately zero. Specifically, accurate-estimators are those participants whose summary trust belief errors measure falls within a small interval around zero [-0.1,0.1], while under-estimators (over-estimators) have trust belief errors that are negative (positive) and fall outside of this interval. We regress senders' earnings on dummies for these three belief errors categories, under-estimators being the excluded category, and report the results in Table 8. The estimates show that accurate-estimators earned about 18 percent more on average than under-estimators, who, in turn, earned about the same as

[36] Each separate day on which the experiment was conducted constitutes a session.

over-estimators. The estimates are robust to controlling for session fixed effects (column 2) and clustering standard errors by session (column 3).[37]

[Table 8]

In summary, Experiment 2 allows us to investigate the economic consequences of miscalibrated trust beliefs. Consistent with a story where miscalibration is due to false consensus, we find again that trust beliefs vary significantly with own trustworthiness and that, moreover, the resulting variation in trust beliefs can have a substantial impact on our participants' earnings. Miscalibrated trust beliefs reduce earnings in our experiment by roughly 20 percent, on average.

5. Conclusions

Large-scale survey evidence suggests that trust beliefs are both extremely heterogeneous across individuals and persistent over age and across generations. In this paper we present the results of two experiments aimed at investigating one prevalent phenomenon that can explain both of these patterns: false consensus. In the first experiment, we find that the relationship between behavior and beliefs is consistent with individuals extrapolating from their own types when forming their trust beliefs about a novel population (false consensus) and that one's own type continues to be a significant predictor of trust beliefs even after considerable opportunities for learning about the population. In our second experiment we use a trust game slightly modified to be more amenable to studying the potential earnings consequences of false consensus than the canonical game by facilitating smooth variation in trust behavior with trust

[37] Results are also robust to using a wider interval---[-0.15,0.15]---or a narrower interval--[-0.05,0.05]--to define accurate-estimators, as well as using a definition of over- and under-estimators defined by the 33rd and 66th percentiles of the observed belief errors.

beliefs. In this one-shot setting, we again find evidence for false consensus---own trustworthiness is highly predictive of trust beliefs. Moreover, we estimate the potential impact of false consensus on earnings to be substantial: miscalibrated trust beliefs lower participants' earnings by about 20 percent.

Einaudi Institute for Economics and Finance

UCLA Anderson School of Management

Einaudi Institute for Economics and Finance

References

Algan, Y. and P. Cahuc, "Inherited Trust and Growth," American Economic Review, 100(5) (2010), 2060-92.

Arrow, K., "Gifts and Exchanges", Philosopy & Public Affairs, 1(4) (1972), 343-362

Berg, J., Dickhaut, J. and K. McCabe, "Trust, Reciprocity and Social History," Games and Economic Behavior, 10 (1995), 122-142.

Bellemare, C. and S. Kröger, "On representative social capital," European Economic Review, 51 (2008), 183-202.

Bisin, A., G. Topa, and T. Verdier, "Cooperation as a Transmitted Cultural Trait," Rationality and Society, 16(4) (2008), 477-507.

Bisin, A. and T. Verdier, "Beyond the Melting Pot: Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits," Quarterly Journal of Economics, 115-3 (2008), 955-988.

Bohnet, I., B. S. Frey and S. Huck, "More Order with Less Law: On Contract Enforcement, Trust, and Crowding," The American Political Science Review, 95 (2001), 131-144.

Bohnet, I. and S. Huck, "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change," American Economic Review, 94 (2004), 362-366.

Buchan, N.R., R.T.A. Croson and S. Solnick, "Trust and gender: An examination of behavior and beliefs in the Investment Game," Journal of Economic Behavior and Organization, 68 (2008), 466-476.

Butler, J.V., P. Giuliano and L. Guiso, "The Right Amount of Trust," EIEF Working Paper, 2012a.

Butler, J. V., P. Giuliano and L. Guiso, "Cheating in the Trust Game," EIEF Working Paper, 2012b.

Camerer, C., Behavioral Game Theory: Experiments in Strategic Interaction (Princeton, NJ: Princeton University Press, 2003).

Camerer, C. and and K. Weigelt, "Experimental tests of a sequential equilibrium reputation model," Econometrica, 56 (1988), 1-36.

Costa-Gomes, M. A., S. Huck and G. Weizsäcker, "Beliefs and Actions in the Trust Game: Creating Instrumental Variables to Estimate the Causal Effect," IZA Discussion Paper No. 4709, 2010

Cox, J. C., "How to Identify Trust and Reciprocity," Games and Economic Behavior, 46 (2004), 260-281

Dohmen, T., A. Falk, D. Huffman, and U. Sunde, "The Intergenerational Transmission of Risk and Trust Attitudes," The Review of Economic Studies, 79(2) (2012), 645-677.

Fehr, E., U. Fischbacher, B. Von Rosenbladt, J. Schupp and G. G. Wagner, "A Nation-Wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Survey," IZA Discussion Paper No. 715, 2003.

Fischbacher U.: "Z-Tree: Zurich Toolbox for Ready-made Economic Experiments", Experimental Economics, 10(2) (2007), 171-178.

Glaeser, E., D. Laibson, J. A. Scheinkman and C. L. Soutter, "Measuring Trust," Quarterly Journal of Economics 115(3) (2000), 811-846.

Guiso, L., P. Sapienza and L. Zingales, "Long Term Persistence," NBER WP 14278, 2008a.

Guiso, L., P. Sapienza and L. Zingales, "Social Capital as Good Culture," Journal of the European Economic Association, 6(2-3) (2008b), 295-320.

Güth, W., H. Kliemt and B. Peleg, "Co-evolution of preferences and information in simple games of trust," Discussion Papers, Interdisciplinary Research Project 373, Quantification and Simulation of Economic Processes, No. 1998,72, (1998)

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

Johnson, N. D. and A. A. Mislin, "Trust Games: A meta-analysis," Journal of Economic Psychology, 32 (2011), 865-889.

Krueger, J. and R. W. Clement, "The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception", Journal of Personality and Social Psychology of Addictive Behaviors, 67 (4) (1994), 596-610.

Massey, C. and R. H. Thaler, "The Loser's Curse: Overconfidence vs. Market Efficiency in the National Football League Draft," University of Chicago, mimeo, 2006.

Rabin, M., "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," Econometrica, 68 (2000), 1281-1292.

Ross, L., Greene, D., and P. House, "The False Consensus Phenomenon: An Attributional Bias in Self-Perception and Social Perception Processes," Journal of Experimental Social Psychology, 13(3) (1977), 279-301.

Sapienza, P., Toldra-Simats, A. and L. Zingales, "Understanding Trust," The Economic Journal, (forthcoming).

Schlag, K. and J. van der Weele, "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality," Theoretical Economics Letters, 3 (2013), 38-42.

Schelling, T., "Uncertainty, Brinkmanship and the Game of `Chicken'," in Strategic Interaction and Conflict, K. Archibald (ed.) (Berkeley, California: Berkeley Institute for International Studies, University of California, Berkeley, 1966)

Smith, V., "Experimental Economics: Induced Value Theory," The American Economic Review: Papers and Proceedings, 66 (1976), 274-279.

Tabellini, G., "The Scope of Cooperation: Values and Incentives," Quarterly Journal of Economics, 123 (3) (2008), 905-950.

Table 1
Descriptive statistics

A. Experiment 1

Variable	mean	st dev
Good Values	0.637	0.199
Initial own trustworthiness (summary measure)	0.320	0.162
Trust beliefs (summary measure)	0.265	0.158
Return proportion	0.211	0.18
Send amount	5.258	3.107
Send amount > 0 (dummy)	0.676	0.469

B. Experiment 2

Variable	mean	st dev
Send amount > 0 (dummy)	0.730	0.446
Send amount	3.934	3.315
Trust beliefs (summary measure)	0.332	0.142
Trustworthiness (summary measure)	0.339	0.160
Trust belief error (summary measure)	-0.007	0.145
Sender earnings	10.950	3.077

Table 2
The effect of own trustworthiness on trust beliefs

A. OLS estimates of trust beliefs on own initial trustworthiness, summary measures

	Rounds 1-3 Trust beliefs	Rounds 4-6 Trust beliefs	Rounds 7-9 Trust beliefs	Rounds 10-12 Trust beliefs
Own initial trustworthiness	0.74***	0.54***	0.47***	0.45***
	(0.04)	(0.07)	(0.07)	(0.08)
Constant	0.08***	0.11***	0.08***	0.07**
	(0.02)	(0.02)	(0.03)	(0.02)
Observations	276	208	171	171
R-squared	0.59	0.31	0.26	0.25

B. OLS estimates of trust beliefs on own initial trustworthiness, by send amount

	Column Heading = Send Amount									
	1	2	3	4	5	6	7	8	9	10
					Rounds 1-3					
Own initial trustworthiness	0.56***	0.64***	0.69***	0.70***	0.83***	0.69***	0.75***	0.77***	0.78***	0.78***
	(0.09)	(0.10)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)	(0.08)	(0.08)	(0.07)
Constant	0.10***	0.09***	0.08***	0.10***	0.07**	0.11***	0.09***	0.08***	0.08***	0.08***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
Observations	276	276	276	276	276	276	276	276	276	276
R-squared	0.36	0.42	0.44	0.49	0.53	0.49	0.49	0.60	0.55	0.55
					Rounds 4-6					
Own initial trustworthiness	0.37***	0.40***	0.44***	0.51***	0.43***	0.52***	0.52***	0.64***	0.58***	0.61***
	(0.10)	(0.14)	(0.13)	(0.12)	(0.12)	(0.12)	(0.13)	(0.12)	(0.13)	(0.12)
Constant	0.15***	0.17***	0.18***	0.18***	0.22***	0.18***	0.18***	0.14***	0.16***	0.15***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)
Observations	208	208	208	208	208	208	208	208	208	208
R-squared	0.19	0.20	0.21	0.25	0.17	0.28	0.23	0.38	0.30	0.33
					Rounds 7-9					
Own initial trustworthiness	0.38***	0.47***	0.52***	0.47***	0.46***	0.45***	0.49***	0.62***	0.54***	0.63***
	(0.14)	(0.16)	(0.15)	(0.16)	(0.15)	(0.16)	(0.17)	(0.15)	(0.16)	(0.14)
Constant	0.16***	0.17***	0.18***	0.21***	0.24***	0.22***	0.21***	0.18***	0.20***	0.18***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
Observations	171	171	171	171	171	171	171	171	171	171
R-squared	0.16	0.21	0.22	0.22	0.19	0.22	0.19	0.34	0.22	0.30
					Rounds 10-12					
Own initial trustworthiness	0.53***	0.60***	0.50***	0.46***	0.48***	0.43***	0.48***	0.53***	0.50***	0.55***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.14)	(0.14)	(0.15)	(0.12)	(0.16)	(0.14)
Constant	0.16***	0.16***	0.19***	0.23***	0.24***	0.24***	0.22***	0.21***	0.21***	0.20***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Observations	171	171	171	171	171	171	171	171	171	171
R-squared	0.27	0.32	0.20	0.20	0.20	0.19	0.19	0.24	0.18	0.22

C. OLS estimates of expected reciprocity on own initial reciprocity, and of expected baseline trustworthiness on initial baseline trustworthiness

Expected Reciprocity as a Function of Own Initial Reciprocity (OLS)				
	Rounds 1-3	Rounds 4-6	Rounds 7-9	Rounds 9-12
Own initial baseline reciprocity	0.74***	0.58***	0.64***	0.66***
	(0.06)	(0.09)	(0.12)	(0.11)
Constant	-0.00	0.00	0.00	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	276	208	171	171
R-squared	0.44	0.18	0.28	0.32

Expected Baseline Trustworthiness as a Function of Own Initial Baseline Trustworthiness (OLS)				
	Rounds 1-3	Rounds 4-6	Rounds 7-9	Rounds 9-12
Own initial baseline trustworthiness	0.71***	0.54***	0.51***	0.57***
	(0.04)	(0.08)	(0.06)	(0.07)
Constant	0.10***	0.09***	0.05**	-0.00
	(0.02)	(0.03)	(0.02)	(0.02)
Observations	276	208	171	171
R-squared	0.46	0.21	0.23	0.31

Notes: [1] For all panels (A, B and C), robust standard errors clustered by participant appear in parentheses, *** significant at 1%, ** significant at 5%, * significant at 10%. [2] Panel A presents OLS regressions of our summary, unidimensional, measure of trust beliefs on our summary measure initial trustworthiness; Panel B presents a similar regression but using trust beliefs and initial trustworthiness measures disaggregated by send amount; Panel C uses the linearized ($r = ms + b$) form for trust beliefs and own initial trustworthiness and presents regressions of beliefs about others' slope (m) and intercept (b) terms on own initial slope and intercept terms, separately.

Table 3
The relationship between received cultural values and own initial trustworthiness
A. OLS estimate of our summary measure of initial trustworthiness on "good values"

	Initial trustworthiness
Good values	0.17*
	(0.09)
Constant	0.21***
	(0.06)
Observations	83
R-squared	0.04

B. OLS estimate of initial trustworthiness on "good values," by send amount

Dependent Variable = Own Initial Trustworthiness										
Column Heading = Send Amount										
	1	2	3	4	5	6	7	8	9	10
Good values	0.22*	0.11	0.18*	0.16	0.19*	0.15	0.17*	0.19**	0.17*	0.15
	(0.12)	(0.11)	(0.09)	(0.09)	(0.10)	(0.09)	(0.10)	(0.09)	(0.10)	(0.10)
Constant	0.11	0.21***	0.19***	0.22***	0.22***	0.25***	0.23***	0.21***	0.23***	0.24***
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Observations	83	83	83	83	83	83	83	83	83	83
R-squared	0.04	0.01	0.04	0.03	0.04	0.03	0.03	0.05	0.03	0.02

C. OLS estimate of initial baseline trustworthiness and reciprocity on "good values"

	Initial Reciprocity	Initial Baseline Trustworthiness
Good values	-0.00	0.17
	(0.01)	(0.10)
Constant	0.01	0.16**
	(0.01)	(0.06)
Observations	83	83
R-squared	0.00	0.03

Notes: [1] Robust standard errors are reported in parentheses, ** significant at 5%, * significant at 10%. [2] To account for possibly multiple observations by participant due to pooling over blocks of rounds, Panels B features standard errors clustered by individual. [3] The numbers of observations falls in later rounds because some sessions, due to time constraints, contained fewer than 12 rounds. [4] The number of observations falls when including our "good values" measure, because some participants did not complete the survey.

Table 4
The relationship between received cultural values and trust beliefs

A. OLS estimates of our summary measure of trust beliefs on good values

	Rounds 1-3 Trust beliefs	Rounds 4-6 Trust beliefs	Rounds 7-9 Trust beliefs	Rounds 10-12 Trust beliefs
Good Values	0.12**	0.13*	0.12*	0.05
	(0.06)	(0.07)	(0.07)	(0.08)
Constant	0.25***	0.20***	0.14***	0.17***
	(0.04)	(0.04)	(0.04)	(0.05)
Observations	339	262	216	216
R-squared	0.03	0.03	0.03	0.00

B. OLS estimates of trust beliefs on good values, by send amount

	Column Heading = Send Amount									
	1	2	3	4	5	6	7	8	9	10
					Rounds 1-3					
Good values	0.14	0.12	0.10	0.11*	0.13**	0.12**	0.13**	0.13**	0.12*	0.13**
	(0.10)	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Constant	0.18***	0.22***	0.25***	0.26***	0.26***	0.26***	0.25***	0.26***	0.27***	0.26***
	(0.06)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Observations	339	339	339	339	339	339	339	339	339	339
R-squared	0.01	0.02	0.01	0.02	0.03	0.03	0.02	0.02	0.02	0.02
					Rounds 4-6					
Good values	0.07	0.14*	0.14	0.14**	0.12*	0.10	0.14**	0.13**	0.12*	0.14**
	(0.12)	(0.09)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)
Constant	0.16**	0.16***	0.17***	0.18***	0.22***	0.23***	0.21***	0.21***	0.22***	0.21***
	(0.08)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
Observations	262	262	262	262	262	262	262	262	262	262
R-squared	0.00	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.02	0.03
					Rounds 7-9					
Good values	0.08	0.13	0.14*	0.17**	0.14*	0.14*	0.15*	0.12	0.10	0.08
	(0.09)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Constant	0.12**	0.10**	0.12**	0.11**	0.15***	0.15***	0.15***	0.17***	0.18***	0.19***
	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Observations	216	216	216	216	216	216	216	216	216	216
R-squared	0.01	0.02	0.03	0.04	0.03	0.03	0.03	0.02	0.01	0.01
					Rounds 10-12					
Good values	-0.01	0.06	0.06	0.06	0.05	0.06	0.07	0.05	0.06	0.05
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Constant	0.13**	0.12**	0.15**	0.16***	0.18***	0.18***	0.18***	0.20***	0.20***	0.21***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)
Observations	216	216	216	216	216	216	216	216	216	216
R-squared	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00

C. OLS estimates of beliefs about others' reciprocity and baseline trustworthiness on good values

Expected Reciprocity as a Function of Good Values

	Rounds 1-3	Rounds 4-6	Rounds 7-9	Rounds 9-12
Good values	0.00	0.00	-0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.01	0.01	0.01	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	339	262	216	216
R-squared	0.00	0.00	0.00	0.00

Expected Baseline Trustworthiness as a Function of Good Values (OLS)

	Rounds 1-3	Rounds 4-6	Rounds 7-9	Rounds 9-12
Good values	0.12	0.11	0.13	0.04
	(0.08)	(0.11)	(0.09)	(0.09)
Constant	0.21***	0.16**	0.09	0.12*
	(0.05)	(0.07)	(0.06)	(0.06)
Observations	339	262	216	216
R-squared	0.01	0.01	0.02	0.00

Notes: [1] Robust standard errors, clustered by participant, are reported in parentheses, ** significant at 5%, * significant at 10%. [2] Clustering by participant is appropriate because there multiple observations for each participant within each three-round block. [3] The numbers of observations falls in later rounds because some sessions, due to time constraints, contained fewer than 12 rounds. [4] The number of observations falls when including our "good values" measure, because some participants did not complete the survey.

Table 5
Trust production function used in experiment 2

Sender sends € s:	1	2	3	4	5	6	7	8	9	10
Receiver receives € f(s):	8.05	11.30	13.85	16.05	17.90	19.60	21.20	22.65	24.05	25.30

Table 6
Evidence for false consensus in experiment 2

A. OLS estimates of trust beliefs on own trustworthiness, summary measures

	Trust beliefs
Own trustworthiness	0.39***
	(0.08)
Constant	0.20***
	(0.03)
Observations	122
R-squared	0.19

B. OLS estimates of trust beliefs on own trustworthiness

	Column Heading = Send Amount									
	1	2	3	4	5	6	7	8	9	10
Own trustworthiness	0.35***	0.21**	0.28***	0.34***	0.21**	0.31***	0.33***	0.32***	0.32***	0.32***
	(0.09)	(0.09)	(0.08)	(0.07)	(0.10)	(0.09)	(0.08)	(0.07)	(0.08)	(0.09)
Constant	0.18***	0.22***	0.21***	0.21***	0.26***	0.25***	0.23***	0.24***	0.25***	0.26***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Observations	122	122	122	122	122	122	122	122	122	122
R-squared	0.21	0.07	0.12	0.19	0.06	0.11	0.16	0.14	0.15	0.12

C. OLS estimates of trust belief errors on own trustworthiness

	Column Heading = Send Amount									
	1	2	3	4	5	6	7	8	9	10
Own trustworthiness	0.35***	0.22**	0.28***	0.31***	0.20*	0.30***	0.32***	0.28***	0.31***	0.32***
	(0.09)	(0.09)	(0.07)	(0.07)	(0.10)	(0.09)	(0.08)	(0.08)	(0.08)	(0.09)
Constant	-0.10***	-0.06**	-0.11***	-0.12***	-0.07*	-0.08**	-0.15***	-0.12***	-0.13***	-0.11***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Observations	122	122	122	122	122	122	122	122	122	122
R-squared	0.21	0.08	0.13	0.15	0.05	0.10	0.14	0.10	0.13	0.12

Notes: [1] Robust standard errors appear in parentheses, *** significant at 1%, ** significant at 5%, * significant at 10%. [2] "Own trustworthiness" is the proportion of the amount received the participant would return if he/she were to be sent s euros and therefore receive f(s) euros (i.e., [return amount conditional on receiving f(s)]/[f(s]); "Trust beliefs" is a participant's beliefs about others' trustworthiness (i.e. beliefs about [others' avg. return amount conditional on receiving f(s)]/[f(s]); "Trust belief errors" is the difference between trust beliefs and others' actual trustworthiness.

Table 7
Senders' earnings and trust belief errors in experiment 2, quadratic specification
OLS estimates of sender's earnings on errors in trust beliefs

	(1)	(2)	(3)
Belief Errors	1.898	2.196	2.196**
	(1.595)	(1.577)	(0.742)
Belief Errors Squared	-24.061***	-23.360***	-23.360**
	(7.353)	(7.945)	(4.798)
Constant	11.465***	10.995***	10.995***
	(0.356)	(0.639)	(0.118)
Session Fixed Effects?	No	Yes	Yes
Session-Clustered Std Errors?	No	No	Yes
Observations	122	122	122
R-squared	0.05	0.07	0.07

Notes: [1] Robust standard errors are in parentheses, *** significant at 1%, ** significant at 5%, * significant at 10%. [2] Belief errors are defined by the difference between a participant's estimate of the proportion of money received that a receiver will return and the actual average return proportion within each session, averaged over each possible amount a receiver could receive. This value excludes the participant's own action in the role of receiver. This yields a number that ranges from -1 to 1 for each participant.

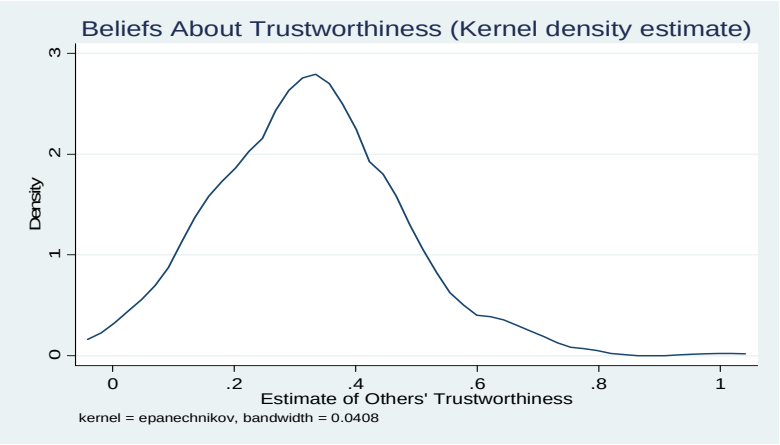
Table 8
Senders' earnings and trust belief errors in experiment 2, 3-category specification
OLS estimates of sender's earnings on dummies for trust beliefs categories

	(1)	(2)	(3)
Accurate Estimators	1.860***	1.773***	1.773**
	(0.663)	(0.657)	(0.500)
Over-estimator	0.311	0.324	0.324
	(0.706)	(0.681)	(0.352)
Constant	9.930***	9.554***	9.554***
	(0.525)	(0.603)	(0.135)
Session Fixed Effects?	No	Yes	Yes
Session-Clustered Std Errors?	No	No	Yes
Observations	122	122	122
R-squared	0.07	0.09	0.09

Notes: [1] Robust standard errors are in parentheses, *** significant at 1%, ** significant at 5%, * significant at 10%. [2] Dependent variable is sender's earnings in euros. [3] The excluded category is "under-estimators." [4] Belief error categories are defined as follows: "Accurate Estimators" had an average belief error within the interval [-0.1, 0.1]; "Over-estimators" had an average belief error in the interval (0.1,1]; "Under-estimators" had an average belief error in the interval [-1,-0.1). [5] We also considered wider and narrower intervals separating the three categories, using [-0.15, 0.15] and [-.05, 0.05] to define accurate estimators. This did not change anything qualitatively; [6] Another specification used the 33rd and 66th percentiles of the error distribution in the data to separate the three categories. This did not change the results.

Figure 1
Heterogeneity in trust priors and own trustworthiness, Experiment 1

A. Trust priors (unidimensional measure)



B. Own initial trustworthiness (unidimensional measure)

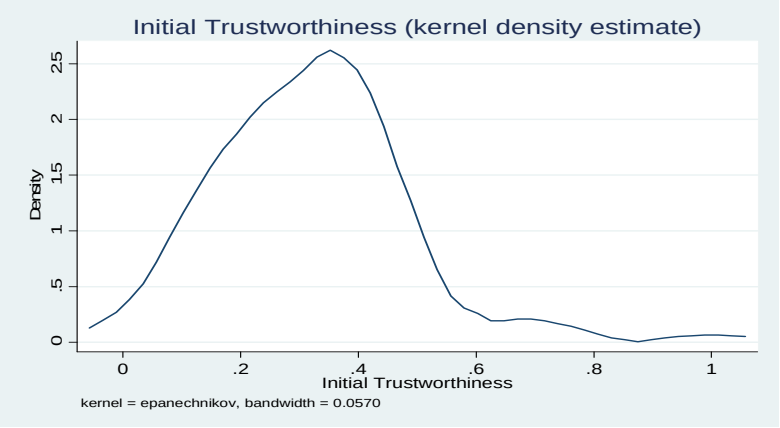


Figure 2
Trust belief errors and own trustworthiness, Experiment 2

Figure 3
Trust belief errors and senders' earnings, Experiment 2