Trust and Cheating*

Jeffrey V. Butler
EIEF

Paola Giuliano
UCLA, NBER, CEPR and IZA

Luigi Guiso
EIEF and CEPR

This version: June 2014

Abstract

When we take a cab we may feel cheated if the driver takes an unnecessarily long route despite the lack of a contract or promise to take the shortest possible path. Is the behavior of the driver affected by beliefs about our cheating notions, and where do his beliefs come from? For that matter, where do our cheating notions come from, and how do they color our own decisions? We address these questions in the context of a trust game by asking participants directly about their personal notions of cheating. We find that both parties to a trust exchange have personal notions of what constitutes cheating; that these notions have a bimodal distribution; and that cheating notions are determined by parentally-transmitted values. We think about cheating notions as moral expectations, which provide a micro-foundation for guilt which extends the scope for empirical applications of guilt aversion theory. We document that cheating notions substantially affect decisions on both sides of the trust exchange.

JEL Classification: A1, A12, D1, O15, Z1
Keywords: Trust, trustworthiness, social norms, culture, cheating

*We thank Roland Bénabou, Gary Charness, Martin Dufwenberg, Andrea Galeotti, Uri Gneezy and three anonymous referees, as well as seminar participants at the EIEF, the Sciences Po/IZA Workshop on Trust, Civic Spirit and Economic Performance, the Florence Workshop on Behavioural and Experimental Economics and the SITE Summer workshop at Stanford University for many helpful comments which have greatly improved the paper.

1 Introduction

When taking a cab we may expect the driver to use a reasonably short route even if neither we nor the driver make explicit mention of it. Despite the lack of explicit promise, we may still feel cheated if the taxi driver takes an unnecessarily long route. Similarly, when we ask for financial advice the advisor does not typically spell out that he will act solely in our best interest, but we may still judge cheating according to this metric. When we book a vacation through a travel agent, search for the best medical insurance at a broker or take our car to a mechanic, we may act on implicit notions of what the behavior of the travel agent, broker or mechanic should be, perhaps feeling cheated or let down when behavior fails to live up to these standards.

Situations like these come up frequently in our daily economic lives: opportunities for mutually beneficial exchanges where complete contracts, agreements or credible communication about what is expected from each side of the exchange are either impossible or infeasible. Considering only our first example above, over $600,000$ taxi rides are taken daily in New York city alone constituting about $1 billion in fares paid per year.[1] And New York is not alone: about one million people use taxis every day in Hong Kong,[2] while a staggering three to four million taxi rides are taken every day in Lima, Peru (Castillo, et al., 2012). Our second example – financial advice from professionals – is also pervasive. According to a broad survey of retail investors in Germany, more than 80 percent of investors consult a financial advisor. Overall, in the UK 91% of intermediary mortgage sales are "with advice" (Chater, Huck and Inderst, 2010). In the US, 73% of all retail investors consult a financial advisor before purchasing shares (Hung, *et al.*, 2008).[3] Given their ubiquity, understanding precisely what drives behavior in such trust-based exchange opportunities is an important undertaking.

In this paper, we focus on one intuitively plausible yet under-explored determinant of behavior on both sides of such exchange opportunities: individuals' personal subjective notions of what constitutes cheating. While individuals may hold widely divergent views on what constitutes cheating and this heterogeneity in cheating notions may, in turn, translate into heterogeneous behavior, economists know virtually nothing about the individual-level

[1] http://en.wikipedia.org/wiki/Transportation_in_New_York_City.
[2] http://www.gov.hk/en/about/abouthk/factsheets/docs/transport.pdf
[3] See also Inderst and Ottaviani (2012) for a general review on financial advice.

relationship between individual cheating notions and behavior in trust-based exchange opportunities. For instance, in the massive body of experimental trust game literature researchers typically *assume* that both involved parties will define cheating according to a single, shared, notion.[4] While this methodology has proven useful for showing that pecuniary concerns alone fail to account for a significant portion of exchange behavior, its ability to provide a detailed understanding of how idiosyncratic cheating notions translate into behavior is obviously limited.

We investigate the role of cheating notions in the context of a trust game (Berg, Dickhaut and McCabe, 1995), a two player sequential moves game of perfect information. In this game, the sender moves first by deciding whether to send some, all or none of a fixed endowment to a co-player, the receiver. Any amount sent is increased by the experimenter before being allocated to the receiver, who then decides whether to return some, all or none of this (increased) amount to the sender. While highly stylized, the trust game is an appropriate context because it captures the essential nature of our motivating examples: a pareto-improving exchange is possible, but comes with the risk of opportunistic counterparty behavior which cannot be eliminated through pre-play promises or contracts.

The timing of our main experiment is as follows: first, we have participants play a slightly-modified trust game; after playing the trust game, we ask participants directly about their personal, subjective, cheating notions;[5] finally, we elicit participants' beliefs about others' cheating notions and behavior, as well as their first- and second-order beliefs about others' behavior and beliefs. The experiment is conducted using a full strategy method. Participants submit their complete contingent strategies for both the sender and receiver trust game roles, their cheating notions, as well as all beliefs before knowing which role they will be assigned. In addition to this main experiment, we run multiple additional experiments (including a direct-response, between-subjects experiment) to provide robustness checks.[6]

[4] For example, in a seminal work in this vein Berg, Dickhaut and McCabe (1995) explicitly posit that trustors will feel cheated by a negative return on their trust-investment. This often unstated assumption continues to pervade the trust literature: the outcome chosen to highlight the existence of aversion to "betrayal," or what we would call cheating, is one that falls just below yielding a positive return on investment (see, e.g., Bohnet and Zeckhauser, 2004).

[5] We realize that asking about cheating notions directly gives rise to concerns about priming. We check for the robustness of directly-revealed cheating notions in additional robustness sessions where cheating notions are elicited indirectly in a way that reduces the likelihood of priming effects.

[6] The between-subjects experiment is described in Section 2.6. Details on all of our other experiments are provided in the Appendix.

In our analysis, we complement the data from our main experiment with data on the values our participants' parents emphasized during their upbringing. These values data were collected for a previously-conducted, unrelated, experiment that took place *from twenty days to sixty days prior* to the trust game experiment. We use these data to investigate one potential source of stable heterogeneity in cheating notions: cultural transmission.

We test several hypotheses. At the most basic level, we test whether trust-based exchanges do indeed engender personal cheating notions. Whether or not this will be the case is not *a priori* obvious: cab drivers, mechanics and financial advisors may very well choose to ignore or downplay the possibility that their customers could ever feel cheated in order to reconcile opportunistic behavior with a positive self-image.[7] Conditional on an affirmative answer to our first question, we test the hypothesis that these implicit cheating notions have an impact in determining behavior in a trust-based exchange situation.

We find that the vast majority of participants articulate a cheating notion even when they can easily refrain from doing so, suggesting they are genuine. We document these notions, showing they are roughly bimodal: many participants define cheating by a positive return on investment rule, as assumed but not tested by Berg, Dickhaut and McCabe (1995); while, contrary to the assumptions of much of the trust game literature, a sizable minority of senders (around 30% of participants) define cheating by a more demanding notion requiring fully half of their co-players' total earnings in order not to feel cheated.[8] We also show that this heterogeneity in cheating notions carries over to beliefs about others' cheating notions and that, moreover, the notion of cheating strongly affects behavior on both sides of the potential exchange.

An important question for our paper is the relationship between our results and guilt aversion. Building on the insights from the theory of guilt aversion, we could expect that receivers' behavior will be substantially constrained by an aversion to guilt arising from falling short of senders' expectations. Therefore, one might wonder: what is the value added by eliciting cheating notions compared to eliciting beliefs about actions? We take this issue seriously and tackle the problem in two ways. First, we explain how cheating notions are conceptually different from beliefs about actions. Second, we show empirically

[7] For evidence that individuals choose their beliefs to avoid cognitive dissonance, we refer the interested reader to the discussion in Akerlof and Dickens (1982).

[8] Because of the way we modified the trust game, this latter rule can be distinguished from previously documented fairness rules such as equal surplus division. For details, see the experimental design section.

that receivers' beliefs about senders' cheating notions have more explanatory power than second-order beliefs in explaining receivers' behavior.

Conceptually, our crucial distinction will be between mathematical and moral expectations. Cheating notions are an example of moral expectations, as they are value judgements about particular behaviors, whereas senders' first-order beliefs are mathematical assessments about the probability of particular events.

With this distinction in mind, empirically, we hypothesize that cheating notions (or *moral* expectations) are a more fundamental determinant of guilt than the *mathematical* expectations upon which guilt aversion theory is constructed (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007), so that moral expectations may represent a micro-foundation for guilt aversion theory. We test which beliefs constrain receivers' behavior more: their beliefs about senders' mathematical expectations or their beliefs about senders' moral expectations. Specifically, we hypothesize that receivers' beliefs about senders' moral expectations will constrain receivers' behavior in the way that guilt aversion models predict receivers' second-order beliefs will, acting as thresholds above which no guilt is engendered. Moreover, if moral expectations are truly a more fundamental determinant of guilt, then receivers' second-order beliefs should exert little influence on receivers' behavior once receivers' beliefs about senders' moral expectations are accounted for.

We find strong support in our data for the notion that beliefs about cheating notions provide a micro-foundation for guilt. In particular, we show that beliefs about others' cheating notions exhibit a close empirical relationship with the second-order beliefs that are central to the current theory of guilt aversion, but that second-order beliefs have little explanatory power for receivers' behavior beyond what is captured by beliefs about senders' moral expectations.

Our paper contributes to the literature in several ways. First and foremost, we provide the first direct evidence on the relationship between personal cheating notions and individual behavior in trust-based exchange opportunities. Secondly, we show that cheating notions provide a micro-foundation for guilt which has strong promise of lending empirical content to theoretical models of guilt aversion. Having documented the empirical and theoretical importance of cheating notions, another contribution is to provide evidence of a substantial role for culturally transmitted values in the formation of cheating notions and related beliefs.

Our data are consistent with the plausible notion that parentally instilled values exert a substantial impact on how individuals define cheating and that own cheating notions shape beliefs about others' cheating notions. Together, these patterns suggest there may be a considerable temporally stable component of cheating notions, adding to their predictive value.

The fourth contribution of our paper is the investigation of how cheating notion beliefs constrain the behavior of the entrusted. We find that the behavior of individuals who refrain from intentional cheating moves in one-to-one correspondence with their beliefs about others' cheating notions. On the other side of the exchange, we document a significant relationship between expected cheating and how much individuals trust.

More generally, our results contribute to the debate over how non-pecuniary preferences affect behavior and where these preferences come from. Receivers in the trust game face a stark trade-off between their pecuniary preferences and moral behavior. Our finding that receivers' behavior is affected by their beliefs about what constitutes cheating lends support to the view put forward by Gneezy (2005) and refined by Lundquist, et al., (2009): moral preferences are affected by the magnitude of damage that immorality inflicts on others.[9] However, because our experiment involves a game with neither communication nor unambiguous moral standards, and hence no literal lying nor deception, we extend Gneezy's findings by showing that the moral forces at work operate outside of the specific context of deception.

Our paper is also related to a nascent literature examining directly the relationship between behavior and social norms exemplified by Krupka and Weber (2013) and Reuben and Riedl (2013). Similar to our study, the aim of this body of research is to complement the copious indirect evidence that social norms affect behavior by directly eliciting these norms and relating the elicited social norms to observed behavior. Our study differs from this vein of research, however, in that we focus on personal cheating notions which may vary widely across individuals and require no tacit or explicit agreement about what is cheating and what is not. In stark contrast, social norms by definition require a "...general social agreement that some actions are more or less socially appropriate" (Krupka and Weber, 2013).

[9] Many popular and intuitive models of moral preferences are inconsistent with this pattern in behavior. For an elaboration of the inconsistencies, see Gneezy (2005).

Our paper is most closely related to Charness and Schram (2013), who distinguish between social and moral norms. Social norms depend on external observability and external sanctions to influence behavior, while moral norms may sway behavior through internal sanctions even without external observability. The authors induce shared norms in a dictator game by transmitting normative advice about what a decision-maker "ought" to do from disinterested "advice groups" to dictators. They find that advice affects dictators' behavior even without the external observability required by social norms, suggesting that moral norms are an important determinant of dictators' behavior. Cheating notions can be thought of as moral norms. In this light, by eliciting moral norms directly from both parties involved in a trust-based exchange we can examine the connection between one's own moral norms, others' moral norms and a host of potentially decision-relevant beliefs, complementing and extending the evidence on the existence of an influence of moral norms provided by Charness and Schram (*ibid*).

Finally, our paper relates to the huge literature investigating behavior in the trust game.[10] The bulk of this vast literature focuses on what drives senders' behavior – interpreting the amount senders send as trust, whence the moniker "the trust game" comes. What, precisely, senders are trusting receivers to do is typically left unspecified, but a common assumption – made explicitly in Berg, Dick and McCabe (1995) and implicitly in much subsequent work (e.g., Bohnet and Zeckhauser, 2004) – is that senders are trusting that receivers will send back at least as much money as they sent. To the best of our knowledge, this assumption has never been tested directly. Differently from most of this literature, rather than assuming a particular cheating notion is operative, we aim to document and study the roles that participants' moral expectations and related beliefs play in determining the behavior of *both* receivers and senders. In doing so, we can shed empirical light on the unresolved question of what it is that senders are trusting receivers to do, what receivers believe senders are expecting of them, and the determinants of receivers' behavior.

The remainder of the paper proceeds as follows: Section 2 details the design of our main experiment and outlines the design of our direct-response experiment; in Section 3 we present our results. Section 4 provides a more general discussion of our findings together with a simple analytical framework to interpret them. Section 5 concludes and suggests

[10] The trust game literature is too large and spans too many disciplines to be summarized here, but for an excellent review see Camerer (2003) and the references therein.

avenues for future research. Additional experimental treatments, analyses conducted to address the robustness of elicited cheating notions and a comparison between behavior in our main experiment – conducted on-line – and a smaller study conducted in a more traditional laboratory environment can be found in Appendix I. Appendix II provides instructions for our main experiment. Appendix III details the design of, and provides instructions for, our direct-response experiment.

2 Experimental Design

A total of 428 individuals participated in our main experiment, all of whom were students in Rome, Italy, enrolled at one of two universities: LUISS Guido Carli University or the University of Rome, La Sapienza. All sessions were conducted on-line.

The experiment consisted of three phases. First, participants played a slightly modified trust game. Responses were collected using the strategy method (Selten, 1967). Participants submitted their complete contingent strategies for both the sender and receiver roles before knowing which role they would be assigned.

The strategy method allows us to gather data on behavior in situations which rarely occur, which may be particularly important in our current context. The main drawback of using the strategy method is a potential "hypothetical bias:" responses to outcomes that have not yet occurred may not accurately reflect underlying preferences. In an early study of this problem, Brandts and Charness (2000) find little evidence for such a bias even in a game where non-pecuniary preferences have been shown to play a large role. In a subsequent analysis of a large number of published studies comparing the strategy method to the direct-response method, Brandts and Charness (2011) find that "...there are significantly more studies that find no difference across elicitation methods than studies that find a difference" (pg. 387). Moreover, behavior in games where players have large action sets – as is ours – are more robust to the elicitation method than games where players make, e.g., binary decisions. Balancing a concern with data quality, for which the best available research provides mixed evidence, against a concern for data quantity led us to employ the strategy method for our main experiment.

A second caveat about our design is that having participants submit strategies for both roles, before knowing which they have been assigned may raise its own concerns. The advantage is, again, the quantity of data we can collect: having participants play only one

role would cut in half the number of observations we could collect for either of the roles. The major drawback of this design choice is that it may change the behavior we observe. For example, Iriberri and Rey-Biel (2011) show that such "role uncertainty" tends to increase costly surplus creation and decrease spiteful behavior in a simple game similar to the trust game we study here. It is not clear which behavior represents true preferences, however, as in the real world we often have experience with both sides of trust-based exchanges. We routinely trust others to pay us for the services we provide (e.g., as professors) and, at the same time, are trusted by others – e.g., our participants – to pay them for their work. In either case, we can partially address concerns about role uncertainty with a subsequent experiment featuring no role uncertainty (described in Section 2.6, below).

As a final caveat, we point out that we adopt a within-subjects design. We feel this design choice was necessary given our objective of analyzing the individual-level relationships between a host of variables at a fine level of detail. Thus, we decided against a between-subjects design largely on the grounds of feasibility. A within-subjects design delivers the internal validity necessary for such analyses without having to rely upon the validity of randomization across what would have been a large number of treatments. Within-subjects designs carry with them a variety of challenges to external validity, however, central among them being the potential for spurious correlation introduced by exposing the same individuals to multiple stimuli (see, e.g., Charness, Gneezy and Kuhn, 2012). We attempt to address the most obvious threats and confounds directly (e.g., ex-post rationalization). Indirectly, the results from our direct-response experiment, which features a between-subjects design, will provide a modicum of reassurance about the external validity of our results.

After participants submitted their trust game strategies, we asked them about their personal cheating notions. Finally, we elicited participants' beliefs about others' behavior and others' cheating notions in an incentive compatible manner. During each of these three phases, participants were unaware of the existence of any of the subsequent phases. After all three phases were completed, participants were randomly paired and within each pairing roles were randomly assigned, determining outcomes.

2.1 Our slightly-modified trust game

Our trust game is standard in most respects: it is a two-player sequential moves game of perfect information involving a sender and a receiver. The sender moves first by deciding

whether to send some, all or none of a fixed endowment to the receiver. Any amount sent is increased by the experimenter before being allocated to the receiver, who then decides whether to return some, all or none of this (increased) amount to the sender. Pairings are random and anonymous.

Our trust game differs, however, in two important ways from the canonical trust game. First of all, we implement an unequal endowment design – senders (receivers) are endowed with 10.50 euros (0 euros). Secondly, while most trust games use a linear function to transform the amount sent into the amount received – typically, if a sender sends s, the receiver receives $f(s) = 3s$ – we implement a concave trust production function. In our trust game, when a sender sends s euros, the receiver receives $8\sqrt{s}$ euros.[11] Both of these modifications will allow us to distinguish among various *a priori* likely cheating notions. For example, a fairness notion that says "I am entitled to half of the surplus created from my actions" coincides with an egalitarian fairness notion ("everybody's final money outcome should be the same") in the standard trust game with equal endowments, but not in our unequal endowment setting. A concave production function will allow us to further distinguish among various common fairness rules which roughly coincide when using a linear function for low send amounts.[12]

An added benefit of using a concave production function is to provide a relatively smooth relationship between behavior and beliefs at the individual level, a feature which will prove useful when we examine the "intensive margin" of trust: how much to send conditional on sending something.[13] Aiding our identification of this intensive margin is one additional, more subtle, feature of our design. We introduce a small (0.50 euro) fixed sending fee in

[11] We restrict the sender's action set to include only integer amounts in order to produce relatively simple values (multiples of 5 cents) while, at the same time, maintaining concavity and surplus creation.

[12] For example, consider two possible cheating notions conditional on sending one euro: a positive return on investment notion; or an equal share of created surplus notion. Irrespective of the trust production function the former notion entails receivers returning 1 euro. The latter notion entails receivers returning $\frac{f(s)}{2}$, which is 1.50 euros when $f(s) = 3s$ but $\frac{8.05}{2} = 4.025$ euros using our concave trust production function. Consequently, these two notions would differ by only 0.50 euros using the standard trust production function, while, in stark contrast, our concave function separates these two cheating notions by just over 3 euros.

[13] For instance, if senders have standard risk-neutral preferences a linear trust production function often implies corner solutions: send the entire endowment if the expected net return from trusting is positive, or nothing if the net return is negative; if the expected return is zero, then all send amounts are optimal. In contrast, our concave production function provides such senders unique internal optimal send amounts that vary continuously with expected return over a wide range of beliefs. In this sense our concave function may provide a more realistic portrait of trusting behavior outside of the lab with stakes large enough for risk aversion to matter. Consequently, an additional justification for using a concave trust production function is to induce risk-averse preferences (Smith, 1976).

some sessions: in "high fee" sessions, senders who choose to send a strictly positive amount incur the fee whereas senders who chose to send nothing do not; in the remaining "low fee" sessions, senders never incur a sending fee. This provides exogenous variation in the cost of sending something versus nothing – the extensive margin of trust – which will allow us to formally model and estimate the intensive and extensive margins of trust separately.

Senders' feasible actions consisted of sending any whole-euro amount, including 0. Conditional on receiving $f(s) > 0$ euros, receivers' feasible actions were $\{0.00, 0.01, \ldots, f(s)\}$. We employed the strategy method to collect participants' trust game decisions: before discovering whether they would play the role of sender or receiver, participants submitted a complete contingent strategy for each role. Each participant specified how much they would send in the role of sender and, for each possible amount they could receive in the role of receiver, how much they would return. The order in which participants submitted their strategies – whether first for sender, then for receiver or first for receiver and then for sender – was randomized. Additionally, to bridge the gap between the strategy method and the direct-response method and to attempt to make each receiver's decision feel as real as possible, participants' receiver strategies were elicited with a series of ten separate screens. Each of these ten screens asked only one question: "if the sender sends s euros and you therefore receive $f(s)$ euros, how much will you return?"[14]

2.2 The cheating notion questions

After participants submitted their complete contingent trust game strategies, we asked them to specify their personal definitions of cheating from the perspective of the sender. For each possible strictly positive send amount, $s \in \{1, 2, \ldots, 10\}$, participants were asked:[15]

> "If you are assigned the role of A [sender] what is the minimum amount you would need to receive back from player B [receiver] in order to not feel cheated? ... If you were to send s euros and B were to therefore receive $f(s)$ euros, you would need back how many euros?"

[14] The order in which receivers faced their ten separate decisions was randomly predetermined but the same for all participants. This maintains comparability across observations without inducing undue consistency in receiver strategies that might arise from, e.g., facing a monotonically increasing or decreasing sequence of send amounts.

[15] In each question "s" and "$f(s)$" were replaced by the appropriate numbers. The words "sender" and "receiver" did not appear on participants' screens.

To respond, participants could either insert a number between 0.00 and $f(s)$ or refrain from specifying a cheating notion by selecting one of two options: "this has nothing to do with cheating;" or "I don't know." Leaving the question blank was also allowed, but not explicitly mentioned as an option.[16]

Some may argue that by asking participants about cheating so directly we may prime them to associate behavior in the trust game with cheating. To address this concern we ran additional sessions in which, rather than asking our direct cheating notion question above, we asked participants to state how they would feel about various send/return combinations if they were to be assigned the role of sender. The results support the idea that priming is not the driver of reported cheating notions (see Appendix I, section A.2.)

Another potential concern with how we elicit cheating notions is that the same individuals who play the game are also asked to report their cheating notions. We made this decision in order to mitigate hypothetical biases stemming from individuals' inability to fully anticipate which outcomes will make them feel cheated without actually playing the game and thereby having pecuniary incentives to understand the consequences of one's own and others' actions. However, some might argue that asking the participants themselves about their cheating notions could bias the reported cheating notions in some other way and that, instead, it would be preferable to ask disinterested parties about what constitutes cheating. The only study we know of that examines this issue directly in the context of a trust game is Rustichini and Villeval (2012). As part of their study the authors describe a trust game to disinterested parties who then, for two specific send amounts, report the interval of return amounts they would consider fair. These same individuals come back the following week, play the trust game and again report their fairness intervals. Comparing the lower bounds of these intervals – the closest analogue to the cheating notions we elicit – between disinterested (first week) and involved (second week) parties reveals little difference. In line with this, we feel that having participants report their cheating notions directly after playing the game, while it is still fresh in their minds, is warranted.

[16] Our design initially did not include the two explicit opt-out responses mentioned above. Although responding to the question was always completely voluntary, we realized that not providing pre-programmed opt-out responses could make some participants feel obligated to supply a cheating notion even if they did not truly have one. To address this concern, we inserted the two opt-out responses described above. The majority of participants – 306 out of 428 – took part in sessions featuring the explicit opt-out responses. The remaining 122 participants took part in sessions with no explicit opt-out opportunity. Unless otherwise specified, our analyses utilize all 428 observations. In Appendix I we show that our results are robust to restricting attention only to sessions with explicit opt-out.

2.3 The beliefs elicitation phase

Following the cheating notions questions, participants discovered there would be a beliefs elicitation phase of the experiment and that they could earn additional money according to the accuracy of their estimates. In this phase, each participant was asked to estimate: i) how much other senders would send on average; ii) how much other receivers would return on average; iii) their beliefs about others' beliefs about how much receivers would return (second-order beliefs); iv) other participants' cheating notions; and v) the proportion of other participants who would not cheat them, according to the respondent's own subjective cheating notion (see Appendix II for exact wording).[17] For all belief elicitation questions, participants were instructed to exclude their own actions from their estimates and were told that the accuracy of their estimates would be calculated excluding their own strategies and cheating notions.[18]

Participants were informed that one estimate from this section would be chosen to count toward their potential earnings. This chosen belief was remunerated according to a randomized quadratic scoring rule (Schlag and van der Weele, 2013) which is both incentive compatible and theoretically robust to risk preferences. The mechanism was explained to participants in detail. Additionally, participants were told that it was monetarily in their best interest to report their true beliefs and provided with an example illustrating this assertion. An exactly correct belief paid 5 euros in most sessions while, in the remaining sessions, an exactly correct belief paid 20 euros. Beliefs were elicited *after* participants submitted their complete contingent strategies, but *before* knowing their assigned roles.

Eliciting beliefs after game-play and after having elicited cheating notions raises several potential concerns. Central among these are ex-post rationalizations of beliefs about others' cheating notions or others' expected returns. For example, participants could ex-post rationalize returning only a little by reporting they believed others expected little back, or by reporting that others needed only a little back in order to not feel cheated. We treat these concerns extensively using several different robustness check exercises. Full details are reported in Appendix I, Section B.

[17] Items ii)-v) were asked for each possible send amount.

[18] This was done to avoid mechanical correlations between reported beliefs and participants' own strategies or cheating notions.

2.4 Payment phase

After all three phases of the experiment were completed, pairings were randomly determined and, within each pair, roles were randomly assigned. Outcomes and potential earnings were determined by combining, within each pair, the sender's strategy with the receiver's strategy.

We randomly selected the approximately 10% of participant pairs who would be paid their potential earnings in the following manner, which was described to participants before they began the experiment. Each participant was randomly assigned a whole number between 0 and 100. Each whole number was equally likely to be selected. If either the participant himself/herself or the participant's co-player was assigned a number weakly less than 5, that pair of participants would be paid their experimental earnings. By selecting participant pairs rather than individual participants to pay, we ensure that decisions are consequential: whenever a decision actually affected a participant's own earnings, it also affected his or her co-player's earnings.

At the end of each session, after outcomes were determined all participants were sent a common e-mail providing a link to a website where they could discover all personally earnings-relevant information about their experimental outcomes: the role they were assigned; the action of their co-player; and which beliefs question was chosen to count as well as how much they earned from this question. Importantly, by entering their own unique experimental code participants could learn their co-player's experimental code as well as the whole number each of them was randomly assigned. This feature provides some credibility to our payment selection procedure: by entering the co-player's experimental code, a participant could verify that his information matched his or her co-player's information.

Irrespective of the credibility of selection, choosing only 10% may seem low. However, the experiment was relatively short and convenient, requiring on average about half an hour of participants' time. Furthermore, note that Italian students' opportunity costs are relatively low. As an example, work-study positions at one university in Rome we are familiar with typically pay students around 5 euros per hour. Given both of these observations, we feel the expected earnings from the experiment are commensurate with participants' opportunity cost of time. Despite this, we also conducted a handful of traditional in-lab sessions. We had participants come to the lab and complete the on-line experiment. In these in-lab sessions, 100% of participants were paid their experimental earnings. Participants' behavior in our in-lab trust game was remarkably similar to behavior in our on-line

study data, providing some reassurance that the monetary incentives in our main study were sufficient. For example, neither average send amounts, nor return proportions nor beliefs about the proportion of non-cheaters in the population differed significantly across these two environments (see Appendix I, Section A.1).

2.5 Instilled values and risk attitudes

For each participant in our main study, we complement the experimental data with data from a previously conducted, unrelated, survey. This survey contains basic demographic information, a (self-reported) measure of the emphasis each participant's parents placed on various normative values during his or her upbringing as well as an incentive-compatible measure of risk aversion (Holt and Laury, 2002).

There was a considerable time lag between the survey and the start of our trust game experiments (from 20 to 60 days) so that survey responses are unlikely to have affected trust game behavior directly. On the other hand, this temporal distance was small enough so that traits, such as risk aversion or instilled values, likely did not change in the meantime. This survey data allows us to control for risk aversion and altruism when examining sender behavior, while instilled values will prove useful in examining what drives receiver behavior.

In Table 1, we summarize key features of the main experiment. Descriptive sample statistics are reported in Table 2.

2.6 Direct-response experiment

One may be worried that the within-subjects design of our main experiment, or the use of the strategy method rather than the direct-response method, introduce spurious correlations among our measures which may be driving our results (see the discussion in Charness, Gneezy and Kuhn, 2012). One could also be concerned about the unintended effects of eliciting individuals' beliefs about the behavior of the experimental population rather than about the behavior of their specific co-players.[19]

To address these concerns simultaneously, we ran an additional experiment, where we used the direct-response method coupled with a between-subjects design. As this drastically reduces the amount of data generated per participant, we implemented a simplified trust game, restricting the sender's action set to $\{0, 5, 10\}$. For comparability with our main

[19] While this could be warranted under the assumption that individuals believe their co-player is representative of the experimental population, we do not know whether this assumption is true.

experiment, however, we retained the quadratic trust production function so that receivers could receive three possible amounts: $\{0, 17.90, 25.30\}$. This simplified trust game balances a concern for generating a reasonable number of observations for each separate send amount against a desire to allow senders meaningful variation in their trust decisions.

We conducted the experiment in the laboratory at the Einaudi Institute for Economics and Finance using pen and paper. The experiment consisted of two treatments: in one we elicited and transmitted senders' cheating notions; in the other, we elicited and transmitted senders' first-order beliefs about their receiver's action.[20]

3 Results

We establish four main results: i) there is substantial heterogeneity in cheating notions and beliefs about others' cheating notions; ii) intergenerationally transmitted values are important determinants of cheating notions; iii) cheating notions affect decisions on both sides of the trust exchange; iv) cheating notions beliefs have more explanatory power than second-order beliefs in explaining receivers' behavior. We therefore show that cheating notions are a more fundamental determinant of guilt and that understanding them may provide a micro-foundation for guilt and guilt aversion theory.

Before starting our analysis, it will prove useful to fix short acronyms for some of our variables. We will typically denote the sender's action by s and the receiver's action by $r(s)$. In addition to behavior in the trust game our experiment produces five (sets of) variables of primary interest.

The first set of variables consists of each participant's cheating notion in the role of sender (described above) which we label *Cheat_notion*. Second, for each $s \in \{1, \ldots, 10\}$, we measure participants' beliefs about other participants' cheating notion, which we denote by *B_Cheat_notion*. The third set of variables of interest is participants' beliefs, one for each $s \in \{1, \ldots, 10\}$, about how much receivers will return if the sender sends s. We label this set of beliefs collectively as *B_Receivers_actions*. Fourth, we elicit each participant's beliefs about other participants' beliefs about receivers' action for each possible s, labeling this set of variables *B_B_Receivers_actions*. Finally, we measure each participant's belief, from the perspective of the sender role, about the chances of not being cheated, denoting

[20] Both the cheating notion question and the (first-order) belief question were similar to the questions used in our main experiment, but adapted to refer only to the sender's chosen send amount and the sender's specific co-player. Details on the questions are provided in Appendix III.

this measure by *B_NotCheated*. The questions associated with each of our variables are described in Table 3.

3.1 Descriptive evidence on cheating notions and related beliefs

We start our analysis by documenting a few results concerning cheating notions and related beliefs directly. We show that: i) the trust game indeed gives rise to well-defined cheating notions (*Cheat_notion*) for the vast majority of our participants; ii) there is considerable across-individual heterogeneity in these cheating notions as well as within-individual consistency across send amounts; and that, iii) the same pattern – across-individual heterogeneity and within-individual consistency – obtains for beliefs about others' cheating notions (*B_Cheat_notion*). The last result is the most important, as it would seem to be a necessary prerequisite for moral expectations to exert a substantial and predictable influence over receiver behavior.

We start by remarking that the vast majority of participants – about 80%, averaging across all send amounts – report a personal cheating notion even in sessions where refraining from specifying a cheating notion is salient and simple (see fn. 16). Restricting attention to sessions involving explicit cheating notion opt-outs, the proportion of senders selecting the option "this has nothing to do with cheating" ranges from a low of 13 percent when considering sending 10 euros, to a high of 20 percent when considering sending one euro. The proportion of senders who opt out of reporting a cheating notion for *any* reason – which includes selecting either "I don't know,"or "this has nothing to do with cheating,"or just leaving the question blank – in these same sessions is also low, ranging from 17 percent to 23 percent. Apparently, few participants have no opinion one way or the other. Moreover, these patterns suggest that for a large majority of our participants being cheated is a well-defined event. These proportions are summarized in Table 4.

Turning from existence to heterogeneity, in Figure 1 we plot histograms of cheating notions for each send amount separately, restricting attention to those who responded with a number. We overlay each histogram with vertical lines representing two *a priori* plausible cheating notions. The first vertical line represents the cheating notion most commonly assumed in the trust literature: a weakly positive return on investment rule.[21] An individual whose cheating notions are consistent with this rule would for each send amount, $s \in$

[21] This is the cheating standard explictly assumed in Berg, Dickhaut and McCabe, 1995 and incorporated into much of the subsequent literature on trust (*cf.* Bohnet and Zeckhauser, 2004).

$\{1, \ldots, 10\}$, report a cheating notion of exactly s, feeling cheated for any return amount strictly less than s but not feeling cheated for any return amount weakly greater than s. The second vertical line represents an equal split of the receivers' entire earnings – i.e., for each s, the line is placed at $\frac{f(s)}{2}$. Accordingly, we call this an "equal split" cheating notion.[22] One justification for this cheating notion is that individuals may generally feel entitled to an equal share of all of the money their actions generate, which *could* be interpreted as the receiver's entire earnings above the receiver's initial endowment.

As is evident from the histograms, there is quite a lot of heterogeneity in personal cheating notions, suggesting that the typical ad-hoc assumption of a uniform standard of cheating in trust-based exchange is unwarranted. The histograms suggest that cheating notions are, instead, roughly bimodal with much of the mass concentrated between the weakly positive return on investment and the typically much more demanding equal split cheating notion. Consequently, while the weakly positive return on investment may serve well as a lower bound on behavior generating the feeling of being cheated, a lot of information on individual heterogeneity is lost by assuming that most individuals' cheating notions coincide *exactly* with this rule.

An important question is whether cheating notions are consistent at the individual level across possible send amounts. Such stability would provide a modicum of reassurance that moral expectations reflect some underlying individually stable trait. To get at this question, we first restrict the attention to individuals whose cheating notion is consistent with an equal split cheating notion for a send amount of 1 – the send amount providing the widest separation between equal split and weakly positive return on investment. About one-third (33%) of our participants report cheating notions consistent with an equal split rule conditional on $s = 1$. For this third of participants, we plot histograms of cheating notions across all other send amounts (Figure 2A), showing a striking amount of consistency.

We repeat this exercise for individuals whose cheating notions are consistent instead with a return on investment rule for send amount 1.[23] We find that, again, cheating

[22] However, recall that since our design features unequal initial endowments, this notion should not be confused with inequality aversion or egalitarianism. Instead, demanding half of the receivers earnings typically implements a lot of inequality in final earnings. For example, if the sender sends 1 euro, the receiver receives 8.05 euros. An individual with an equal split cheating notion would feel cheated by receiving less than 4.02 euros back which would correspond to (sender earnings, receiver earnings)= $(13.02, 4.03)$.

[23] To be generous to the idea that participants define being cheated according to some notion of return on investment, we expand the definition to allow for a strictly positive, yet reasonable, return on investment of no greater than 10% and, at the same time, take into account whether or not a session involved a sending

notions are consistent with a positive return on investment for about one-third (31%) of our participants when $s = 1$. For this approximately one-third of participants, we report in Figure 2B histograms of cheating notions for all other send amounts. There is still a considerable amount of consistency across send amounts.

Having documented both the heterogeneity and individual consistency of the cheating notions engendered by trust-based exchange, we next ask whether these features carry over to individuals' beliefs about others' cheating notions (B_Cheat_notion) and beliefs about senders' beliefs about receivers' actions, i.e., second-order beliefs ($B\_B\_Receivers\_actions$). Beliefs about others' cheating notions and second-order beliefs follow much the same distribution as cheating notions themselves (Figures 3 and 4). In the appendix, we also report analogous within-individual consistency exercises for both first-order and second order beliefs, finding a remarkable amount of consistency.[24]

All together, the data suggest considerable heterogeneity and consistency in cheating notions, first order and second order beliefs.

3.2 What Determines Cheating Notions?

So far our data suggest the existence of substantial heterogeneity in cheating notions ($Cheat\_notion$) and beliefs about others' cheating notions (B_Cheat_notion). Our data also reveal that individuals tend to expect a positive relationship between their own cheating notions and others' cheating notions. We try to understand the reason for this link, emphasizing the relevance of a substantial stable, culturally transmitted component.

We start with a plausible conjecture based on previous research about belief formation: in novel situations introspection substitutes for information so that through the well-established psychological phenomenon known as "false consensus" one's own cheating notion may become a significant determinant of beliefs about others' cheating notions (see, e.g., Ross, Green and House, 1977; in a trust game context, see Butler, Giuliano and Guiso, *forthcoming*). If own cheating notions themselves are persistent – perhaps being based on moral values which tend to be culturally transmitted from parents to children (see e.g. Bisin and Verdier, 2010) – then cheating notion beliefs may also persist over time and context. There are two links in this chain: i) values to cheating notions; and ii) cheating notions to

fee.

[24] See Figures A1a and A1b for first-order beliefs and Figures A2a and A2b for second-order beliefs (all in the Appendix).

cheating notion beliefs. We provide evidence on both links.

Starting from the second link, there is an abundance of evidence in our data suggesting that own cheating notions contribute substantially to cheating notion beliefs. The correlation between *Cheat_ notion* and *B_ Cheat_ notion* ranges from 0.53 ($s = 4$) to 0.66 ($s = 1$) and is always highly significant ($p < 0.01$).[25]

Proceeding backwards, to investigate the first link we test directly for a relationship between our participants' cheating notions and the values their parents emphasized during their upbringing while controlling for a variety of demographic variables. We use data from a previously conducted unrelated survey (described in Section 2.5) which included a rich set of parentally instilled normative values. For each normative value in this set, survey participants were asked to state how much emphasis their parents placed on this value during their upbringing which we take to be a proxy for received cultural values.[26] Valid responses ranged from 0, which indicates no emphasis, to 10 which indicates quite a lot of emphasis.[27] For our estimates, we select a relevant subset of these normative values and organize them into two categories: "cooperative" and "competitive." The former category includes such values as helping others and honesty. The latter category includes, for example, the value of striving to be better than others.[28] We construct an index of parents' emphasis on "cooperative" and "competitive" values by taking the average emphasis over all the values constituting each category. This yields a measure for each category theoretically ranging from 0 to 10. We divide each of these measures by 10 obtaining an index on a 0 to 1 scale.

To get a summary measure of the relationship between instilled values and own cheating notions, we pool over all send amounts and regress cheating notions on cooperative and

[25] Similar results are obtained by regressing cheating notion beliefs on own cheating notions for each send amount separately, while controlling for available demographics. The coefficient on own cheating notions ranges from 0.52 to 0.57 and is always significant at better than a 1% level. Results are available from the authors.

[26] We acknowledge that such self-reported retrospective questions are likely to be noisy or biased 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.

[27] Participants could also respond "I don't know," which we code as missing.

[28] The full set of "cooperative" values is: i) behave as a model citizen; ii) help others; iii) group loyalty; iv) always give others their fair share; v) always tell the truth; vi) always keep your word. "Competitive" values are: i) always extract the maximum advantage from every situation; ii) seek to be better than others; iii) act so as to induce good in others (e.g., scold somebody who litters).

competitive values. Since pooling in this manner results in multiple observations for each participant we incorporate individual-level random effects in our model. As the presence of an investment fee may directly affect cheating thresholds we also include a dummy for sessions with no investment fee. Finally, because our trust production function is non-linear (concave) in money sent, we allow cheating notions to vary non-linearly with money sent by adding a quadratic term to the estimated equation.

The estimates reported in Table 5 reveal a substantial relationship between values and cheating notions. Interestingly, our data suggest that the two classes of values we consider pull in opposite directions. Instilled cooperative values significantly lower cheating notions: the more emphasis parents placed on cooperative values, the fewer euros senders need back in order to not feel cheated. Competitive values, on the other hand, have the opposite effect, raising cheating notions significantly.

While these relationships may appear counter-intuitive at first glance, they are both plausible. For example, if more cooperative individuals internalize others' outcomes to a greater extent than less cooperative individuals, then *ceteris paribus* more cooperative individuals may be satisfied with lower own-earnings if this raises others' earnings. At the same time, if competitive individuals place a high value on earning more than others, then they may be less satisfied with, and more likely to feel cheated by, outcomes where others earn more and they earn less. The latter pattern will tend to raise cheating notions while the former pattern will lower cheating notions, which is what we observe in the data.[29]

Controlling for instilled values, cheating notions tend to move one-for-one with the amount sent, suggesting that a positive return on investment rule is the baseline cheating notion and that values determine how far individuals deviate from this baseline. Finally, there is little evidence that cheating notions vary by demographics once we control for values.

Summing up, our data suggest that parentally instilled values are significant predictors of cheating notions and that cheating notions, in turn, are highly significant predictors of cheating notion beliefs, lending some credence to the idea that cheating notions and related beliefs are stable predictors of behavior. Consequently, in the next subsection, we focus on the relationship between cheating notion beliefs and behavior.

[29] For corroborating evidence that instilled values affect behavior, see Butler, Giuliano and Guiso (*forthcoming*).

3.3 The relationship between cheating notions beliefs and behavior

In this section we look at the effect of cheating notion beliefs on the behavior of both receivers and senders.

3.3.1 Receivers' Decision to Intentionally Cheat

One advantage of focusing on cheating notion beliefs directly is that we can study what drives receivers' decision to intentionally cheat. We can address this latter question directly because we know when receivers cheat according to their own estimates of others' cheating definitions.

Since our measure of cheating notions asks about cheating directly, and because our *B_ Cheat_ notion* measure asks participants for their beliefs about *others'* cheating notions, explicitly instructing respondents to omit their own cheating notions from consideration, we can be somewhat confident that *B_ Cheat_ notion* reflects what *receivers themselves* believe senders will consider cheating.

As a first pass, we construct a dummy variable taking the value of 1 whenever $r <$ *B_ Cheat_ notion* and 0 otherwise, for each amount $s \in \{1, \ldots, 10\}$, interpreting this dummy as an indicator of intentional cheating. We then relate this intentional cheating indicator to receivers' demographic characteristics and their own cheating notions (*Cheat_ notion*) as well as to their beliefs about senders' cheating notions, *B_ Cheat_ notion*.

Table 6 presents our estimates of receivers' propensities to intentionally cheat for each possible send amount. Participants' demographics have few consistent effects on cheating across different send amounts: older participants generally cheat less for lower send amounts; smarter participants – those who have higher math scores – are less likely to cheat for high send amounts. Interestingly, gender plays no role. On the other hand, controlling for receivers' expectations about senders' cheating notions (*B_ Cheat_ notion*), receivers that have higher own standards – i.e., who would feel cheated unless they were given back a lot when playing as senders – are consistently less likely to cheat across all send amounts. We interpret this finding as saying that more demanding people tend to refrain from cheating others, behaving according to the principle "do not do to others what you would not want others to do to you." Notice, however, that conforming to this principle is cheaper when amounts sent are low and the temptation to deviate from it (and doing to others what you would not want them do to you) is thus weaker. Consistent with this we find that the

effect of receiver's own cheating notions (*Cheat_notion*) is stronger for lower levels of s: the reported probit coefficients imply that the marginal effect of an increase in *Cheat_notion* at send amount 10 is half that at send amount 1 (1.6 percentage points vs 3.6 percentage points, respectively).

3.3.2 The Effects of Cheating Beliefs on Senders' Behavioral Trust

As a second step, we consider whether and how the specter of being cheated affects senders' behavior. While previous research suggests that expected cheating or betrayal may affect the likelihood of trusting behavior (e.g., Bohnet and Zeckhauser, 2004), it is an open question whether the likelihood of being cheated affects the intensive margin of trust – i.e., *how much* to trust, conditional on trusting at all. This is an important distinction as it speaks to the potential benefits that may obtain in terms of surplus creation from policies aimed at reducing cheating. For example, if expected cheating determines the extensive margin of trust only, then there may be little to gain from reducing cheating in environments where most people already exhibit at least some small amount of trust.

To examine whether anticipating being cheated affects the intensive margin of trust, for each participant we construct a unidimensional measure of his or her beliefs about the proportion of non-cheaters in the (experimental) population. We do this by constructing each sender's average response to our set of ten *B_NotCheated* measures. The resulting measure of beliefs about population trustworthiness theoretically ranges from 0 to 1, with 1 indicating the sender believes no receiver will cheat for any send amount (all are trustworthy) and 0 indicating all receivers will cheat for every send amount (none is trustworthy).[30] The measure can therefore be interpreted as subjective probability of not being cheated. We call this measure *Pr(NotCheated).*

Figure 5 plots the kernel density of this probability separately for opt-out and no-opt-out sessions. We document a modal value at around 0.5 (almost equal to the fraction of non-cheaters in the pool – see Table 2, bottom row) irrespective of opt-out opportunities. In sessions with opt-out (the dashed line), a second mass of observations centers around a

[30] Individuals who did not report a cheating notion conditional on sending s euros are coded as missing. In this case our elicitation mechanism is not incentive compatible since we cannot observe whether such an individual will feel cheated. Given these caveats, we construct a unidimensional measure of beliefs about population trustworthiness for 401 (out of 428) participants. For those individuals who respond that sending S euros "...has nothing to do with cheating," we assume that they *cannot* feel cheated regardless of the receiver's decision. Therefore, we code such individuals' population trustworthiness belief conditional on sending s euros as 1 before constructing their summary measure.

value of 1, reflecting (mechanically) the small minority of participants who report the trust game "has nothing to do with cheating" consistently.

In an analogous fashion, we construct for each participant a summary measure of his or her beliefs about the proportion of the money they send that will be returned to them. For each $s \in \{1, \ldots 10\}$ we divide the participant's estimated return *amount*, *B_receivers_actions*, conditional on sending s euros by s to get their estimated (gross) return proportion. We then average their 10 return proportion estimates to get a unidimensional measure of return proportion beliefs. We interpret this index, *B_return_proportion*, as a measure of senders' expected (gross) return proportion.[31]

Finally, using these two summary measures we estimate a model of how much senders send as a function of the senders' expected return proportion, their beliefs about being cheated and an interaction between these two variables. We control for our standard set of demographics. To account for selection into sending a positive amount we estimate a Heckman model and exploit variation in the investment fee across sessions to construct the selection equation. Specifically, the exclusion restriction for the selection equation consists of a dummy for "Low fee" sessions where the investment fee was zero. Importantly, because two common alternative explanations for senders' behavior in the trust game are risk preferences and altruism, among our demographic controls we include an incentive-compatible measure of risk aversion collected from the survey described in Section 2.5 as well as a proxy for altruism obtained from that same survey.[32]

Table 7 presents the estimates. The second column presents the selection equation, which is a probit model estimate of the decision to send something versus nothing, i.e., of the extensive margin of trust. As desired, this extensive margin depends significantly on the presence of a sending fee. The first column presents the main equation which estimates the intensive margin of trust formally accounting for selection into sending a positive amount. Here, the estimate implies that the specter of being cheated plays a significant role in the intensive margin of trust: the positive and significant coefficient on our measure of the expected probability of *not* being cheated indicates that when senders believe it is less likely that they will be cheated, they send more. The implied effect of non-cheating beliefs on

[31] This summary measure, which ranges from a low of 0.00 to a high of 4.02 with a mean of 1.27 and a standard deviation of 0.64,is also nearly identical to actual gross return proportions (Table 2).

[32] We use as our measure of altruism the emphasis, on a scale from 0 to 10, participants' parents placed on the value of "helping others" during their upbringing.

behavioral trust is non-trivial: increasing *Pr(NotCheated)* from 0.1 to 0.9 is associated with an increase in the average amount sent equal to 51% of the sample mean; ignoring interaction effects and non-linearities, increasing this belief by 50 percentage points is roughly equivalent to decreasing our measure of risk aversion from its maximum value of ten (very risk averse) to its minimum value (risk loving). The coefficient on senders' expected (gross) return proportion is also positive and significant, indicating that standard pecuniary concerns also drive senders' behavior. Finally, the negative and (marginally) significant coefficient on the interaction between expected returns and non-cheating beliefs suggests that as expected pecuniary returns increase, the negative impact on trust of expected cheating subsides. In other words, the sting of expected betrayal can be soothed by money.[33]

3.4 Cheating notions and guilt aversion

A central piece of guilt aversion theory (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) is the relevance of second-order beliefs. In this literature, guilt is the result of disappointing others with respect to their formal, mathematical, expectations of counter-party behavior: person A is disappointed whenever person B's action falls short of A's expectations of B's action. Consequently, B's second-order beliefs – B's beliefs about A's beliefs – shape the set of possible equilibria.

The idea that violating others' expectations can give rise to guilt has strong intuitive appeal. This may be partially due to the fact that "expectation" is often used in two, easily conflated, ways. As in the description of second-order beliefs above, the term "expectation" can denote a formal mathematical construct – the probability weighted average of possible outcomes. At the same time, "expectation" also has a less mathematical – more subjective and moral – meaning. For example, the *Oxford English Dictionary* lists "[t]o look for as due from another" as one meaning of expect, while *Merriam-Webster* offers the definition "to consider bound in duty or obligated" along with the example sentence "[t]hey expect you to pay your bills."

For clarity of exposition we will refer to the first meaning of expectations as "mathematical expectations" and the second as "moral expectations." We consider cheating notions to be an example of *moral* expectations, while senders' first-order beliefs about receivers' actions are an example of *mathematical* expectations. Using this terminology, existing the-

[33] The results are virtually the same if we estimate a Tobit model of send amounts, which intuitively models selection as censoring.

ories of guilt aversion rely on beliefs about others' *mathematical* expectations to define the threshold of behavior engendering guilt. We, on the other hand, hypothesize that beliefs about others' *moral* expectations are a more fundamental determinant of guilt. Consequently, we would argue that moral expectations may provide a micro-foundation for guilt and guilt aversion theory.

Mathematical and moral expectations are clearly conceptually distinct: the former is an assessment about the likelihood of possible outcomes, while the latter is a value judgment about particular outcomes. Still, one might anticipate that these two types of expectations are empirically correlated. A correlation may come about through several channels. For example, as most of our daily interactions do not involve being cheated, induction or Bayesian updating may lead individuals to *mathematically* expect to not be cheated.[34] Consequently, when individuals construct their mathematical expectations, outcomes satisfying the individual's moral expectations may receive the lion's share of the weight, inducing a mechanical correlation between moral and mathematical expectations. Alternatively, correlations between senders' moral and mathematical expectations can be generated from a simple fixed cost of cheating model with common knowledge of cheating notions. Taking this logic one step further, if senders' mathematical expectations are correlated with their moral expectations, then receivers' (second-order) beliefs about senders' mathematical expectations and receivers' (first-order) beliefs about senders' moral expectations should be empirically correlated as well.

In this section we show that not only are senders' mathematical and moral expectations correlated but that this correlation is reflected in receivers' beliefs – *B_Cheat_notion* and *B_B_Receivers_actions* – as well. Establishing the existence of an empirical correlation between moral expectations, mathematical expectations and related beliefs and, in the process, demonstrating that moral expectations and related beliefs are an important source of expectations about how others will behave is important: if first-order (second-order) beliefs about others' actions (beliefs) are closely empirically related with personal cheating notions, then knowledge about the distribution of personal cheating notions in a population can provide insight into which of the multiple equilibria typically predicted by guilt

[34] We provide evidence in Appendix I, Section C, that reported cheating notions are not "reverse-caused" in this respect: i.e., that participants do not form beliefs about the amounts participants return and then simply report this belief as their cheating notion as to avoid, e.g., looking liking a sucker. Essentially, we show that cheating notions are no more correlated with beliefs for outcomes which may actually happen – where looking foolish is a possibility – than for outcomes that are impossible.

theoretical models are most likely to occur.

We test for the conjectured correlations between i) own cheating notions and beliefs about receivers' action; and ii) beliefs about others' cheating notions and receivers' second order beliefs. For the sake of brevity, we report details and results of these tests in the appendix.[35] The main lesson from our exercise is that senders' own cheating notions, *Cheat_ notion*, are consistently highly significant predictors of senders' beliefs about receivers' actions (*B_ Receivers_ actions*) and that receivers' beliefs about senders' cheating notions (*B_ Cheat_ notion*) exhibit a strong positive relationship with receivers' second-order beliefs (*B_ B_ Receivers_ actions*). Having seen that receivers' beliefs about senders' cheating notions (*B_ Cheat_ notion*) and receivers' second-order beliefs (*B_ B_ Receivers_ actions*) are closely related empirically, the question arises: do second-order beliefs contain predictive power for receivers' behavior beyond what is contained in cheating notion beliefs? We turn to this question next.

3.4.1 What constrains receivers behavior more?

In this section we investigate if moral expectations are a more fundamental determinant of guilt, by investigating whether there is any influence of mathematical expectations, after moral expectations are taken into account. Specifically we test:

Hypothesis 1: When estimating receivers' behavior, $r(s)$, as a function of both *B_ Cheat_ notion* and *B_ B_ receivers_ actions* simultaneously, *B_ Cheat_ notion* will be a significant predictor of $r(s)$ while *B_ B_ receivers_ actions* will have little explanatory power.

In Table 8 we present estimates of receivers' behavior as a function of both *B_ Cheat_ notion* and *B_ B_ receivers_ actions* for each $s = 1, \ldots, 10$, separately.[36] We find strong support for Hypothesis 1. Our estimates reveal that receivers' beliefs about senders' moral expectations (*B_ Cheat_ notion*) are almost always highly significant predictors of receivers' behavior while their beliefs about senders' mathematical expectations (*B_ B_ receivers_ actions*) almost never are.[37]

[35] For details about the empirical strategy to test for these correlations and the corresponding results, see part D of the appendix and Tables A14-A15.

[36] In each of these ten regressions we also control for a host of demographics to isolate the impact of the beliefs in question on behavior. Since we provide evidence in a later section that beliefs about others' moral expectations may be extrapolated from one's own moral expectations – a process that may not be available to individuals who have no moral expectations of their own – we insert a dummy for individuals who refrained from specifying *Cheat_ notion*. As lacking one's own moral expectations may affect *B_ Cheat_ notion* and *B_ B_ receivers_ actions* in different ways, we include interactions between both of these variables.

[37] Obviously, one may be worried that the lack of significance of *B_ B_ Receivers_ Actions* is due to col-

A second way to show that moral expectations are a more fundamental determinant of guilt than mathematical expectations is to test whether failing to live up to the sender's moral expectations is less likely than failing to live up to senders' mathematical expectations; in other words, we would like to see that senders' moral expectations constrain "cheating" – returning strictly less than the senders' relevant expectation – more than senders' mathematical expectations. One way to do that would be to provide some receivers with their sender's mathematical expectations and other receivers with their sender's moral expectations and show that moral expectations behave more like the type of threshold we would expect from guilt aversion models. We can perform this exercise using the data from our direct-response experiment and test the following:

Hypothesis 2: The event $[r < Cheat\_notion]$ in the cheating notion treatment (DR-CN) will be less likely than the event $[r < B\_receivers\_actions]$ in the first-order beliefs treatment (DR-FOB).[38]

In other words, we test whether violating senders' moral expectations is less likely than violating their mathematical expectations. Since we have a directional hypothesis, a one-sided test is appropriate. Because senders' decisions do not differ by treatment ($\chi^2(2) = 0.60, p = 0.74$), we compare receivers' average behavior across treatments. Specifically, we compare the proportion of observations in DR-CN in which receivers returned strictly less than the sender's moral expectations to the proportion of DR-FOB receivers returning less than their sender's mathematical expectations.

The results again support the notion that moral expectations are a more fundamental determinant of the guilt threshold. Only 32% of DR-CN receivers violated their sender's

linearity between *B_ B_ Receivers_ Actions* and *B_ Cheat_ notion*. However, notice that the standard errors associated with the coefficient on *B_ B_ Receivers_ actions* are of the same order of magnitude as those associated with *B_ Cheat_ notion* so that lack of significance of the former appears to be driven by the fact that the point estimates of the coefficients on *B_ B_ Receivers_ actions* are simply smaller. More formally, we also compute the variance inflation factors (VIFs) for both variables. For every s, the VIF was always less than 2 for both *B_ B_ Receivers_ actions* and *B_ Cheat_ notion*, whereas it typically takes a VIF greater than 10 to indicate collinearity may be an issue.

[38] In our direct-response experiment, half of the 112 participants played only the role of sender, submitting both a send amount and either their cheating notions (DR-CN) *or* their first-order beliefs about their co-player's actions (DR-FOB). The remaining 56 participants played only the role of receiver. Of these, 29 receivers participated in DR-CN and were informed of their co-player's cheating notions (*Cheat_ notion*) when deciding how much to return. The remaining 27 receivers participated in DR-FOB and were provided with their co-player's first-order beliefs (*B_ receivers_ actions*). Senders' information was transmitted to receivers in a credible way. Rather than eliciting receivers' beliefs, we assume that receivers' beliefs match the information they had at their disposal when making their decisions: in DR-CN (DR-FOB) we assume that each receiver's *B_ Cheat_ notion* (*B_ B_ receivers_ actions*) equals his or her sender's reported *Cheat_ notion* (*B_ receivers_ actions*).

moral expectations, while 52% of DR-FOB receivers violated their sender's mathematical expectations. This 20 percentage point increase in cheating represents 47% of the unrestricted sample mean (42%) and is marginally statistically significant ($p = 0.074$, one-sided difference-in-proportions test).

The third hypothesis we test is another way of asking which type of expectation acts more like a guilt threshold. We ask: conditional on satisfying the sender's expectation, do receivers *exactly* satisfy the expectation? If the expectation in question is a true guilt threshold, then returning more would not reduce guilt, theoretically, but would reduce the receiver's earnings, so that no receiver would willingly return strictly more. Under the plausible assumption that receivers' beliefs about their specific senders matched the information they had at their disposal when making their decision ($B\_B\_receivers\_actions$ equals the sender's reported $B\_receivers\_actions$ in DR-FOB; B_Cheat_notion equals the sender's reported $Cheat\_notion$ in DR-CN) we can test:

Hypothesis 3: Conditional on returning at least as much as their sender's mathematical or moral expectation, receivers' behavior will more closely mirror moral expectations than mathematical expectations: $r - B\_Cheat\_notion < r - B\_B\_receivers\_actions$.

To test this hypothesis, we restrict attention to those observations in our direct-response experiment where a receiver returned at least as much as their sender's moral (DR-CN) or mathematical (DR-FOB) expectation and ask: conditional on returning weakly more than required by the sender's expectation, how close do receivers come to returning *exactly* the relevant expectation. Put another way, how close does each notion come to resembling a threshold for avoiding guilt?

We start by pooling over all send amounts and restricting attention to observations where receivers returned at least as much as their sender's (moral or mathematical) expectation. For these observations, we compute the distance between each receiver's action and his or her sender's expectation: $r-B\_Cheat\_notion$ (DR-CN) or $r-B\_B\_Receivers\_actions$ (DR-FOB). We find that the average distance between a receiver's action and his or her sender's expectation conditional on not cheating is 0.50 ($s.e = 0.35$) in DR-CN, while in DR-FOB this distance is almost three times as large ($1.43, s.e. = 0.55$). Even with the few observations we have, we can reject the null hypothesis that these distances are equal across treatments ($p = 0.069$, one-tailed non-parametric permutation test). To provide corroborating graphical evidence, in Figure 6 we plot the raw data from the direct-response experiment, this time

restricting attention to observations where $s > 0$. We overlay the plot with a 45^o line. We use solid markers to indicate observations above the 45^o line, where receivers returned at least as much as their sender's expectation. We use hollow markers for observations below the line, where receivers cheated – returning less than the sender's expectation. From the figure it is apparent that receivers who know their sender's moral expectations are keen to exactly match them as observations in DR-CN not involving cheating typically lie quite close to the 45^o line. Receivers who live up to their sender's mathematical expectations, on the other hand, often exceed these expectations by a considerable amount. This is consistent with mathematical expectations being only a noisy measure of senders' disappointment threshold so that by returning strictly more receivers seek to avoid the risk of actually disappointing their sender. When receivers know sender's moral expectations, however, there is no risk of such accidental cheating so that receivers who intentionally choose to refrain from cheating need to return no more than the sender's moral expectation.

On the other hand, receivers who return strictly less than their sender's moral or mathematical expectations return substantially less: conditional on cheating, the distance between the sender's expectation and the receiver's action ranges from a minimum of 1.80 euros to a maximum of twelve euros with an average of 4.99 ($s.e. = 0.55$). These latter distances do not vary significantly across treatment ($p = 0.230$, one-tailed non-parametric permutation test). The discrete jump in return amounts conditional on cheating is also consistent with a story where senders' expectations serve as a guilt threshold. The discrete increase in earnings may be necessary to offset the discrete decrease in utility from triggering guilt.

Overall we do find that receivers who are given their sender's cheating notions and refrain from cheating tend to do so minimally: returning more than necessary to avoid cheating does not reduce guilt but does reduce own money earnings.

It would be reassuring to find this same pattern in the data from our main experiment where our data are more extensive but also more fraught with potential confounds. To provide such evidence, we split the sample between cheaters ($r < B\_Cheat\_notion$) and non-cheaters ($r \geq B\_Cheat\_notion$) and estimate the amount receivers return as a function of their beliefs about senders' cheating notions and our standard set of demographics. To formally account for selection into cheating or not cheating, we exploit our relatively large sample size and estimate Heckman models. Using the interpretation of $Cheat\_notion$

as a measure of how much participants care about morality, together with the evidence that own moral standards are predictive of cheating, our Heckman estimates use as their exclusion restrictions in the selection equations *Cheat_notion*. The results reported in Table 9 are broadly consistent with intentional cheating giving rise to guilt.[39] For those who choose to refrain from cheating, return amounts vary essentially one-to-one with their beliefs about senders' cheating notions, *B_Cheat_notion*, suggesting that receivers' beliefs about senders' cheating notions are acting as thresholds for non-cheaters. On the other hand, receivers who cheat their co-players are much less sensitive to these same beliefs. The estimated coefficients on *B_Cheat_notion* are consistently around half as large as for non-cheaters.

All together, the evidence from both our main experiment and the complementary evidence from our direct-response experiment support the idea that beliefs about senders' moral expectations are a more fundamental determinant of receivers' behavior than their beliefs about senders' mathematical expectations. The interpretation we favor is that violating senders' moral expectations is a primary determinant of guilt in trust-based exchanges.

Wrapping up, in this section we have shown that cheating notions may constitute a micro-foundation for models of guilt aversion. Providing a micro-foundation for guilt is important for two reasons. First of all, while the theory of guilt aversion is an elegant and self-contained theory, its equilibrium predictions depend crucially on mathematical expectations. Because the theory offers no guidance on which mathematical expectations are likely or plausible, equilibria often proliferate. Proliferation of equilibria limits the ability of the theory to provide clear predictions about behavior, limiting the scope for empirical applications. If the relevant expectations are moral and not *purely* equilibrium constructs, existing research can offer hints and hypotheses about which expectations, and hence which equilibria, are most likely. Furthermore, as moral expectations may be temporally persistent and culturally determined, understanding how such expectations vary across individuals and cultures may extend the empirical relevance, predictive ability and scope for impact of guilt aversion models.

A second reason micro-founding guilt may be of interest is practical. Eliciting even first-order beliefs often strains the limits of practicality as theoretically proper elicitation

[39] Ignoring selection issues and estimating simple OLS models of return amounts yields qualitatively similar results.

mechanisms typically require participants to be somewhat familiar with probability theory. Eliciting beliefs about elicited beliefs may require participants to have an even deeper understanding of probability theory – a requirement which is unlikely to be met outside of the usual college student subject pools. On the other hand, the feeling of being cheated is an emotional reaction many have experienced and consequently may be something a quite general population can easily comprehend, anticipate and reason about. Since receivers' (first-order) beliefs about senders' moral expectations (*B_Cheat_notion*) appear to be the most relevant driver of guilt, there may be little reason to incur the added complexity associated with eliciting second-order beliefs.

4 Discussion and interpretation

In this section we attempt to provide a more general view on the type of preferences that could explain receivers' cheating decisions.

We start by plotting (Figure 7) the fraction of receivers who intentionally cheat at each send amount after partialing out the effect of *B_Cheat_notion*, thus purging the data from the mechanical effect this has on the probability of cheating.[40] The declining propensity to cheat as receivers receive larger sums from senders is inconsistent with both purely selfish preferences, which imply that receivers would always cheat, and fixed-cost of cheating models, that would predict a *non-decreasing* relationship between amount sent and cheating propensity, since potential pecuniary gains from cheating increase in the amount sent.

Patterns in our data also appear to be inconsistent with literal interpretations of many influential social preferences models. For example, inequality averse individuals (Fehr and Schmidt, 1999) lose utility from unequal outcomes, while individuals with social welfare preferences (Charness and Rabin, 2002) place weight in their utility calculations on the outcome of the worst-off individual in their reference group as well as the total amount of money being distributed. Both of these models predict that receivers should never willingly put themselves behind in terms of final monetary payoffs. However, a large fraction of receivers in our study do exactly that. For example, 82% of receivers willingly put themselves further behind than necessary when sent 1 euro and 47% of the receivers put themselves

[40] For each $s \in \{1, \ldots, 10\}$ we estimate a linear probability model using our cheating dummy as the dependent variable and *B_Cheat_notion* as the lone independent variable. The estimated constants from these regressions are the cheating fractions we plot in Figure 7.

behind when sent 4 euros. The patterns suggest that receivers' behavior is unlikely to be explained by purely distributional concerns.

The estimation results in Table 9 and the cheating pattern in Figure 7 could be justified in (at least) two ways. The standard justification is positive or negative reciprocity: sending more is a nicer action and/or sending less is a meaner action, so reciprocity demands responding in kind with a nice action (not cheating) or a mean action (cheating). An alternative explanation comes directly from the definition of trust: trust entails *vulnerability*. At the same time, a widespread and intuitive moral standard is that, irrespective of what constitutes cheating, it is *more* wrong to cheat the more vulnerable. For example, cheating the elderly or the very young is commonly viewed as particularly reprehensible.[41] This is the point made by Gneezy (2005). In the context of the trust game, sending more plausibly makes senders more vulnerable. Consequently, it is reasonable to assume that the moral costs of cheating increase in amount sent. For ease of exposition, we will assume that vulnerability is a driving motive with the important caveat that we cannot rule out reciprocity as the driving motive: reciprocity and vulnerability provide observationally equivalent predictions.

In light of these patterns, a unified way to model both senders' and receivers' preferences that is consistent with our data is to augment standard pecuniary preferences with a moral cost function. Individuals incur disutility from immoral actions, either when they are the perpetrator or the victim of such actions. Receivers lose utility when they cheat because, e.g., they might suffer guilt. Senders lose utility when receivers cheat them by not living up to sender's cheating notions. This is consistent with our finding that the likelihood of ending up feeling cheated has a direct negative impact on the amount senders send. Beyond implying disutility from being cheated, our data do not say much about what senders' moral cost function might look like. On the receiver side, however, our data provide a bit more bite. For the rest of this section, therefore, we focus on receivers' preferences.

As a flexible specification for receivers' moral cost function m, we assume it has three arguments: the vulnerability of the sender as measured by the amount sent, s; a fixed cost of cheating term (K_j); and a term measuring the degree with which the receiver cheats, as

[41] Of course, the elderly or very young do not choose to be vulnerable, as our trust game senders do. However, there are some examples of injunctions against cheating even those who *choose* to be vulnerable in widely used moral codes such as those contained in the Bible. See, e.g., Deuteronomy 27:19 or Jeremiah 22:3 for injunctions against cheating foreigners.

defined by the distance between r and B_Cheat_notion (called $c_j(s)$ to ease notation).

A receiver's utility is then given by:

$$U_j(r, s, c_j, K_j) = u(f(s) - r)) - \mathbb{I}[r < c_j(s)] \times m(s, K_j, dist(r, c_j(s))) \quad (1)$$

In (1), $f(s)$ denotes how much the receiver receives when the sender sends s and $\mathbb{I}[r < c_j(s)]$ is an indicator function taking the value of 1 whenever the receiver intentionally cheats. We assume that the moral cost function $m(s, K_j, dist(r, c_j(s)))$ is increasing in s, the vulnerability of the sender. We also assume that the fixed cost of cheating, $K_j \geq 0$, is a random draw from a common non-degenerate distribution function, $F(K)$. Finally, we assume that m is increasing and convex in its last argument, $dist(r, c_j(s))$, so that higher degrees of cheating are increasingly morally costly.

To be more concrete, a simple utility specification satisfying these assumptions is given by:

$$U_j(r, s, c_j, K_j) = u(f(s) - r)) - \alpha_j \mathbb{I}[r < c_j(s)] \times \{K_j + v(s) + \gamma_j (c_j(s) - r)^2\} \quad (2)$$

In equation 2, we assume that $u(f(s) - r)$, the receiver's standard pecuniary utility, is increasing and concave. The rest of the utility function captures the moral cost of cheating. The parameter α_j captures how much receiver j cares about morality. The parameter γ_j captures how much the receiver cares about degrees of cheating. A sender's vulnerability or niceness is captured by $v(s)$ which we assume is increasing.

There are three points to notice about this utility specification. First of all, setting $\alpha_j = 0$ reduces receivers' utility to standard (amoral) preferences; Secondly, notice that whenever $\alpha_j > 0$, setting $\gamma_j = v(s) \equiv 0$ implies receivers have simple fixed-cost-of-cheating preferences. Finally, if we assume that receivers' beliefs about senders' cheating notions – $c_j(s)$ in our model – are informative of receivers' second-order beliefs, as we have seen in the data, then one way to think about this model is as a reduced-form guilt aversion model.

The specification for receiver utility given in equation 2 can explain: a) why the decision to cheat depends on others' expected cheating notions; b) why cheating depends on the intensity of moral preferences as proxied for by receivers' own cheating notions; and c) why the probability of cheating decreases in amounts sent as shown in Figure 7. This latter feature would be implied, for instance, whenever there are sufficiently many receivers with

$\alpha_j > 0$ and when $v(s)$ is sufficiently steep in s. Intuitively, as $v(s)$ becomes steeper, cheating more vulnerable senders requires a larger offsetting pecuniary utility gain.

This simple preference specification can also account for another feature of the data: conditional on cheating, receivers on average do not go so far as to return nothing. Instead, they send *something* back. In our model, the amount returned by cheaters should depend positively on expected cheating notions, but – and this is the key prediction – it should *not* move one-to-one with the expected notion of cheating. On the other hand, conditional on not cheating, receivers should return the minimum amount consistent with satisfying the sender's notion. Non-cheaters' return amounts should therefore move one-to-one with the expected cheating notion. Only the latter prediction is shared by both our model and the fixed cost of cheating model. As we have seen, both predictions find support in our data.

5 Concluding Remarks

Many real life exchanges require the "trustor" to decide whether and how much to trust a "trustee" who makes no promise on how he will behave in response to the trust received. This paper investigates what individuals' personal, subjective notions of what constitutes cheating – their moral expectations – can tell us about behavior in such situations. Our study takes place in the context of a trust game where we elicit participants' definitions of being cheated and a wide array of related beliefs.

In this context, our data suggest several patterns. First of all, participants have personal cheating definitions when playing the trust game. We find that these moral expectations and beliefs about others' moral expectations are quite heterogeneous but roughly bimodal, clustering around an equal-split rule and a positive return on investment rule. We provide evidence that (first-order) beliefs about others' cheating notions may provide a micro-foundation for guilt, which potentially extends the scope for empirical applications of guilt aversion theory. Finally, we document evidence consistent with cheating notions being culturally transmitted, and hence stable, which is important since we also find that stability in one's own moral expectations may translate into stability in beliefs about others' cheating notions through false consensus. All together, our results suggest that studying cheating notions and related beliefs can help us understand and predict behavior in trust-based exchange.

An interesting question which we cannot address with our current data is how *knowing*

that there are multiple notions of cheating affects sender and receiver behavior, either in the one-shot context here or when, more realistically, individuals interact repeatedly. One may wonder whether individuals adapt their own cheating notions to be more in line with the average population cheating notions causing an eventual convergence to one normative cheating standard; or, rather, whether those with high cheating notions cease to interact with the general population because they feel cheated more often in their interactions. We leave these and related questions for future research.

References

[1] Akerlof, George A. and William T. Dickens (1982), "The Economic Consequences of Cognitive Dissonance," *The American Economic Review*, 72, 307-319.

[2] Balafoutas, L., A. Beck, R. Kerschbamer, and M. Sutter (*forthcoming*), "What drives taxi drivers? A field experiment on fraud in a market for credence goods." *Review of Economic Studies*.

[3] Battigalli, Pierpaolo and Martin Dufwenberg (2007). "Guilt in Games," *American Economic Review*, 97, pp. 170-176.

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

[5] Bohnet, Iris and Richard Zeckhauser (2004), "Trust, Risk and Betrayal." *Journal of Economic Behavior and Organization*, 55(4), pp. 467-484.

[6] Bisin, Alberto and Thierry Verdier (2010), "The Economics of Cultural Transmission and Socialization." In Jess Benhabib, Alberto Bisin and Matthew O. Jackson editors: Handbook of Social Economics, Vol. 1A, The Netherlands: North-Holland, 2011, pp. 339-416.

[7] Brandts, Jordi and Gary Charness (2000), "Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics*, 2: 227-238.

[8] Brandts, Jordi and Gary Charness (2011), "The strategy method vs. the direct-response methodL a first survey of experimental comparisons," *Experimental Economics*, 14: 375-398.

[9] Butler, Jeffrey V., Paola Giuliano and Luigi Guiso (*forthcoming*), "Trust, Values and False Consensus." *International Economic Review*

[10] Castillo, Marco, Ragan Petrie, Torero Ragan, Maximo A. Torero and Lise Vesterlund (2012), "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination." NBER Working Paper No. w18093

[11] Charness, Gary and Martin Dufwenberg (2006), "Promises and Partnership." *Econometrica*, 74(6), pp. 1579-1601.

[12] Charness, Gary, Uri Gneezy and Michael A. Kuhn (2012), "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior & Organization*, 81: 1-8.

[13] Charness, Gary and Matthew Rabin (2002). "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117(3), pp. 817-869.

[14] Charness, Gary and Arthur Schram (2013), "Social and Moral Norms in Allocation Choices in the Laboratory," *Economics Working Paper Series*, Department of Economics, UC Santa Barbara.

[15] Chater, Nick, Steffen Huck and Roman Inderst (2010), "Consumer Decision-Making in Retail Investment Services: A Behavioral Economics Perspective", Report to the European Commission/SANCO.

[16] Cox, James C. (2004), "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46, 260-281

[17] Dufwenberg, Martin and Gneezy, Uri (2000). "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior*, 30, pp. 163-182.

[18] Eagly, A.H. and M. Crowley (1986). "Gender and Helping Behavior: A Meta-Analytic Review of the Social Psychological Literature," *Psychological Bulletin*, 100, pp. 283-308.

[19] Eckel, Catherine C. and Philip J. Grossman (1998). "Are Women Less Selfish Than Men?: Evidence from Dictator Experiments," *Economic Journal*, 108, pp. 726-35.

[20] Ermisch, John and Diego Gambetta (2006). "People's trust: the design of a survey-based experiment," *ISER Working Paper Series 2006-34*, Institute for Social and Economic Research.

[21] Fehr, Ernst (2009), "On the Economics and Biology of Trust", *Journal of the European Economic Association*, 7, pp. 235-266.

[22] Fehr, Ernst and Urs Fischbacher (2004), "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25, pp. 63-87.

[23] Fehr, Ernst and K.M. Schmidt (1999). "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114 (3), pp. 817-868.

[24] Geanakoplos, John, David Pearce and Ennio Stacchetti (1989), "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, pp. 60-79.

[25] Glaeser, Edward, David Laibson, Jose A. Scheinkman and Christine L. Soutter (2000), "Measuring Trust," *Quarterly Journal of Economics* 115(3), 811-846.

[26] Gneezy, Uri (2005), "Deception: The role of consequences," *American Economic Review*, March 2005, 384-394.

[27] Hung, Angela A., Clancy Noreen, Jeff Dominitiz, Eric Talley, Calude Berrebi and Farrukh Suvankulov (2008), "Investor and Industry Perspectives on Investment Advisers and Broker-Dealers", Technical Report, Rand Institute for Civil Justice.

[28] Inderst, Roman and Marco Ottaviani (2012), "Financial Advice," *Journal of Economic Literature*, 50(2): 494-512.

[29] Iriberri, Nagore and Pedro Rey-Biel (2011), "The role of role uncertainty in modified dictator games," *Experimental Economics*, 14(2): 160-180.

[30] Krupka, Erin L. and Roberto A. Weber (2013), "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, 11(3): 495-524.

[31] Lundquist, Tobias, Tore Ellingsen, Erik Gribbe and Magnus Johannesson (2009), "The aversion to lying," *Journal of Economic Behavior & Organization*, 70: 81-92.

[32] Rabin, Matthew (1993), "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5), pp. 1281-1302.

[33] Reuben, Ernesto and Arno Riedl (2013), "Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations," *Games and Economic Behavior*, 77(1): 122-137.

[34] Reiss, Michelle C. and Kaushik Mitra (1998). "The Effects of Individual Difference Factors on the Acceptability of Ethical and Unethical Workplace Behaviors, *Journal of Business Ethics*, 17(14), pp. 1581-93.

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

[36] Rousseau, Denise and Sim B. Sitkin and Ronald S. Burt and Colin Camerer (1998), "Introduction to Special Topic Forum: Not So Different After All: A Cross-Discipline View of Trust," *The Academy of Management Review*, 23(3), pp. 393-404.

[37] Rustichini, Aldo and Marie-Claire Villeval (2012), "Moral Hypocrisy, Power and Social Preferences," *GATE Working Paper No. 1216.*

[38] Schlag, Karl and Joel J. van der Weele (2013), "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality," *Theoretical Economics Letters*, 3(1), 38-42.

[39] Sapienza, Paola, Anna Toldra and Luigi Zingales (2007), "Understanding Trust," NBER WP 13387

[40] Selten, Reinhard (1967), "Die Strategiemethode zur Erforschung des eingeschrï¿œnkt rationalen Verhaltens im Rahmen eines Oligopolexperiments," in H. Sauermann (Ed.), Beitrï¿œge zur experimentellen Wirtschaftsforschung, Tï¿œbingen, Mohr: 136 - 168.

[41] Smith, Vernon L. (1976), "Experimental economics: Induced value theory." *American Economic Review*, 66(2): 274-279.

Table 1
Experimental design

	Number of sessions	Explicit cheating notion question opt-out	Investment fee	Max belief pay	Obs
Initial study	4	No	0.50 euro	5 euro	122
Additional sessions	4	Yes	0.50 euro (2 sessions) 0.00 euro (2 sessions)	20 euro	306

Table 2
Descriptive statistics

	Mean	Std Dev	Min	Max	N
Male	0.46	0.499	0	1	420
Age	23.73	4.171	18	58	420
Math score	7.66	1.251	3	10	402
Inc<30K	0.29	0.455	0	1	391
30≤Inc<45	0.24	0.426	0	1	391
45≤Inc<70	0.25	0.431	0	1	391
70≤Inc<120	0.16	0.366	0	1	391
Inc≥120K	0.07	0.249	0	1	391
Risk aversion	5.71	2.193	1	10	417
Send decision (binary)	0.81	0.392	0	1	428
Send amount	4.31	3.232	0	10	428
Average return proportion	1.28	0.697	0	4.02	427
B_return_proportion	1.27	0.637	0	4.02	425
Competitive values emphasis	0.62	0.196	0	1	410
Good values emphasis	0.76	0.149	0.17	1	404
Pr(NotCheated)	0.42	0.232	0	1	427
Average proportion of non-cheaters	0.49	0.376	0	1	428

Table 3
Variable Description

Variable Name	Question
Cheat_notion	This is shorthand for "Cheating notion" and is a participant's answer to the question "If you are assigned the role of A [sender] what is the minimum amount you would need to receive back from player B [receiver] in order to not feel cheated? …If you were to send €[s] and B were to therefore receive €[f(s)], you would need back how many euros?"
B_Cheat_notion	This is shorthand for "Beliefs about Cheating notions". They are the answers to the set of questions: "What is the minimum amount (on average) that A's will need back from B's in order to not feel cheated? If A sends €[s] and B therefore receives €[f(s)], to not feel cheated A will need back from B at least: €__.__"
B_Receivers_actions	This is shorthand for "My Belief about Receivers' Actions" and is the answer to the set of questions: "How much, on average, will B's return to A's? If A sends €[s] and B therefore receives €[f(s)], B's will return on average: €__.__"
B_B_Receivers_actions	This is shorthand for "Beliefs about Others' Beliefs about Receivers' Actions." These are the answers to the set of questions "How much money (on average) do other participants in the role of A believe will be returned to them by B's? If A sends €[s] and B therefore receives €[f(s)], how much money does A believe B will return? €__.__"
B_NotCheated	This is shorthand for "Beliefs about the Probability of Not Feeling Cheated" These are participants' answers to the set of questions: "What percent of participants in the role of B will return enough money to you (if you are assigned the role of A) so that you will not feel cheated? …If you send €[s] and B therefore receives €[f(s)], what percent of B's will return enough so that you will not feel cheated? ."

Note: Each variable listed in this table is actually a set of ten variables, one for each possible send amount s = 1, …, 10. However, as in the table, we will typically suppress the dependence on s for ease of exposition.

Table 4
Proportion of participants in sessions who opt-out of reporting a cheating notion in sessions with explicit opt-out opportunities

	Send Amount										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Obs
	Proportion who selected "this has nothing to do with cheating"										
Mean	0.20	0.18	0.17	0.15	0.13	0.13	0.13	0.13	0.14	0.13	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
	Proportion who did not report a cheating notion for any reason										
Mean	0.23	0.21	0.21	0.17	0.15	0.15	0.15	0.16	0.17	0.17	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	

Notes: [1] In sessions with an explicit "opt-out" possibility participants could refrain from specifying an explicit personal cheating notion and instead respond either "I don't know" or "this has nothing to do with cheating." [2] The top row of Table 4 presents the proportion of participants who chose "this has nothing to do with cheating," while the lower row presents the proportion of participants who chose either of these two "opt-outs" or left the question entirely blank.

Table 5
Determinants of cheating notions

Dependent variable = *Cheat_notion*										
Cooperative values	Competitive values	€ sent	(€ sent)^2	Male	Age	Math score	Risk aversion	Cons	Obs	Individuals
-2.55**	1.63**	1.07***	-0.02***	-0.47	0.00	-0.02	-0.11	3.55***	3496	354
(1.09)	(0.64)	(0.07)	(0.01)	(0.43)	(0.03)	(0.11)	(0.08)	(1.31)		

Notes: [1] Estimates are from an individual-level random effects regression model. [2] Variables present in the regression, but omitted for readability: full set of income dummies; dummy for sessions with no investment fee; dummy for sessions comprising the initial study. None of these variables had significant coefficients. [3] Robust standard errors, clustered by session, appear in parentheses.

Table 6
Receivers' decision to intentionally cheat, by send amount

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	-0.08**	-0.13***	-0.08*	-0.07***	-0.04	-0.04*	-0.04	-0.05**	-0.03*	-0.04*
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
B_Cheat_notion	0.22***	0.21***	0.23***	0.22***	0.17***	0.15***	0.16***	0.17***	0.12***	0.14***
	(0.04)	(0.06)	(0.05)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
Male	0.07	-0.02	0.18	0.06	0.03	-0.05	-0.16	-0.07	0.01	-0.06
	(0.07)	(0.16)	(0.12)	(0.14)	(0.20)	(0.15)	(0.14)	(0.13)	(0.13)	(0.18)
Age	-0.03	-0.04***	-0.03***	-0.02**	-0.02**	-0.03**	-0.01	0.01	-0.01	-0.01
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Math score	-0.04	-0.03	0.04	0.05	-0.04	-0.03	-0.06***	-0.06	-0.08**	-0.03
	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)	(0.09)	(0.02)	(0.06)	(0.04)	(0.04)
Risk aversion	-0.00	0.03	0.01	-0.00	-0.02	-0.02	0.00	-0.01	-0.02	-0.07***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)
30≤ Inc <45	-0.02	0.18	0.24**	0.22	0.09	0.25*	0.50*	0.06	0.10	0.16
	(0.19)	(0.16)	(0.12)	(0.21)	(0.23)	(0.15)	(0.26)	(0.19)	(0.12)	(0.21)
45≤ Inc<70	0.12	0.01	0.06	0.10	0.29*	0.23***	0.43	0.24**	0.12	0.15
	(0.16)	(0.08)	(0.13)	(0.17)	(0.17)	(0.07)	(0.28)	(0.12)	(0.14)	(0.20)
70≤ Inc <120	0.17	0.17	0.07	-0.05	0.33*	0.41*	0.58**	0.70***	0.04	0.16
	(0.33)	(0.18)	(0.21)	(0.18)	(0.19)	(0.22)	(0.24)	(0.16)	(0.20)	(0.33)
Inc ≥120	0.00	-0.21	-0.07	0.01	-0.51	0.02	-0.21	-0.44	-0.04	-0.56*
	(0.35)	(0.28)	(0.21)	(0.20)	(0.32)	(0.28)	(0.40)	(0.31)	(0.29)	(0.33)
Constant	0.45	0.85	-0.69	-1.02*	-0.07	-0.10	-0.76*	-0.97	-0.05	-0.42
	(0.76)	(0.52)	(0.52)	(0.60)	(0.41)	(0.79)	(0.41)	(0.81)	(0.70)	(0.63)
Obs	369	366	366	369	371	370	371	369	366	366

Notes: [1] Each column presents estimates from a Probit model. Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated. [2] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%. [3] Math score is individual's self-reported score on required math exams taken during the final year of high school in Italy. [4] Income variables refer to self-reported annual family income from all sources, in thousands of euros, net of taxes. The excluded category is "below 30 thousand euros annually". [5] Observations vary over columns because not all participants reported a cheating notion for every send amount. This is discussed in the text. Additionally, we do not have demographics for all participants.

Table 7
Senders' decisions, Heckman estimates

	Main equation (1)	Selection equation (2)
Pr(NotCheated)	2.76**	0.57
	(1.38)	(0.65)
B_return_proportion	1.34***	0.28**
	(0.45)	(0.12)
Pr(NotCheated)x B_return_proportion	-1.57*	-0.07
	(0.85)	(0.46)
Low fee (dummy)	--	0.68***
		(0.09)
Age	0.11***	0.00
	(0.03)	(0.02)
Male	0.36	0.35**
	(0.32)	(0.14)
Math score	-0.00	0.12***
	(0.09)	(0.04)
Risk aversion	-0.14***	0.04
	(0.05)	(0.03)
Altruism	0.03	0.04
	(0.12)	(0.04)
30≤ Income <45	-0.29	0.13
	(0.42)	(0.25)
45≤ Income <70	-0.22	-0.04
	(0.59)	(0.23)
70≤ Income <120	-0.62**	-0.08
	(0.29)	(0.13)
Income ≥120	-0.63	0.74*
	(0.70)	(0.40)
Constant	1.45	-1.62***
	(2.16)	(0.61)
Obs	350	350
Mills Ratio	0.33	
	(0.18)	

Notes: [1] Robust standard errors, clustered by session, appear in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] For the Heckman model (cols 1-2): the dependent variable in the main equation is *how much* the sender sends; the dependent variable in the selection equation takes the value of 1 if the sender sends a positive amount and 0 otherwise; [4] The exclusion restriction for the selection equation consists of a dummy for "Low fee" sessions, a dummy taking the value of one if the observation came from a session where senders were charged nothing to send a positive amount, and 0 if the observation came from a session where senders were charged € 0.50 to send a positive amount [5] "Pr(NotCheated)" is our measure of probability about not being cheated, described in the text; "B_return_proportion" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send

amounts; "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse); "*Altruism*" is how much emphasis participants' parents placed on the value "help others" during their upbringing. [6] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table 8
Predicting receiver behavior: second-order beliefs or cheating notion beliefs?

Dependent variable = return amount conditional on send amount in column heading										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.25*	0.33***	0.31***	0.18***	0.32***	0.21**	0.19*	0.25***	0.17**	0.24***
	(0.11)	(0.09)	(0.07)	(0.04)	(0.06)	(0.07)	(0.08)	(0.05)	(0.06)	(0.06)
B_B_receivers_actions	0.31**	0.07	0.11	0.15*	0.11	0.13	0.16	0.11	0.21**	0.08
	(0.12)	(0.12)	(0.07)	(0.07)	(0.12)	(0.10)	(0.11)	(0.07)	(0.07)	(0.10)
No Personal Cheat Notion (NPCN)	-0.17	-0.76	-1.12**	-2.01***	-1.62	-1.20	-3.23**	-1.45	-4.15***	-3.83**
	(0.41)	(0.43)	(0.39)	(0.54)	(0.98)	(1.79)	(1.13)	(1.18)	(0.97)	(1.59)
NPCN X B_Cheat_notion	-0.03	-0.25	-0.23	-0.09	-0.33**	0.09	0.00	0.24	0.41**	0.08
	(0.19)	(0.19)	(0.15)	(0.10)	(0.12)	(0.16)	(0.35)	(0.13)	(0.12)	(0.13)
NPCN X B_B_receivers_actions	0.16	0.41	0.49**	0.35**	0.66**	0.11	0.37	0.01	0.06	0.33
	(0.27)	(0.29)	(0.14)	(0.10)	(0.24)	(0.41)	(0.38)	(0.17)	(0.21)	(0.23)
Constant	-0.02	0.70	2.82**	4.23***	2.59*	2.43*	4.01***	4.80**	3.78**	4.16*
	(1.51)	(0.90)	(0.90)	(0.82)	(1.30)	(1.05)	(0.95)	(1.62)	(1.56)	(1.78)
Demographics?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	375	375	375	375	375	375	375	375	375	375
R-squared	0.20	0.15	0.16	0.13	0.19	0.12	0.14	0.15	0.17	0.14

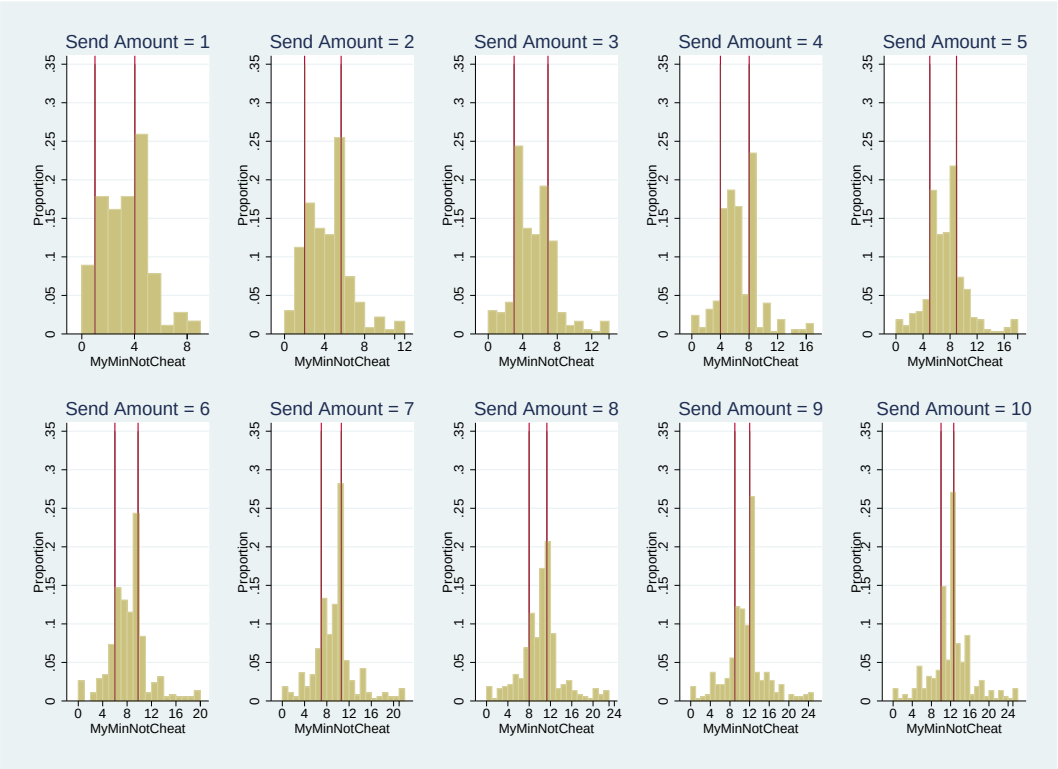
Notes: [1] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents an OLS estimate using the dependent variable r(s), where s is specified in the column heading. [3] The reported independent variables in column *i* are: "B_Cheat_notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, *i*=1,…,10; "B_B_receivers_actions" is each participant's belief about the average amount of money the sender believes the receiver will send back when the sender sends *i* euros, *i*=1,…,10. [4] Each estimate includes demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion.

Table 9
Sensitivity of amounts returned to beliefs about senders' cheating notions by decision to cheat, Heckman models

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (r ≥ B_Cheat_notion)										
B_Cheat_notion	1.17***	1.02***	0.97***	1.19***	1.07***	0.95***	0.88***	0.89***	1.09***	1.02***
	(0.17)	(0.13)	(0.14)	(0.25)	(0.15)	(0.11)	(0.16)	(0.13)	(0.22)	(0.13)
Constant	3.83**	4.98**	3.54**	4.68*	5.06**	4.88***	5.96***	4.97**	8.15**	9.26***
	(1.95)	(2.25)	(1.73)	(2.78)	(2.39)	(1.85)	(2.10)	(2.00)	(4.06)	(3.02)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
	Wald test: *B_Cheat_notions* coefficient = 1 (p-value)									
	0.32	0.86	0.84	0.43	0.63	0.66	0.43	0.39	0.67	0.84
Conditional on cheating (r < B_Cheat_notion)										
B_Cheat_notion	0.42***	0.37***	0.57***	0.38***	0.44***	0.43***	0.49***	0.58***	0.63***	0.53***
	(0.07)	(0.06)	(0.10)	(0.10)	(0.10)	(0.09)	(0.13)	(0.12)	(0.14)	(0.12)
Constant	-0.16	0.93	0.18	0.95	1.68	0.03	1.09	0.09	-0.38	-0.01
	(0.92)	(1.02)	(1.51)	(1.92)	(1.91)	(2.01)	(2.87)	(3.02)	(3.36)	(3.31)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
	Wald test: *B_Cheat_notions* coefficient = 0.5 (p-value)									
	0.23	0.04	0.47	0.24	0.52	0.43	0.95	0.52	0.34	0.79

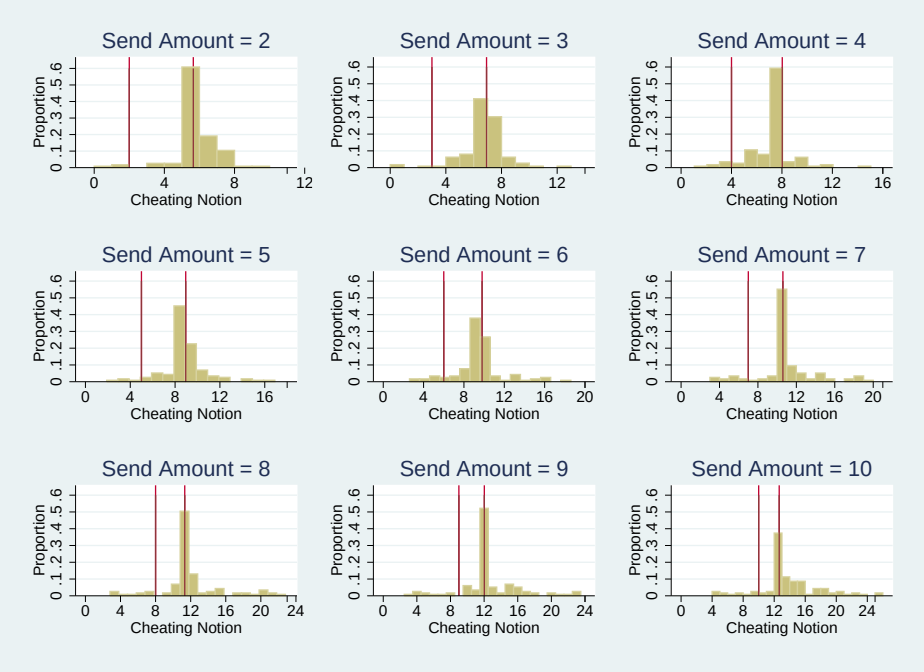
Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a Heckman model estimate using as its exclusion restriction participants' own cheating notions. [3] The dependent variable in column i is the amount a participant will send back if the sender sends i euros, i=1,…,10. [4] The reported independent variables in column *i* are: "B_Cheat_notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, i=1,…,10. [5] Each estimate includes our standard set of demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion.

Figure 1
Own Cheating Notions (Cheat_notion)



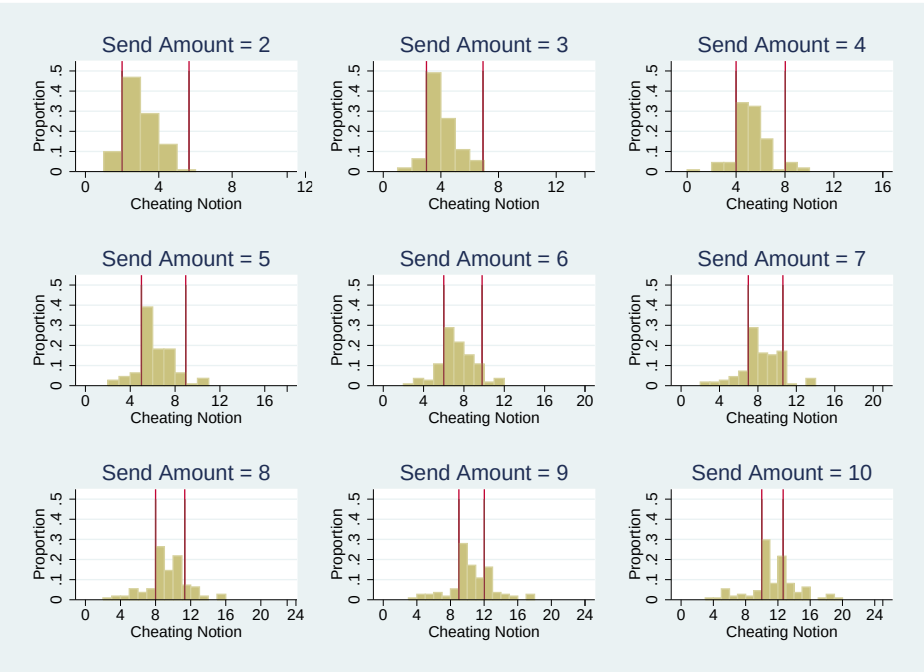
Notes: [1] The figure reports histograms of participants' personal cheating notions for each send amount s=1,….,10. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a *weakly positive return on investment* cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an *equal split* cheating definition.

Figure 2A: Within-individual Consistency of Cheating Notions across Send Amounts, equal split



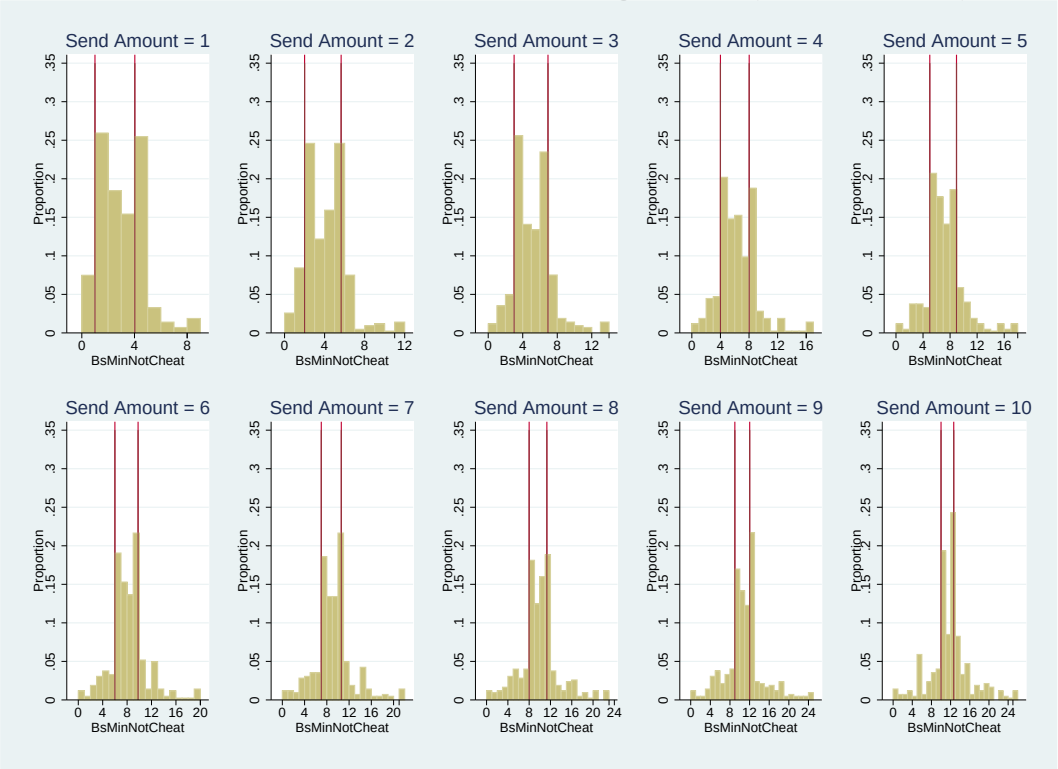
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 2B: Individual-level Consistency of Cheating Notions across Send Amounts, strictly positive return on investment



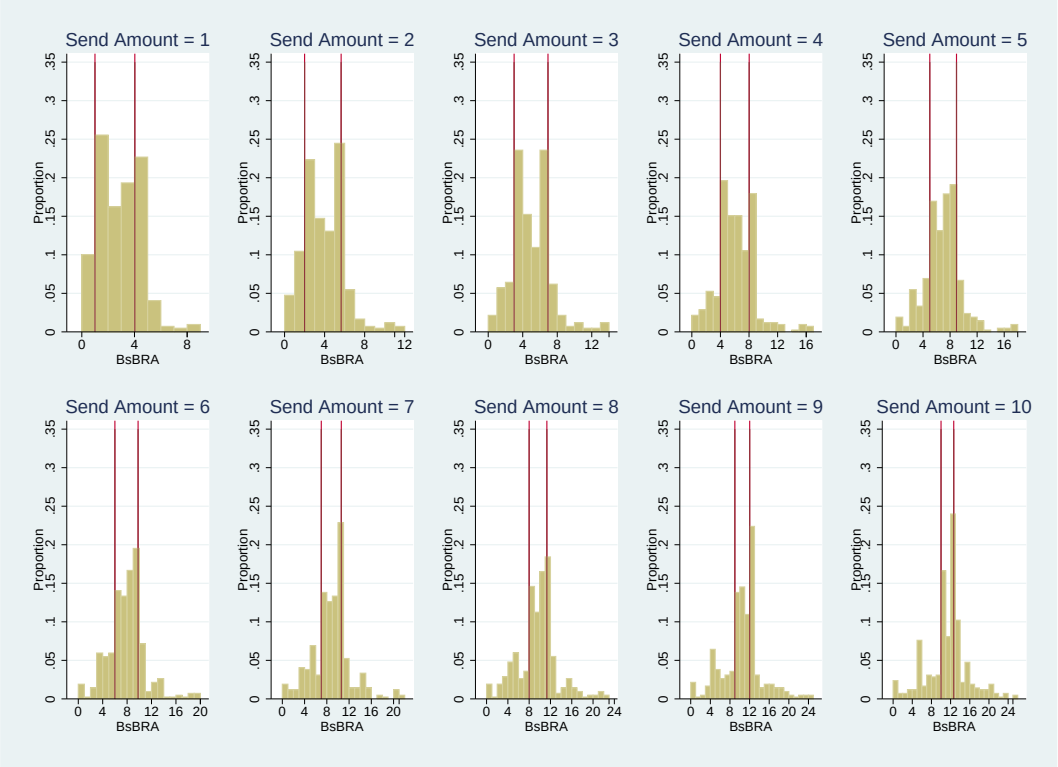
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 3
Participants' Beliefs about Others' Cheating Notions (B_Cheat_notion)



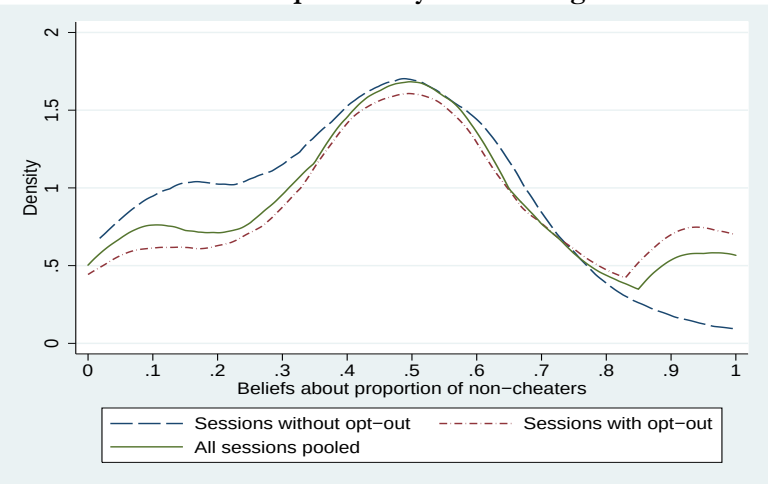
Notes: [1] The figure reports histograms of participants' beliefs about other participants' cheating notions (B_Cheat_notion). [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 4
Second-order beliefs (B_B_receivers_actions)



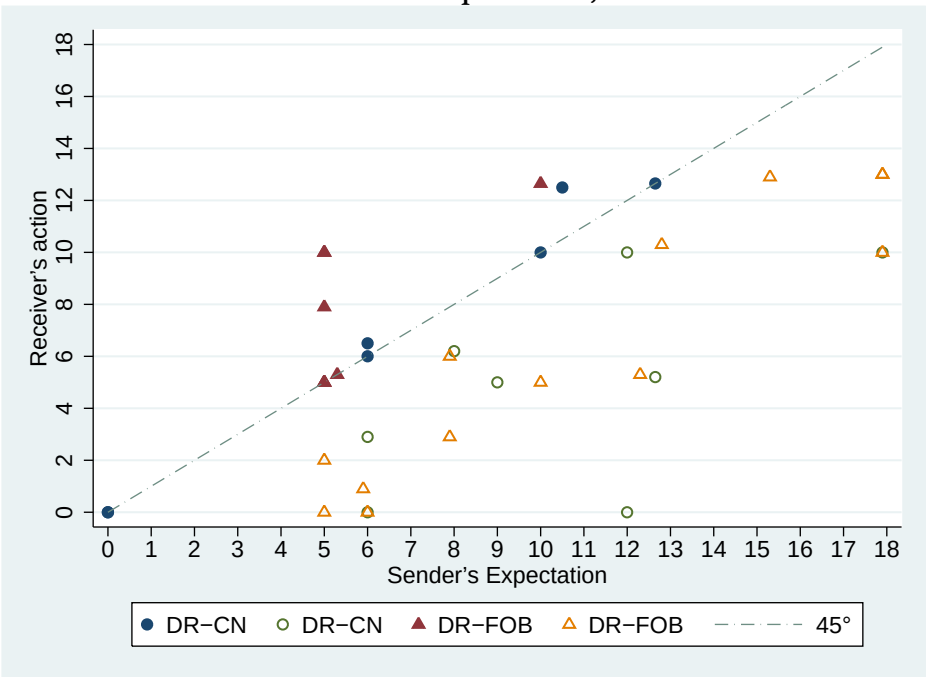
Notes: [1] The figure plots participants' beliefs about senders' beliefs about receivers' actions (B_B_receivers_actions). [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 5
Beliefs about the probability of not being cheated



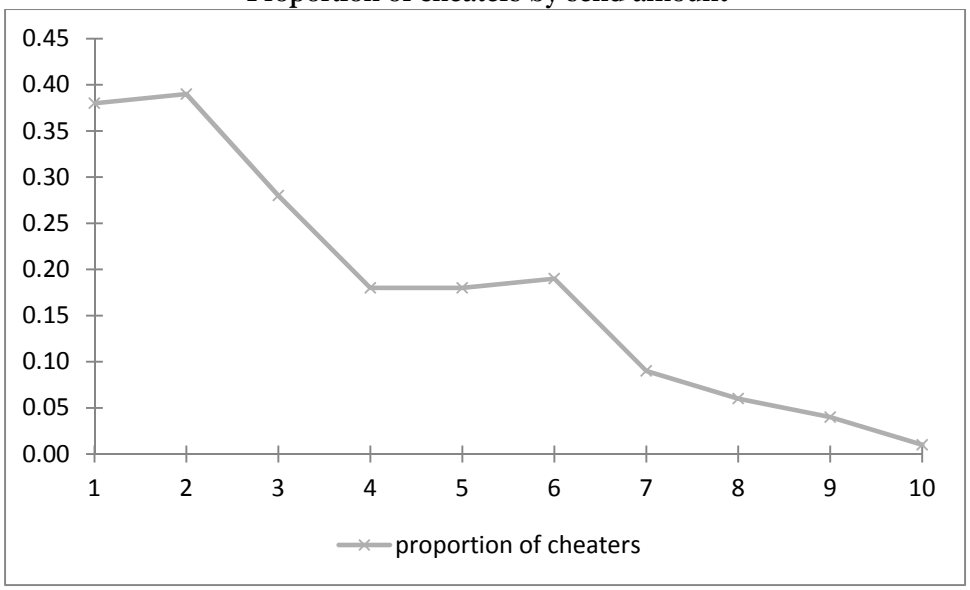
Notes: Observations in the sessions with opt-out (short-dash line) are restricted to individuals who have a cheating notion for every possible amount a sender could send. This is to ensure our summary measure of beliefs about the probability of being cheated is well-defined. Thus the density plot for the additional sessions is based on 207 (out of 306) observations.

Figure 6
Receiver's actions vs. sender's expectations, DR-CN and DR-SOB



Notes: [1] The figure restricts attention to observations in the direct-response experiment where s > 0 and plots each receiver's action against his or her sender's moral (DR-CN) or mathematical (DR_FOB) expectation. [2] Solid markers correspond to observations where the receiver did not cheat – i.e., returned at least as much as their sender's expectation – while hollow markers correspond to observations where the receiver cheated. [3] The dashed line is a 45-degree line along which a receiver's action exactly matches his or her sender's expectation.

Figure 7
Proportion of cheaters by send amount



Notes: The figure reports the proportion of cheaters (y-axis), after partialling out the effect of expectations of others' cheating notions, for each possible send amount (x-axis).

Not for publication

Appendix I: Robustness Checks

A Additional Robustness check treatments

In addition to our main experiment described in Appendix II, two further treatments were conducted for robustness. First of all, to check whether there is something peculiar about the on-line environment driving our results or whether paying only 10 percent of participants provides incentives that are too weak, we ran two sessions in the laboratory where 100 percent of participants were paid. As a second robustness exercise, we conducted sessions in which our direct cheating notion question was omitted and replaced with a series of questions asking participants how they would feel about various possible outcomes in the trust game from the point of view of the sender. The purpose of this latter treatment is to address the concern that our direct cheating notion question might prime participants to associate cheating with the trust game.

A.1 In-lab sessions

In total, 36 individuals took part in two sessions conducted in the experimental laboratory at the Einaudi Institute for Economics and Finance in Rome, Italy. Participants were recruited from the same subject pool as were the on-line sessions. There was no overlap in actual participants—i.e., no participant took part in both an on-line session and an in-lab session. All in-lab participants were paid based on their choices in the experiment and the accuracy of the their reported beliefs.

Apart from taking place in the laboratory, the design of this treatment and the materials used were exactly the same as the on-line treatments. Participants simply completed the on-line experiment in the laboratory. All sessions of the in-lab experiment allowed participants to opt out of specifying a cheating notion by selecting one of two responses: "I don't know" or "this has nothing to do with cheating." Neither session featured a fee to send a positive amount.

In Table A1 we report summary statistics for both the in-lab and most comparable on-line sessions. Receivers' behavior does not change much across these two environments: average return proportions and the propensity to intentionally cheat are all quite similar. Beliefs about these return proportions (*B_return_proportion*) and the likelihood of being cheated are also quite similar across the two environments. On the other hand, in-lab senders were slightly more likely to send a positive amount than their on-line counterparts, raising the average amount sent by in-lab senders. However, conditional on sending a positive amount average send amounts were again quite similar: 5.36 in on-line low fee sessions; 5.43 in the laboratory; with standard errors 0.25 and 0.44, respectively.

In terms of cheating notions (*Cheat_notion*), the picture is also quite similar in the lab and on-line experiments: the vast majority of participants have a cheating notion for all possible send amounts (Table A2); the vast majority have a cheating notion *at least as demanding as* the weakly positive return on investment (TableA3). Considering the proportion of participants whose cheating notions are consistent with various definitions

(Table A4), we again see that the weakly positive return on investment describes a small minority of participants, while a similar but relaxed notion, a strictly positive return on investment, describes a substantial minority of participants for most send amounts, as does an equal split rule: over all send amounts, these two rules each account for about 27%-29% of participants' reported cheating notion. We also, again, find that literal inequality aversion fits very few participants' definitions of cheating. We find the same patterns when considering beliefs about others' cheating notions (Table A5), which is also consistent with our on-line findings.

Considering next the relationship between second-order beliefs (*B_ B_ Cheat_ notion*) and cheating notions and related beliefs, the in-lab environment delivers similar patterns as those found in the on-line environment. Own cheating notions are again highly predictive of beliefs about how much receivers will return (*B_ Receivers_ actions*) (Table A6). In-lab beliefs about others' cheating notions (*B_ Cheat_ notion*) are highly predictive of in-lab second-order beliefs (*B_ B_ Receivers_ actions*) (Table A7). As in the on-line data, *Cheat_ notion* is typically negatively related to intentional cheating while *B_ Cheat_ notion* is usually positively related to intentional cheating (Table A8).

In Table A9, we replicate the pattern suggesting that beliefs about others' cheating notions (*B_ Cheat_ notion*) function as thresholds for those who refrain from cheating. Because we have many fewer observations here, to show this we take a more straightforward approach and do not model selection explicitly. Instead, we simply split the data into those who refrain from intentional cheating (top panel) and those who intentionally cheat (bottom panel) and run simple univariate OLS regressions of return amounts on beliefs about others' cheating notions. We find that, just as in the main data, for those who refrain from intentionally cheat, return amounts vary essentially one-to-one with *B_ Cheat_ notion* for most send amounts. For those who intentionally cheat, return amounts are consistently much less sensitive to *B_ Cheat_ notion* which is, again, consistent what we find in the on-line data.

Considering the sender's side of the exchange, next we consider how send amounts vary with cheating and monetary return beliefs (Table A10). Because we have few observations and lack the exogenous variation in senders' incentives which we exploited in the analysis of our on-line data, we account for selection into sending a positive amount here by estimating a Tobit model rather than a Heckman model. The results paint a picture qualitatively similar to the on-line data: amounts sent vary positively and significantly with both expected (lack of) cheating (*Pr(NotCheated)*) and expected return (*B_ return_ proportion*).

A.2 Treatments without cheating notion question

We also conducted (on-line) sessions of a treatment in which we dropped our direct cheating notion question and replaced it with a section where participants were asked to indicate how they would feel, as a sender, about various send/return amount scenarios. In total, 170 participants took part in this treatment. As with the main study, ten percent of participants were randomly chosen to be paid their experimental earnings.

To keep the number of individual questions reasonable, we selected three common send amounts—$S = 1, 5$ and 10—and, for each of these, asked participants how they would "feel" if the receiver returned four specific amounts: $0, \frac{S}{2}$, S and $\frac{f(S)}{2}$. These send/return

scenarios were chosen to line up with the cheating notions common in the data from our main study. In terms of feelings, for each send/return amount scenario participants were asked to select exactly two options from a list of several options that best described how they would feel if the scenario were realized. The list of options included positive evaluations ("[the receiver] was generous," "[the receiver] treated me fairly"), neutral evaluations ("[the receiver] was intelligent," "I have no particular opinion of [the receiver's] behavior") and negative evaluations ("[the receiver] cheated me," "[the receiver] disappointed me"). A free-form response option was also available.

To compare the qualitative data we have in this treatment with data from our main sessions, for each send/return scenario investigated in this treatment we calculate the proportion of participants in our main treatment who would feel cheated according to their own reported cheating notions. We compare this proportion to the proportion of respondents in the "feelings" treatment reporting feeling "disappointed" or "cheated." To maximize comparability, from our main treatment data we use only sessions where participants were allowed to opt out of specifying a cheating notion. We find a strong positive relationship between the proportion of participants expressing negative feelings in particular scenarios and the implied proportion of participants feeling cheated in those scenarios in the data from the main treatment (Figure A1). We interpret this as support for the view that trust game participants have well-defined cheating notions and evidence against the view that the cheating notions they report can be mainly attributed to priming.

A.2.1 Evidence on receivers' motivations

In sessions without a direct cheating notion question, at the end of the experiment we added a section in which participants were asked to descibe the rationale they used, if any, for deciding how much to return in the role of receiver. Participants were asked:

> *Describe, in general, how you arrived at your decisions concerning how much to return when you played role B [receiver] for each amount A could have sent you*

Participants could select among four pre-programmed options, or, if none on the list suited them they could select "other" and specify their own rationale. Three of the four pre-programmed responses were meant to capture positive reciprocity, ("the more A [the sender] sent, the more I returned in order to reward nice behavior"); negative reciprocity ("the less A [the sender] sent, the less I returned, in order to punish bad behavior"); vulnerability ("the more A [the sender] sent, the more I returned in order to compensate A [the sender] for being at the mercy of my actions"). The fourth pre-programmed option was essentially a decline to state option ("I did not have any particular rationale in mind.").

Table A11 presents the results. Overall, 83 percent of participants selected one of the four pre-programmed option. The modal response, selected by 42 percent of participants, was that receivers return more when senders send more to compensate senders for their vulnerability. The second most common response reflected positive reciprocity. Almost nobody (6 percent) selected negative reciprocity as their primary rationale, while a similarly low percentage selected the pre-programmed decline to state option (6 percent).

B Robustness checks on beliefs

A common concern whenever beliefs are elicited is the extent to which the elicitation mechanism itself colors reported beliefs. Monetary incentives meant to ensure that participants report beliefs truthfully may give rise to other potential confounds, such as hedging motives: by shading reported beliefs toward bad outcomes, individuals may reduce the variance of their experimental earnings. On the other hand, monetary incentives that are too weak can allow reported beliefs to be non-truthful for various reasons. In particular, one may worry that the significant correlation between *B_Cheat_notion* and receivers' return amounts arises because of a tendency for participants to ex-post rationalize their receiver strategies: by reporting believing that whatever they return is enough to not cheat others, participants can maintain a positive moral self-image.

First we consider ex-post rationalization. If ex-post rationalization is driving beliefs about others' cheating notions (*B_Cheat_notion*), then quadrupling the incentives for belief accuracy in the additional sessions should make this motive less relevant. Evidence of ex-post rationalization would be a consistently smaller correlation between return amounts and *B_Cheat_notion* in the "high belief pay" sessions.

As a simple test for ex-post rationalization, Table A12 (panel A) presents panel regressions of *B_Cheat_notion* as a function of return amounts incorporating a dummy for high belief pay and an interaction with return amounts. The coefficient of interest is on the interaction between high belief pay and return amount: if ex-post rationalization is important when belief pay is low, and diminished for high belief pay, we would expect this coefficient to be consistently negative and significant. Instead, the estimated coefficient on the interaction term is positive and marginally significant providing evidence against ex-post rationalization. Adding our standard set of demographics does not change the results. Moreover, restricting to the subset of observations where the receiver does not intentionally cheat—where the ex-post rationalization argument has the most bite—changes nothing qualitatively. We omit these last two robustness checks to save space, but they are available on request. It should also be noted that variation in belief pay could not have directly affected receivers' actions, since participants did not know there would be a belief elicitation section until after they had submitted their strategies.

Next, consider hedging motives. As a concrete example, consider a sender who has chosen to send 10 euros. If the sender believes the receiver is trustworthy and reports this belief, then in the good state of the world where the receiver *is* trustworthy, the sender earns a lot—both beliefs and actions pay off. However, in the bad state of the world, say, where the receiver returns nothing, the sender loses quite a lot—neither actions nor beliefs pay off. By shading reported beliefs downward—towards a higher likelihood of an untrustworthy sender—the sender can shift some earnings out of the good state of the world into the bad state of the world, reducing earnings variance, i.e., risk.

To test for hedging motives in beliefs, we estimate participants' stated beliefs about the amount of money receivers will return (*B_Receivers_actions*) for each possible send amount. We present panel regressions, where we control for whether a sender actually chose to send a particular amount, risk aversion and an interaction between these two variables. Since hedging motives can only (literally) apply to the send amount a sender actually chooses, one measure of the hedging motive is the coefficient on the dummy for

actually-chosen send amounts. A secondary prediction is that more risk averse individuals care about hedging more, so the interaction term should be negative. Table A12 (panel B) presents our estimates, which provide no support for the importance of hedging. In fact, contrary to hedging motives, reported beliefs about return amounts are marginally significantly *higher* for the amount a sender actually chose to send as evidenced by the coefficient on "Chosen send amount." Risk aversion plays no significant role. Controlling for demographics and/or the level of belief pay does not change anything qualitatively, so we omit these specifications.

C Additional Robustness checks on cheating notions

One additional concern with cheating notions is that they may be (reverse) caused by beliefs. Although priming is not an issue here, as we elicited beliefs after cheating notions, one explanation for the strong correlation between *Cheat_ notion* and *B_ Receivers_ actions* could be that that individuals simply report how much they expect back from receivers as their cheating notion. One reason this could happen is through an individual's desire to maintain a positive self-image and to avoid appearing, to themselves or to the experimenters, as "foolish" for allowing themselves to be cheated. To be clear, if senders expect not to be cheated and hence their cheating notion affects their reported beliefs, that is fine for our purposes. However, if participants first form beliefs about how much receivers will return and then report this belief as their cheating notion because of, e.g., a desire to not appear like a "sucker," then this calls into question the informativeness of the reported cheating notion.

In the latter case, it seems likely that such processes would affect reported cheating notions much more strongly for situations which could *actually* occur—i.e., for the one send amount an individual actually chooses. For concreteness, suppose an individual chooses to send $s = 3$ in the role of sender. Since this is an event that may actually occur, when asked about his or her cheating notion for $s = 3$ an individual may report his or her belief about how much the receiver will return instead of his or her cheating notion in order to avoid looking like a sucker if the event actually occurs. This might be particularly likely if *B_ Receivers_ actions* is less than *Cheat_ notion*. Such a process would tend to inflate reported cheating notions and, at the same time, overstate the correlation between *Cheat_ notion* and *B_ Receivers_ actions*. However, for all other send amounts ($s = 1, 2, 4, \ldots, 10$), since they cannot actually occur, such processes should have little effect on *Cheat_ notion* or its relationship with *B_ Receivers_ actions*.

To test for this effect, we report in Table A13 the results of ten separate regressions—one for each send amount—using *Cheat_ notion* as the dependent variable. On the right hand side, we include an individual's beliefs about the amount the receiver will return (*B_ Receivers_ actions*), a dummy indicating whether the individual chose to send the amount listed in the column heading and an interaction between these two variables. We control for our usual set of demographics, but as they have little explanatory power here we do not report them for ease of exposition.

We find that whether an individual actually chooses a particular send amount has no consistent effect on his or her reported cheating notion: half of the estimated coefficients on

Chosen send amount are positive, half are negative, and only one out of the ten coefficents is significant at conventional levels. Similarly, whether an amount was actually chosen has no consistent effect on the relationship between *B_ Receivers_ actions* and *Cheat_ notion*: five of the ten coefficients on the interaction between *B_ Receivers_ actions* and *Cheat_ notion* are positive, the other five are negative and only one out of the ten is statistically significant. Considered together, our results provide little evidence for cheating notions being reverse-caused by beliefs because, e.g., participants want to avoid looking like a sucker.

D Cheating notions and guilt aversion theory

In this section we test for the conjectured correlations between: i) *Cheat_ notion* and beliefs about receivers' actions (*B_ Receivers_ actions*); and ii) beliefs about others' cheating notions (*B_ Cheat_ notion*) and second-order beliefs (*B_ B_ Receivers_ actions*).

In Table A14 we report ten separate regressions—one for each send amount—using *B_ Receivers_ actions* as the dependent variable and, as the main explanatory variable, an individual's own personal cheating notion (*Cheat_ notion*). We control for available demographics and relevant experimental design features. In this latter category, we include a dummy for whether there was a sending fee in the session as this might factor into a sender's definition of return on investment. As a simple check on whether the reported beliefs are true beliefs, or rather whether the relationship between beliefs and cheating notions is driven by nuisance factors (e.g., ex-post rationalization), we include a dummy indicating sessions where we *quadrupled* belief elicitation incentives as well as an interaction term between this dummy and own cheating notions. The main lesson from this exercise is that one's own cheating notion is consistently a highly significant predictor of senders' first-order beliefs (*B_ Receivers_ actions*). The strength of the relationship is large in magnitude as well: a one euro increase in *Cheat_ notion* translates into a roughly 50 cent increase in *B_ Receivers_ actions*. Examining the coefficient on the interaction between cheating notions and belief elicitation incentives, we find that much stronger incentives have no consistent impact on this relationship and that, moreover, the impact is almost never significant. These patterns suggest that reported beliefs are true beliefs. Finally, it is worth noting that demographics have little explanatory power with one exception: gender. Male participants consistently expect about 40 to 50 cents less back from receivers than female participants.

In Table A15, we estimate receiver's second-order beliefs (*B_ B_ Receivers_ actions*) as a function of their beliefs about others' cheating notions (*B_ Cheat_ notion*). As before, we control for available demographics, relevant experimental design features, beliefs incentives and an interaction between beliefs incentives and reported beliefs about others' cheating notions. We find that beliefs about others' cheating notions are always highly significant predictors of second-order beliefs and that this relationship is also large in magnitude: a one-euro increase in *B_ Cheat_ notion* translates into a 34 to 83 cent increase in second-order beliefs with an average increase, over all ten send amounts, of about 60 cents. Strengthened belief incentives, again, have no consistent impact on this relationship and, moreover, their effect is almost never significant at conventional levels. Demographics play a slightly larger role here: being male or having more mathematical ability tends to lower second-order

beliefs; being older tends to raise them. The main lesson from Table 6, however, is that beliefs about others' cheating notions exhibit a strong positive relationship with second-order beliefs.

Appendix II: Experiment Instructions

In this experiment, you will be randomly paired with another participant and assigned randomly one of two roles: A or B. This pairing will be anonymous. Neither the person in the role of A nor the person in the role of B wil know with whom they have been paired.

The role of A

The player in the role of A is given 10.50 euros and must decide whether to send some all or none of this money to the player in the role of B, the person with whom A has been paired. [If A decides to send some of the this money, A will be charged a fee of 0.50 euros.] For every euro that A sends, B will receive more than 1 euro according to the table below.

If A sends €	1	2	3	4	5	6	7	8	9	10
B receives €	8.05	11.3	13.85	16.05	17.9	19.6	21.2	22.65	24.05	25.3

The role of B

After A makes his or her decision about how much to send to B, B decides how much of the money he or she receives—the amounts in the table above (8.05 euros, 11.30 euros, etc.)—to return to A. The player in the role of B will specify an amount to return for each possible amount they could receive. For example, if A sends 4 euros and B therefore receives 16.05 euros, B must decide how much of this 16.05 euros to return to A; and a decision must be made for every amount A could send (1,2,3,. . . ,10 euros).

Your earnings

For every pair of participants, one in the role of A and one in the role of B, the decisions that both A and B make determine the pairs earnings. Both A and B will be informed of the outcome determined by their choice.

In general:

- If A sends a positive amount to B:
    1. A's earnings will be: € 10.50 – (euros sent to B) + (euros returned by B) – (€ 0.50 fee)
    2. B's earnings will be: (euros received by B according to the table above) – (euros returned to A)
- If A sends nothing to B:
    1. A's earnings will be € 10.50
    2. B's earnings will be € 0.

Specifically, for every pair of players the result of this situation will be determined as follows:

i Every participant specifies their decision for each possible role (A and B).

ii The computer will randomly assign a role to each participant and randomly and anonymously pair each participant assigned the role of A with a participant assigned the role of B.

iii Within each pair, A's decisions will be combined with B's decision to determine the outcome for both A and B.

A Experiment Screens

A.1 Sender decision screen 1

If you are assigned the role of A, do you want to send money to B? If you send money, you will be charged a € 0.50 fee.

Choose "send" or "don't send" on this screen. If you choose "send", you will specify the amount to send on the next screen.

__ Send money
__Don't send money

A.2 Sender decision screen 2

How much money do you want to send if you are assigned the role of A?

__ € 1
__ € 2
...
__ € 10

A.3 Receiver decision screens

[There are 10 separate screens. A representative question is below.]

Imagine that you have been assigned the role of B ...
How much will you send back to A if A sends € 7 and you therefore receive € 21.20?

A.4 Cheating definition screen

If you are assigned the role of A, what is the minimum amount you would need to receive back from B in order to not feel cheated?

If you send €1 and therefore B receives €8.05, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating

__ I do not know

...
If you send €10 and therefore B receives €25.30, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating
__ I do not know

A.5 Belief elicitation

A.5.1 Instructions, screen 1

Now, we begin a new section. In this section as in the previous section, each question can contribute to your potential earnings.

Specifically, in this section you will be asked to estimate the choices other participants made in the previous section. Every question is about the choices of other participants, so please exclude your own actions from your estimations. The accuracy of your estimates will be calculated excluding your own actions as well.

Your earnings from this section will be determined by choosing one of your estimations at random and paying you according to the accuracy of this randomly chosen estimation. Every estimate has the same chance of being chosen by the computer. Your potential earnings from this experiment will be the sum of your earnings in this section and in the previous section.

The formula used to calculate your earnings from the randomly-chosen estimate is detailed on the next page.

A.5.2 Belief compensation formula screen

The method used to calculate your earnings from your estimates is detailed below. The most important thing to notice is that more accurate estimates have higher chances of earning money.

- Your estimate, R, is inserted into the following formula where "r" stands for the true value of the thing being estimated and "r_{max}" is the maximum value this true value can attain.

$1 - \left(\frac{R-r}{r_{\max}}\right)$

- This produces a number between 0 and 1. Call this number "z".
- The computer chooses a number between 0 and 1 with each number in between 0 and 1 being equally likely. Call this number "y".

- If $y \leq z$, you will earn €5.00 [€20.00] for your estimate.
- If $y > z$, you will earn €0.00 for your estimate.

An example

Suppose you are asked to estimate the average amount participants in the role of A send in the previous section of this experiment. And, imagine that this average turns out to actually be €4.00. The maximum value this average could have taken is €10. Therefore "r_{max}" in the equation above is 10 and r is 4. The equation therefore becomes:
$1 - \left(\frac{R-4}{10}\right)$
Notice that the closer your estimate, R, is to the actual value of 4 in our hypothetical example, the larger is z and therefore the larger is the probability of earning €5 [€20.00] for your estimate rather than €0.

- If your estimate is exactly correct, then (R-4)/10 = 0 and therefore z=1. Because the number chosen by the computer is at most one, an exactly correct estimate always pays €5 [€20.00].
- On the other hand, the probability with which your estimate earns you €5 [€20.00] diminishes the farther away from the true value your estimate is: z becomes smaller and so does the chances that $y < z$.

Click continue to begin start the estimation section

A.5.3 Beliefs elicitation screen 1

How much, on average, will players in the role of A send to B's? Insert a number between 0.00 and 10.00 : _ _ _

A.5.4 Beliefs elicitation screen 2

How much, on average, will B's return to A's?

If A sends €1 and B therefore receives €8.05, B's will return on average: _ _ _
...
If A sends €10 and B therefore receives €25.30, B's will return on average: _ _ _

A.5.5 Beliefs elicitation screen 3

What is the minimum amount (on average) that A's will need back from B's in order to not feel cheated?

If A sends €1 and B therefore receives €8.05, to not feel cheated A will need back from B at least: _ _ _

...

If A sends €10 and B therefore receives €25.30, to not feel cheated A will need back from B at least: ___

A.5.6 Beliefs elicitation screen 4

What percent of participants in the role of B will return enough money to you (if you are assigned the role of A) so that you don't feel cheated?

If you send €1 and B therefore receives €8.05, what percent of B's will return enough so that you don't feel cheated?: ___

...

If you send €10 and B therefore receives €25.30, what percent of B's will return enough so that you don't feel cheated?: ___

A.5.7 Beliefs elicitation screen 5

How much money (on average) do other participants in the role of A believe will be returned to them by B's?

If A sends €1 and B therefore receives €8.05, how much money does A believe B will return? ____

...

If A sends €10 and B therefore receives €25.30, how much money does A believe B will return? ____

Appendix III: Direct Response Experiment

Section 1: Experimental design and procedures

This appendix describes the procedures and provides instructions for the direct-response experiment.

The experiment was conducted in the laboratory at the Einaudi Institute for Economics and Finance using pen and paper. It consisted of two treatments: DR-CN and DR-FOB. The sole difference between the two treatments was what we elicited from senders and subsequently transmitted to receivers. In DR-CN we elicited and transmitted senders' cheating notions; in DR-FOB we elicited and transmitted senders' first-order beliefs about their receivers' actions.

Both treatments proceeded as follows. After arriving at the lab but before being seated all participants were presented instructions for our simplified trust game. Participants were told that the experiment they would participate in would involve this game. They were then publicly randomly assigned either the sender role or the receiver role.[1] Receivers were escorted to a separate waiting room where they were instructed to wait quietly for senders to make their decisions. Once all receivers had left the room, senders were assigned experiment codes in a transparently random fashion—by drawing numbered chips from an opaque bag. Each code corresponded to a seat in the lab. Seats were separated from each other by opaque dividers, essentially creating private cubicles.

After drawing a code, each sender was handed a decision sheet and instructed to go to their cubicle to fill out their sheet. Each decision sheet asked for only two pieces of information: i) the participant's experiment code; and ii) whether they would send 0, 5 or 10 euros to their co-player. The latter piece of information was supplied by ticking a box next to one of the three options. When all senders were finished making their decisions, decision sheets were collected and another sheet of paper was handed out. This sheet asked for three pieces of information: i) their experiment code; ii) their chosen send amount;[2] and iii) either their cheating notion (DR-CN) or how much money they believed their co-player would return to them (DR-FOB).

Both the cheating notion question and the (first-order) belief question were similar to the questions used in our main experiment, but adapted to refer only to the sender's chosen send

[1] For a session with N participants, (N/2) red poker chips and (N/2) blue poker chips were placed in an opaque bag and then each participant, without looking, drew one poker chip from the bag. Those who drew a red (blue) poker chip were assigned the role of sender (receiver). As in all of our experiments for this paper, more neutral wording was used. The sender role was always referred to as "Role A" while the receiver role was "Role B." If an odd number of participants showed up, one was randomly selected to be sent home and paid a 5 euro show-up fee.

[2] If a participant asked, they were instructed to simply check the same box they had checked before. Very few participants asked.

amount and the sender's specific co-player. The cheating notion question was: "How much money would you need back from player B [the receiver] in order to not feel cheated?" As in our main experiment, participants could specify a number or select either "I don't know" or "this has nothing to do with cheating." The first-order belief question was "How much money will player B [the receiver] send back to you?" Participants could insert a number or select "I don't know." As in our main experiment, proper incentives were provided for truthful belief reporting.[3] To enhance the credibility of our beliefs elicitation mechanism, we used a physical randomizing device to resolve uncertainty.[4]

When all senders had completed this final sheet they were escorted to the waiting room. At the same time, the receivers who had been waiting there were escorted to the laboratory. Upon entering the lab, receivers were randomly assigned an experiment code by drawing a chip from among the remaining chips in the opaque bag, which insures there was no duplication in code numbers. Each receiver was handed their own blank decision sheet as well as a decision sheet from one randomly selected sender and instructed to sit in their assigned cubicle. Each receiver's decision sheet asked for five pieces of information: i) the receiver's experiment code; ii) the experiment code of the sender with whom the receiver had been paired; iii) how much money their sender chose to send to them; iv) their sender's cheating notion (DR-CN) or first-order belief (DR-FOB); and, finally, v) the receiver's decision about how much money to return. Receivers could return any amount € 0.00 ≤ r ≤ € f(s).

Once all receivers had completed their decision sheet, they were escorted back to the waiting room. In the waiting room, experimental earnings were calculated. After each participant was paid individually in cash he or she was instructed to leave the premises before the next person would be paid. This design implements a nearly double blind procedure and ensures that each participant's decision is as consequential as possible. In addition to their experimental earnings, all participants were paid a 5 euro show-up fee.

[3] Differently from our main experiment, to ameliorate hedging motives senders were instructed that either the belief question or their trust game outcome would determine their earnings. Senders were informed that we would randomly draw a number from 1 to 100, with a number larger than 75 dictating that senders' earnings would be determined by the accuracy of their beliefs. As in our main experiment we used a randomized quadratic scoring to determine senders' potential earnings from their reported belief. Senders were provided with details of this scoring rule as well as a numerical example.

[4] At the front of the room was a miniature bingo blower containing balls numbered from 1 to 100. To decide whether beliefs would be remunerated we extracted a number from this bingo blower in front of all senders. This number was extracted after all senders had submitted their beliefs but before they left the room.

Section 2: Experimental materials

Sheet 1: General game description provided to all participants before role assignment

The Game

Your experiment code is _____

General Instructions

In this experiment, you will be paired randomly with one other participant and randomly assigned one of two roles: **A** or **B**. This pairing will be anonymous. Neither the person assigned the role **A** nor the person assigned the role **B** will discover with whom they have been paired.

*The role of **A***:
The player assigned the role **A** is given €10.50 and must decide whether to send some, all or none of this money to the player assigned the role **B**, the player with whom **A** has been paired. For every euro that **A** sends, **B** receives more than one euro as reported in the table below.

If **A** sends:	€ 0	€ 5	€ 10
B receives:	€ 0	€ 17.90	€ 25.30

*The role of **B***:
After **A** makes his or her decision about how much to send to the player assigned the role of **B**, **B** must decide how much of the money he or she receives to send back to **A**. The possible amounts **B** can receive are reported in the table above. For example, if **A** sends € 5 and **B** therefore receives € 17.90, **B** must decide how much of this € 17.90 to send back to **A**.

Your earnings:
For every pair of participants, one assigned the role of **A** and one assigned the role of **B**, the decision of **A** together with the decision of **B** will determine both **A**'s and **B**'s earnings. Both **A** and **B** will be informed of the outcome determined by their decisions. However, you will not discover who your co-player was and your co-player will not discover who you are.

In general:

- if **A** sends a positive amount,
  - **A's** earnings will be: (€ 10.50) - (euro sent to **B**) + (euro sent back by **B**);
  - **B's** earnings will be: (the euro value associated with **A**'s send amount reported in the table above) - (the amount returned to **A**).

- If **A** sends zero euros to **B**,
  - **A's** earnings are € 10.50;
  - **B's** earnings are € 0.

Sheet 2: Sender's initial decision sheet (DR-CN and DR-FOB)

ROLE A

Your experiment code is _____

After you have read the game instructions on the previous page carefully, please respond to the key question below.

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

Thank you for participating in this role.

Sheet 2: Sender's final decision sheet (DR-CN)

ROLE A

Your experiment code is _____

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

QUESTION: *What is the minumum amount you would need to receive back from player B in order to not feel cheated? [leave the space blank if you chose to send € 0]*

YOUR RESPONSE:

Insert a number: € _ _ . _ _

… or choose one of the following two options:

[] I don't know
[] this has nothing to do with cheating

Sheet 2: Sender's final decision sheet (DR-FOB)

ROLE A

Your experiment code is _____

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

QUESTION: *How much money will player B will return to you? [Leave blank if you chose to send € 0]*

YOUR RESPONSE:

Insert a number: € _ _ . _ _

… or choose the following option:

[] I don't know

NB:

- Your earnings from this latter question will depend on how accurate your guess is (for details, see the next page).
- You will be paid <u>either</u> your earnings from this question <u>or</u> your earnings from the game.
- To determine whether this question determines your earnings, before you leave this room we will extract a number from 1 to 100 using the randomizing device at the front of the room. If the extracted number is greater larger than 75, your earnings will be determined by this question.

Back of Sheet 2 (DR-FOB)

How we will calculate your earnings from this question:

We use the following method to calculate your earnings from the latter question in euros. The most important feature to notice is that more accurate estimates yield higher a probability of earning money.

- Your estimate, call this "R", is inserted into the following formula where "r" denotes the true value of the number being estimated and "r_{max}" denotes the maximum value the number being estimated can attain.

$$1 - \left(\frac{R - r}{r_{max}}\right)^2$$

- This produces a number between 0 and 1. We will multiply the number produced by 100 to obtain a number between 0 and 100. Call this number "z".
- At the same time, we will choose randomly a number between 0 and 100. Call this number we randomly select "y".

If y ≤ z, you will earn € 15 for your estimate,

If y > z, you will earn € 0 for your estimate.

If this question is chosen to determine your earnings, "y" will be chosen by extracting a second number using the randomizing device at the front of the room.

An example:

Imagine you are estimating the average amount that participants in the role of A will send in this game. To be concrete, suppose this average actually turns out to be 4. The maximum value this average could attain is 10, so that "r_{max}" = 10. Plugging both of these facts into the equation above yields:

$$\frac{z}{100} = 1 - \left(\frac{R - 4}{10}\right)^2$$

Now, notice that the closer your estimate, R, comes to the actual value, 4, the higher "z" will become and, consequently, the larger will be the probability that you will earn € 15 for your estimate instead of nothing.

For example, if your estimate is exactly correct, i.e., R = 4, then $\left(\frac{R-4}{10}\right)^2$ = 0 and therefore z = 100. Since the number we will randomly draw, "y," is always less than 100, your exactly correct estimate would earn you € 15 with certainty.

On the other hand, the farther away your estimate R is from the the true value, the larger z will become. Since this means that the probability that y ≤ z also increases, your chances of earning € 0 instead of € 15 from your estimate also increase.

Sheet 3: Receiver's decision sheet

Role B

Your experiment code is _____

Please read the instructions for the game. Then, read through the additional materials provided to discover: i) your co-player's code; ii) how much money your co-player in Role A decided to send to you; and [Treatment CN: *iii) how much your co-player needs back in order to not feel cheated*.] [Treatment FOB*: iii) how much your co-player believes you will send back*.] *Please write these facts in the spaces below.*

My co-player's code is: _____

My co-player sent € _ _ . _ _ , so that I received € _ _ . _ _.

[*Treatment CN:* My co-player needs back in order to not feel cheated: € _ _ . _ _]
[*Treatment FOB:* My co-player believes I will send back: € _ _ . _ _]

Next, if your co-player sent you some money please choose how much you will return.

KEY QUESTION: How much will you return to A?

YOUR RESPONSE:

I will send back to A € _ _ . _ _

Thank you for participating in this role.

Table A1: Comparison of behavior in the lab and on-line, summary statistics

	Send > 0	Send amount	Return proportion	B_return_proportion	Proportion of non-cheaters	Pr(NotCheated)
			In-lab sessions			
	0.97	5.28	1.25	1.36	0.43	0.56
	(0.03)	(0.45)	(0.10)	(0.10)	(0.06)	(0.03)
Obs	36	36	36	36	36	36
			On-line low fee sessions			
	0.90	4.83	1.28	1.22	0.53	0.53
	(0.03)	(0.26)	(0.06)	(0.06)	(0.03)	(0.02)
Obs	150	150	149	148	150	135

Table A2: Proportion of participants with a cheating notion (Cheat_notion), in-lab sessions

	Send amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ cheating notion	0.72	0.86	0.83	0.92	0.94	0.94	0.97	0.97	0.97	0.97
	(0.08)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Obs	36	36	36	36	36	36	36	36	36	36

Notes: [1] Raw proportions reported. [2] Standard errors appear in parentheses

Table A3: Proportion of participants who would feel cheated by (return amount) < (send amount), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ (cheating notion) ≥ (send amt)	0.88	0.84	0.97	0.94	0.94	0.97	0.89	0.83	0.83	0.86
	(0.06)	(0.07)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.06)	(0.06)	(0.06)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. [2] Standard errors appear in parentheses

Table A4: Proportion of participants for whom Cheat_notion is consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Weakly positive return on investment	0.08 (0.05)	0.10 (0.05)	0.30 (0.09)	0.18 (0.07)	0.18 (0.07)	0.12 (0.06)	0.20 (0.07)	0.14 (0.06)	0.11 (0.05)	0.23 (0.07)
Strictly positive return on investment	0.15 (0.07)	0.16 (0.07)	0.50 (0.09)	0.39 (0.08)	0.29 (0.08)	0.32 (0.08)	0.29 (0.08)	0.29 (0.08)	0.23 (0.07)	0.31 (0.08)
Inequality Aversion	0 --	0.03 (0.03)	0 --	0.03 (0.03)	0.06 (0.04)	0.03 (0.03)	0 --	0.03 (0.03)	0.23 (0.07)	0.37 (0.08)
Equal split	0.35 (0.10)	0.23 (0.08)	0.20 (0.07)	0.18 (0.07)	0.32 (0.08)	0.32 (0.08)	0.34 (0.08)	0.23 (0.07)	0.23 (0.07)	0.37 (0.08)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. Classifications are not mutually exclusive so that, e.g., the same cheating notion can be labeled as consistent with both SPROI and Inequality aversion. [2] Standard errors are in parentheses. [3] A weakly positive return on investment (WPROI) cheating notion entails reporting exactly the send amount (s) as one's cheating threshold in sessions without a sending fee. [4] "SPROI" (strictly positive return on investment) is a more generous definition of WPROI taking into account a reasonable interest rate, r = 10%. We multiply the send amount by 1+r to get an "exact SPROI" definition. To be as generous as possible to this notion, and to account for the fact that experimental participants typically have a well-known predilection to state whole-number values, we then calculate the least integer greater than this exact value, denoted by ceiling("exact SPROI"). For each send amount, *s*, We label as SPROI all cheating thresholds falling within the interval with integer end-points: [s, ceiling("exact SPROI")]. [5] "Inequality Aversion" refers to a cheating notion which requires equal monetary outcomes, and we label a cheating notion as consistent with inequality aversion if it lies within the smallest closed interval with integer endpoints containing this outcome. As an example, consider s = 1. The total surplus in this case is 10.50 – 1 + 8.05 = 17.55, and half of this surplus is 8.775. Any cheating notion in the interval [8, 9] would therefore be labeled as consistent with inequality aversion. [6] An "Equal-split" (ES) cheating notion entails a cheating threshold of half of the entire amount allocated to the receiver. As with SPROI and Inequality Aversion above, to account for participants' predilection for whole numbers, the definition of ES for each send amount, s, includes all cheating thresholds falling within the smallest interval with whole-number end-points containing a precisely-equal split of the receivers' total earnings: i.e., $\frac{f(s)}{2} \in$ [n, n+1]. For example, if a sender sends s = 3, a receiver receives f(s) = 11.30, and $\frac{f(s)}{2}$ = 5.65. Consequently, ES for s = 3 would include all cheating thresholds within the interval [5, 6].

Table A5: Proportion of participants whose beliefs about others' cheating notions (B_Cheat_notion) are consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Weakly positive return on investment	0.28	0.22	0.17	0.17	0.25	0.17	0.14	0.17	0.14	0.19
	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
Strictly positive return on investment	0.39	0.36	0.39	0.28	0.31	0.28	0.28	0.28	0.22	0.33
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)
Inequality Aversion	0	0	0	0	0.06	0.08	0	0.14	0.22	0.31
	--	--	--	--	(0.04)	(0.05)	--	(0.06)	(0.07)	(0.08)
Equal split	0.25	0.31	0.33	0.14	0.33	0.36	0.36	0.25	0.22	0.31
	(0.07)	(0.08)	(0.08)	(0.06)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)
Obs	36	36	36	36	36	36	36	36	36	36

Table A6: Beliefs about the amount receivers will return (B_Receivers_actions) as a function of own cheating notions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	0.75***	0.75***	0.79***	0.53***	0.57***	0.51***	0.62***	0.84***	0.53***	0.64***
	(0.10)	(0.10)	(0.08)	(0.11)	(0.09)	(0.11)	(0.19)	(0.24)	(0.18)	(0.18)
Constant	0.17	0.48	0.74	2.01**	2.21***	2.74**	2.47	0.99	3.94*	3.04
	(0.30)	(0.46)	(0.50)	(0.78)	(0.78)	(1.04)	(1.79)	(2.26)	(2.02)	(2.13)
Observations	26	31	30	33	34	34	35	35	35	35
R-squared	0.72	0.65	0.76	0.42	0.53	0.42	0.24	0.28	0.21	0.27

Table A7: Beliefs about senders' beliefs about amount receivers will return (B_B_Receivers_actions), as a function of beliefs about others' cheating notions (B_Cheat_notion), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.62***	0.61***	0.62***	0.57**	0.59***	0.63***	0.68***	0.64***	0.65***	0.67***
	(0.18)	(0.20)	(0.21)	(0.24)	(0.21)	(0.20)	(0.20)	(0.17)	(0.17)	(0.17)
Constant	0.85	1.40*	1.67	2.42	2.89*	3.12*	2.81	3.59*	3.64*	3.80*
	(0.54)	(0.81)	(1.12)	(1.51)	(1.50)	(1.66)	(1.83)	(1.80)	(1.95)	(2.09)
Observations	36	36	36	36	36	36	36	36	36	36
R-squared	0.25	0.22	0.20	0.14	0.19	0.23	0.26	0.28	0.29	0.30

Table A8: Intentional cheating (reduced form), in-lab sessions

	Sent Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	-0.78*	-0.35**	-0.08	-0.05	0.02	-0.08	-0.05	-0.02	-0.25**	-0.13*
	(0.43)	(0.17)	(0.12)	(0.11)	(0.07)	(0.09)	(0.11)	(0.11)	(0.12)	(0.07)
B_Cheat_notion	1.08**	0.32	0.15	0.16	0.18	0.05	0.25**	0.11	0.28*	0.34***
	(0.50)	(0.20)	(0.17)	(0.14)	(0.13)	(0.13)	(0.11)	(0.13)	(0.16)	(0.11)
Constant	-0.68	0.27	-0.36	-0.73	-1.13	0.30	-1.49	-0.41	-0.11	-1.98*
	(0.64)	(0.62)	(0.74)	(0.87)	(0.94)	(0.85)	(1.09)	(1.06)	(1.01)	(1.11)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Each column presents estimates from a Probit model, with the (binary) dependent variable being "receiver intentionally cheats if sent relevant amount." Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated, i.e., by the event r < B_Cheat_notion. This threshold amount is also inserted as a control in each estimate by the variable "B_Cheat_notion." [3] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%.

Table A9: Intentional cheating (reduced form), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (r ≥ B_Cheat_notion)										
B_Cheat_notion	1.24***	1.09***	0.90***	1.08***	1.02**	1.27***	0.44	1.00***	0.93***	0.76
	(0.32)	(0.32)	(0.26)	(0.18)	(0.34)	(0.32)	(0.60)	(0.27)	(0.24)	(0.48)
Constant	0.30	0.51	1.37	1.05	1.36	-0.42	6.27	1.95	2.12	4.74
	(0.77)	(1.16)	(1.32)	(1.07)	(2.28)	(2.61)	(4.90)	(2.58)	(2.53)	(5.05)
Obs	15	16	17	19	15	18	14	11	15	14
R-squared	0.53	0.46	0.44	0.68	0.40	0.50	0.04	0.60	0.55	0.17
Conditional on cheating (r < B_Cheat_notion)										
B_Cheat_notion	0.44**	0.43**	0.20	0.37	0.45	0.25	0.44*	0.71***	0.53**	0.40
	(0.19)	(0.18)	(0.21)	(0.22)	(0.28)	(0.22)	(0.21)	(0.21)	(0.23)	(0.28)
Constant	-0.35	0.10	1.84	1.47	0.85	2.84	1.96	-0.19	1.61	2.63
	(0.61)	(0.81)	(1.13)	(1.42)	(2.16)	(1.86)	(2.13)	(2.28)	(2.70)	(3.53)
Obs	21	20	19	17	21	18	22	25	21	22
	0.22	0.24	0.05	0.16	0.12	0.07	0.17	0.32	0.22	0.09

Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a simple OLS regression of return amount conditional on beliefs about others' cheating notion for the send amount listed in the column heading. [3] The top panel is restricted to observations not involving intentional cheating, while the bottom panel is restricted to observations involving intentional cheating.

Table A10: Send amount (Tobit), in-lab sessions

	Dependent variable = send amount		
	(1)	(2)	(3)
Pr(NotCheated)	4.29*	4.94**	6.97***
	(2.19)	(2.25)	(1.90)
B_return_proportion	1.54*	1.57**	1.41*
	(0.81)	(0.75)	(0.77)
Male		1.53*	0.93
		(0.81)	(0.75)
Age		-0.16*	-0.30**
		(0.09)	(0.11)
Math score		-0.34	0.07
		(0.42)	(0.37)
Risk aversion			-0.47**
			(0.17)
Altruism			0.04
			(0.21)
30≤ Income <45			-1.48
			(0.98)
45≤ Income <70			0.03
			(1.10)
45≤ Income <70			1.76
			(1.47)
Income ≥120			-2.99**
			(1.26)
Constant	0.86	5.85	8.66*
	(1.52)	(4.60)	(5.01)
Obs	36	34	32

Notes: [1] Robust standard errors in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] Each column presents a Tobit model estimate where the dependent variable is *how much* the sender sends and censoring below 0 is taken into account. [5] "Pr(NotCheated)" is our measure of participants' subjective beliefs about not being cheated, described in the text. [6] "B_return_proportion" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send amounts. [7] "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse). [8] Altruism is how much emphasis participants' parents placed on the value "help others" during their upbringing. [9] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table A11: Proportion of receivers specifying a particular rationale

	Overall	High fee sessions	Low fee sessions
Sender vulnerability	0.42	0.40	0.45
	(0.04)	(0.05)	(0.06)
Positive reciprocity	0.29	0.31	0.27
	(0.04)	(0.05)	(0.05)
Negative reciprocity	0.06	0.06	0.05
	(0.02)	(0.03)	(0.03)
No motive	0.06	0.05	0.08
	(0.02)	(0.02)	(0.03)
Obs	170	93	77

Notes: [1] Raw proportions reported; [2] Standard errors in parentheses; [3] Proportions in each column sum to less than one, with the unaccounted for observations being participants who elected to supply their own rationale rather than one of the four pre-programmed rationale; these self-supplied rationale varied widely and are not easily classifiable.

Table A12: Robustness checks on beliefs, main study data

Panel A: checking for ex-post rationalization

Dependent variable = *B_Cheat_notion*

Return amount	Amount sent	High belief pay	(High belief pay) X (Return amt)	Cons	Obs	Individuals	R^2
0.11***	0.85***	0.12	0.05*	1.58***	4254	428	0.5
(0.02)	(0.03)	(0.24)	(0.03)	(0.20)			

Panel B: checking for hedging motives in beliefs

Dependent variable = *B_Receivers_actions*

Amount sent	Chosen send amount	Risk aversion	(Chosen send amt) X (Risk aversion)	Cons	Obs	Individuals	R^2
0.82***	0.29*	-0.00	-0.02	1.61***	4146	417	0.34
(0.02)	(0.17)	(0.04)	(0.03)	(0.23)			

Notes: [1] Both the top and bottom panel report individual random effects regressions pooling observations across all send amounts. [2] Robust standard errors, clustered by session, appear in parentheses. [3] "High belief pay" is a dummy taking the value of one if the session involved a 20 euro maximum belief pay, and 0 if the maximum possible belief pay was 5 euros; "Chosen send amount" is a dummy variable indicating the amount a participant actually chose to send in the role of sender; "Risk aversion" is an incentive-compatible index of risk aversion obtained from a previous experiment. [4] We drop observations for which we have no measure of risk aversion.

Table A13: Robustness check on own cheating notion, main study data

Dependent variable = *Cheat_notion*										
	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Receivers_actions	0.66[A]	0.66[A]	0.68[A]	0.58[A]	0.48[A]	0.54[A]	0.57[A]	0.50[A]	0.54[A]	0.52[A]
	(0.06)	(0.09)	(0.08)	(0.10)	(0.08)	(0.11)	(0.07)	(0.09)	(0.08)	(0.07)
Chosen send amount	-0.18	0.42	-0.30	1.30	-2.22[B]	-1.16	0.73	1.02	-1.62	1.04
	(0.75)	(0.79)	(0.88)	(1.07)	(0.69)	(1.21)	(0.97)	(2.46)	(1.52)	(2.29)
Chosen send amount X B_Receivers_actions	0.11	-0.32	0.18	-0.29	0.31[B]	0.07	0.00	-0.02	-0.03	-0.08
	(0.29)	(0.32)	(0.25)	(0.23)	(0.09)	(0.18)	(0.08)	(0.35)	(0.11)	(0.18)
Low Fee	-0.08	0.00	-0.11	-0.26	-0.10	-0.06	0.41	0.23	0.43[C]	0.05
	(0.15)	(0.18)	(0.11)	(0.20)	(0.24)	(0.13)	(0.27)	(0.27)	(0.22)	(0.26)
Constant	1.99C	2.70[B]	2.19[C]	3.89[B]	6.36[B]	4.92[B]	2.55	6.74[B]	4.92[C]	5.68[B]
	(0.96)	(0.91)	(0.93)	(1.48)	(2.09)	(1.81)	(2.01)	(2.17)	(2.23)	(1.97)
Demographic controls?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.33	0.39	0.30	0.25	0.30	0.32	0.25	0.31	0.29

Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' personal cheating notions (*Cheat_notion*). [2] Robust standard errors, clustered by session, appear in parentheses. [3] Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "B_Receivers_actions" is a participant's belief about how much a receiver will return for the send amount indicated in the column heading; "Chosen send amount" is a dummy variable indicating the participant actually chose to send the amount in the column heading in the role of sender. [5] Demographic controls are included but not reported for readability. The set of demographic controls is identical to the set reported in Table 6 in the manuscript. "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros. [6] Observations vary over columns because we do not have demographics for all participants and because not all participants reported a cheating notion for all send amounts.

Table A14: Beliefs about the amount receivers will return as a function of own cheating notions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	0.61^{A}	0.58^{A}	0.52^{A}	0.46^{A}	0.36^{A}	0.46^{A}	0.57^{A}	0.50^{A}	0.51^{A}	0.52^{A}
	(0.06)	(0.02)	(0.02)	(0.06)	(0.06)	(0.04)	(0.03)	(0.10)	(0.10)	(0.10)
Male	-0.30^{C}	-0.49^{B}	-0.34^{C}	-0.53^{A}	-0.43^{A}	-0.37^{C}	-0.23	-0.22	-0.49	-0.28
	(0.15)	(0.17)	(0.17)	(0.13)	(0.10)	(0.19)	(0.22)	(0.32)	(0.31)	(0.25)
Age	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.02	0.04	-0.00	-0.00
	(0.01)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)
Math score	-0.04	-0.08^{C}	-0.06	0.01	0.10	0.02	0.01	0.05	-0.01	-0.04
	(0.05)	(0.03)	(0.05)	(0.07)	(0.09)	(0.10)	(0.07)	(0.15)	(0.13)	(0.17)
Risk aversion	0.03	0.02	0.04	0.02	0.02	0.07	0.01	0.11	0.06	0.17
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	(0.09)	(0.10)	(0.11)	(0.10)	(0.15)
30≤ Inc <45	0.18	0.47^{B}	0.34	0.62^{B}	0.12	0.07	-0.06	-0.31	-0.08	-0.17
	(0.19)	(0.15)	(0.24)	(0.25)	(0.30)	(0.37)	(0.32)	(0.48)	(0.54)	(0.66)
45≤ Inc<70	0.24	0.31	0.32	0.57^{B}	0.29	0.38^{C}	-0.04	-0.25	-0.15	-0.19
	(0.13)	(0.25)	(0.29)	(0.24)	(0.26)	(0.19)	(0.43)	(0.32)	(0.34)	(0.37)
70≤ Inc <120	-0.03	0.17	0.39	0.66^{B}	0.34	0.52	-0.14	0.06	0.01	-0.26
	(0.19)	(0.18)	(0.25)	(0.27)	(0.31)	(0.46)	(0.68)	(0.68)	(0.74)	(0.91)
Inc ≥120	0.26	0.19	0.14	0.39	0.01	-0.36	-0.08	-0.18	-0.42	-0.83
	(0.26)	(0.21)	(0.21)	(0.30)	(0.38)	(0.53)	(0.67)	(0.59)	(0.64)	(0.85)
Low Fee	-0.13	-0.23	-0.30^{B}	-0.14	-0.20	-0.30^{B}	-0.52^{B}	-0.48^{C}	-0.61^{B}	-0.25
	(0.11)	(0.21)	(0.12)	(0.22)	(0.14)	(0.12)	(0.20)	(0.24)	(0.23)	(0.26)
High belief Incentives	0.22	0.71^{A}	0.21	0.09	-0.49	-0.35	0.62	0.08	-0.41	-0.71
	(0.15)	(0.16)	(0.32)	(0.74)	(0.81)	(0.87)	(0.46)	(1.31)	(1.31)	(1.48)
Own cheating notion X High belief Incentives	-0.12	-0.16^{B}	0.00	0.00	0.08	0.05	-0.06	-0.02	0.04	0.05
	(0.06)	(0.06)	(0.05)	(0.11)	(0.09)	(0.09)	(0.06)	(0.12)	(0.12)	(0.11)
Constant	1.00	1.27	1.90^{C}	2.45^{C}	2.24^{C}	2.96^{C}	3.00^{B}	1.70	3.72^{C}	3.44
	(0.81)	(0.74)	(0.84)	(1.05)	(1.06)	(1.30)	(1.16)	(1.36)	(1.71)	(1.84)
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.34	0.38	0.28	0.23	0.28	0.31	0.25	0.30	0.29

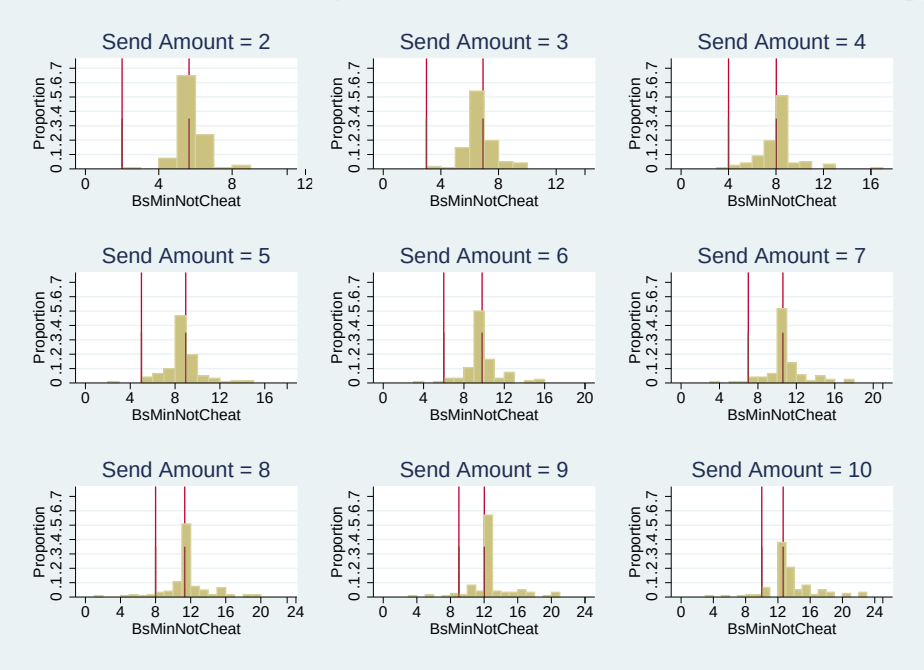
Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' beliefs about the amount receivers will return (*B_Receivers_actions*). [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable is a participant's own cheating notion. Additional demographic controls include: "Math score" = self-reported score on required math exams taken during the final year of high school in Italy; "Risk aversion" = an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism from a prior, unrelated, experiment, which takes values from 1 (risk loving) to 10 (very risk averse); "Inc" = self-reported annual family income from all sources, in thousands of euros, net of taxes. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because not all participants reported a cheating notion for every send amount and because we do not have demographics for all participants. [7] The coefficients and significance levels on the main explanatory variable, "Own cheating notion," are virtually identical if demographics are omitted. From s = 1, …, 10, the coefficients and significance levels are: 0.59^{A}, 0.59^{A}, 0.54^{A}, 0.46^{A}, 0.37^{A}, 0.45^{A}, 0.58^{A}, 0.49^{A}, 0.50^{A}, 0.51^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 2.

Table A15: Second-order beliefs (B_B_Receivers_actions) as a function of beliefs about others' cheating notions (B_Cheat_notion)

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.83^{A}	0.66^{A}	0.69^{B}	0.84^{A}	0.58^{A}	0.65^{A}	0.51^{B}	0.34^{A}	0.46^{A}	0.45^{A}
	(0.12)	(0.14)	(0.21)	(0.07)	(0.12)	(0.08)	(0.18)	(0.08)	(0.08)	(0.06)
Male	-0.26^{B}	-0.54^{A}	-0.57^{A}	-0.55^{B}	-0.56^{B}	-0.64^{A}	-0.77^{A}	-0.73^{A}	-0.92^{B}	-0.96^{B}
	(0.09)	(0.08)	(0.12)	(0.16)	(0.18)	(0.13)	(0.17)	(0.16)	(0.37)	(0.31)
Age	0.05^{C}	0.05^{B}	0.07^{B}	0.07^{B}	0.07^{B}	0.08^{B}	0.10^{C}	0.09^{C}	0.08	0.06
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)
Math score	-0.10^{B}	-0.13^{B}	-0.09^{C}	-0.18^{C}	-0.06	-0.03	-0.11	-0.19C	-0.07	-0.04
	(0.04)	(0.04)	(0.04)	(0.08)	(0.07)	(0.08)	(0.10)	(0.08)	(0.13)	(0.12)
Risk aversion	-0.01	-0.02	-0.00	-0.00	-0.02	-0.03	-0.04	0.03	-0.00	0.03
	(0.02)	(0.03)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.08)
30≤ Inc <45	-0.28	-0.36^{B}	-0.16	-0.39^{C}	-0.11	-0.28	-0.19	-0.14	-0.14	-0.61
	(0.17)	(0.14)	(0.17)	(0.20)	(0.20)	(0.15)	(0.32)	(0.30)	(0.36)	(0.33)
45≤ Inc<70	0.06	0.04	0.26	0.35	0.52C	0.31	0.23	0.25	0.31	0.16
	(0.09)	(0.29)	(0.26)	(0.26)	(0.26)	(0.37)	(0.23)	(0.29)	(0.38)	(0.46)
70≤ Inc <120	-0.15	-0.30	-0.04	-0.02	0.05	0.08	0.02	-0.14	0.14	0.09
	(0.09)	(0.31)	(0.25)	(0.31)	(0.29)	(0.37)	(0.33)	(0.45)	(0.48)	(0.65)
Inc ≥120	-0.17	-0.65^{C}	-0.83	-0.72	-0.81	-0.57	-0.45	-1.08	-0.55	-0.64
	(0.18)	(0.30)	(0.62)	(0.62)	(0.82)	(0.89)	(0.87)	(0.93)	(0.99)	(0.99)
Low Fee	0.03	-0.14	-0.24	-0.23	-0.18	0.02	-0.14	-0.21	-0.14	-0.24
	(0.07)	(0.16)	(0.20)	(0.19)	(0.23)	(0.09)	(0.21)	(0.25)	(0.28)	(0.31)
High Belief Incentives	0.21	-0.20	-0.14	0.43	-0.88	-0.45	-1.57	-4.01^{A}	-2.53	-2.54^{C}
	(0.46)	(0.58)	(1.08)	(0.54)	(1.06)	(0.78)	(1.73)	(0.95)	(1.37)	(1.12)
Est. others' cheating notion X High Belief Incentives	-0.20	0.00	-0.07	-0.20^{C}	0.05	-0.02	0.11	0.34^{A}	0.19	0.19^{C}
	(0.14)	(0.14)	(0.21)	(0.09)	(0.15)	(0.09)	(0.19)	(0.09)	(0.11)	(0.08)
Constant	0.59	1.67	1.28	1.63	2.35	1.76	3.71	6.34^{A}	4.91^{C}	5.57^{B}
	(0.86)	(1.15)	(1.66)	(1.45)	(1.78)	(1.71)	(2.55)	(1.68)	(2.11)	(1.99)
Observations	375	375	375	375	375	375	375	375	375	375
R-squared	0.45	0.40	0.34	0.39	0.33	0.34	0.32	0.32	0.32	0.34

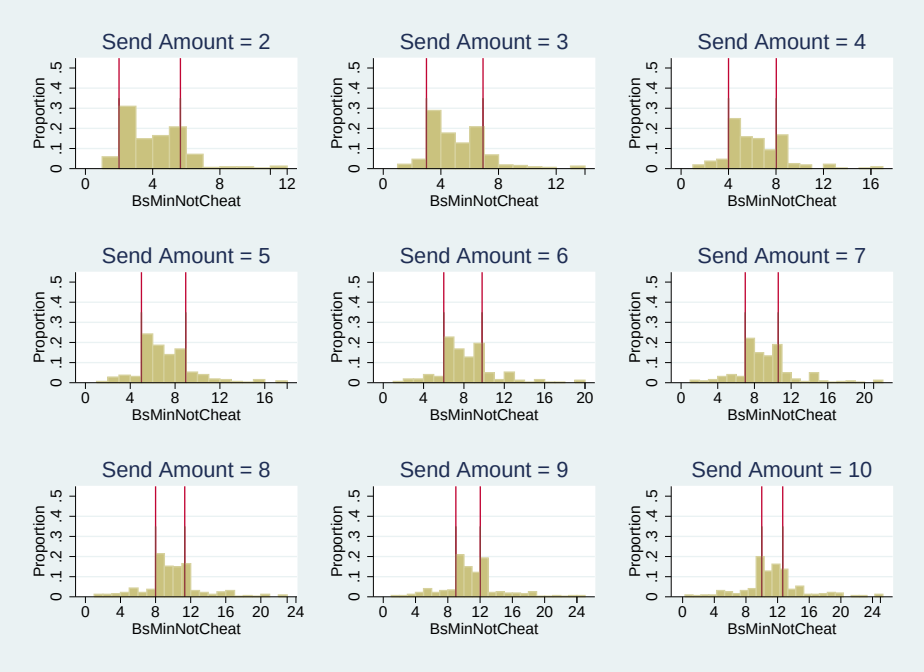
Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' second-order beliefs *B_B_Receivers_actions*. [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "B_Cheat_notion" is a participant's belief about others' cheating notions. Other demographic controls are identical to those in Table 6, above. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because we do not have demographics for all participants. [7] If demographics are omitted, the coefficients and significance levels on the main explanatory variable, "B_Cheat_notion," are virtually identical. From s = 1, …, 10, the coefficients and significance levels are: 0.84^{A}, 0.69^{A}, 0.73^{A}, 0.83^{A}, 0.63^{A}, 0.68^{A}, 0.55^{A}, 0.40^{A}, 0.49^{A}, 0.48^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 8.

Figure A1a: Individual-level Consistency of B_Cheat_notion across Send Amounts, equal split



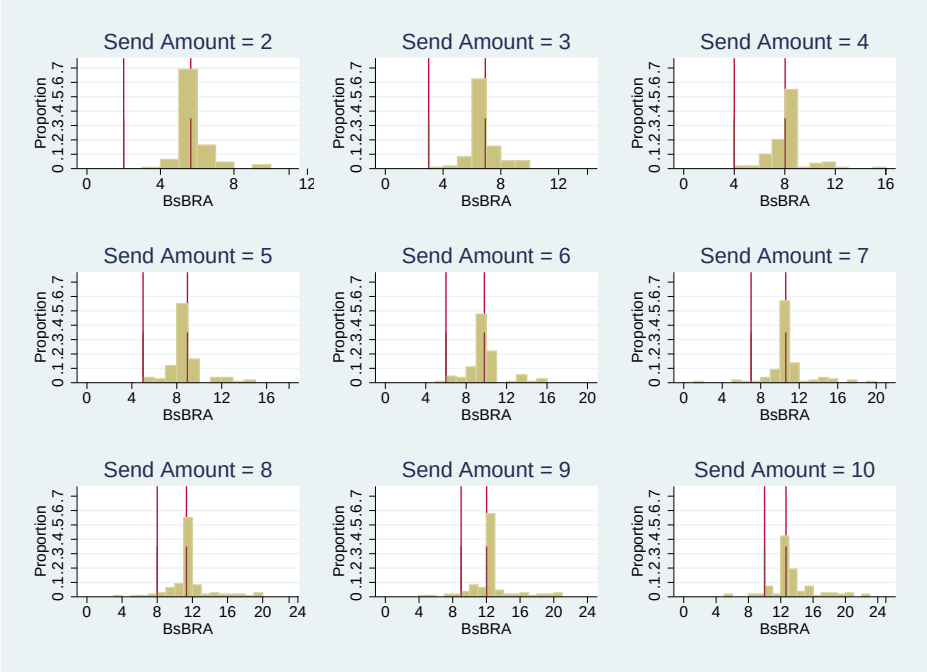
Notes: [1] The figure restricts attention to participants whose beliefs about others' cheating notions (*B_Cheat_notion*) were consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' beliefs about others' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A1b: Individual-level Consistency of B_Cheat_notion across Send Amounts, strictly positive return on investment



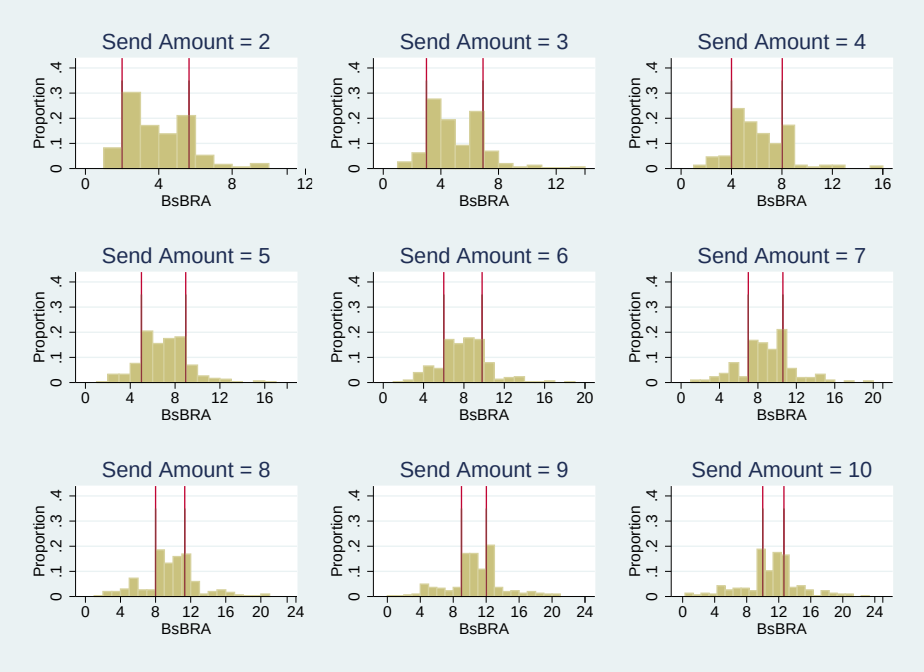
Notes: [1] The figure restricts attention to participants whose beliefs about others' cheating notions (*B_Cheat_notion*) were consistent with strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' beliefs about others' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A2a: Individual-level Consistency of B_B_Receivers_actions across Send Amounts, equal split



Notes: [1] The figure restricts attention to participants whose second-order belief (*B_B_Receivers_actions*) was consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' *B_B_Receivers_actions* for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A2b: Individual-level Consistency of B_B_Receivers_actions across Send Amounts, strictly positive return on investment



Notes: [1] The figure restricts attention to participants whose second-order belief (*B_B_Receivers_actions*) was consistent with a strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' *B_B_Receivers_actions* for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A3: Comparison of proportion feeling cheated by elicitation method

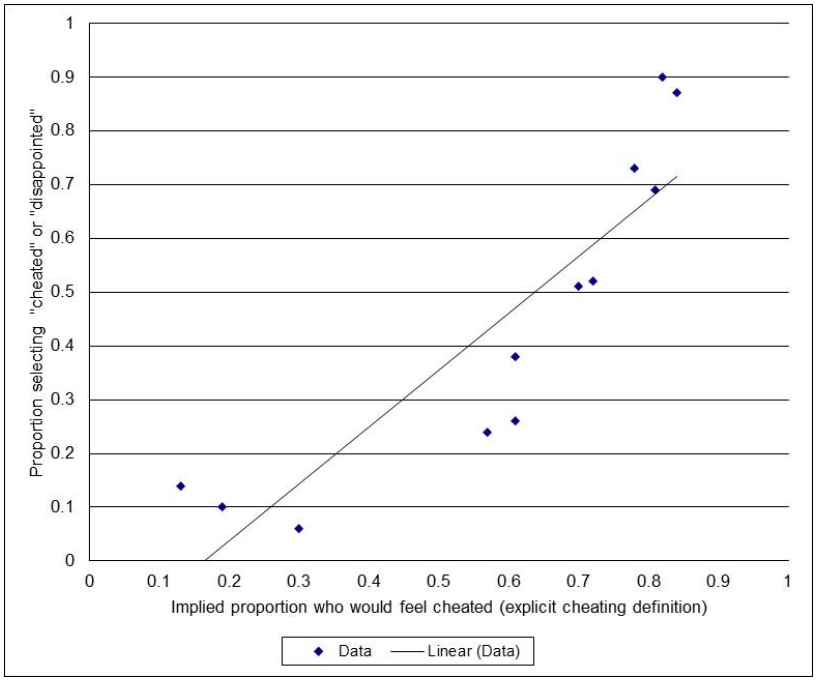


Table 1
Experimental design

	Number of sessions	Explicit cheating notion question opt-out	Investment fee	Max belief pay	Obs
Initial study	4	No	0.50 euro	5 euro	122
Additional sessions	4	Yes	0.50 euro (2 sessions) 0.00 euro (2 sessions)	20 euro	306

Table 2
Descriptive statistics

	Mean	Std Dev	Min	Max	N
Male	0.46	0.499	0	1	420
Age	23.73	4.171	18	58	420
Math score	7.66	1.251	3	10	402
Inc<30K	0.29	0.455	0	1	391
30≤Inc<45	0.24	0.426	0	1	391
45≤Inc<70	0.25	0.431	0	1	391
70≤Inc<120	0.16	0.366	0	1	391
Inc≥120K	0.07	0.249	0	1	391
Risk aversion	5.71	2.193	1	10	417
Send decision (binary)	0.81	0.392	0	1	428
Send amount	4.31	3.232	0	10	428
Average return proportion	1.28	0.697	0	4.02	427
B_return_proportion	1.27	0.637	0	4.02	425
Competitive values emphasis	0.62	0.196	0	1	410
Good values emphasis	0.76	0.149	0.17	1	404
Pr(NotCheated)	0.42	0.232	0	1	427
Average proportion of non-cheaters	0.49	0.376	0	1	428

Table 3
Variable Description

Variable Name	Question
Cheat_notion	This is shorthand for "Cheating notion" and is a participant's answer to the question "If you are assigned the role of A [sender] what is the minimum amount you would need to receive back from player B [receiver] in order to not feel cheated? …If you were to send €[s] and B were to therefore receive €[f(s)], you would need back how many euros?"
B_Cheat_notion	This is shorthand for "Beliefs about Cheating notions". They are the answers to the set of questions: "What is the minimum amount (on average) that A's will need back from B's in order to not feel cheated? If A sends €[s] and B therefore receives €[f(s)], to not feel cheated A will need back from B at least: €__.__"
B_Receivers_actions	This is shorthand for "My Belief about Receivers' Actions" and is the answer to the set of questions: "How much, on average, will B's return to A's? If A sends €[s] and B therefore receives €[f(s)], B's will return on average: €__.__"
B_B_Receivers_actions	This is shorthand for "Beliefs about Others' Beliefs about Receivers' Actions." These are the answers to the set of questions "How much money (on average) do other participants in the role of A believe will be returned to them by B's? If A sends €[s] and B therefore receives €[f(s)], how much money does A believe B will return? €__.__"
B_NotCheated	This is shorthand for "Beliefs about the Probability of Not Feeling Cheated" These are participants' answers to the set of questions: "What percent of participants in the role of B will return enough money to you (if you are assigned the role of A) so that you will not feel cheated? …If you send €[s] and B therefore receives €[f(s)], what percent of B's will return enough so that you will not feel cheated? ."

Note: Each variable listed in this table is actually a set of ten variables, one for each possible send amount s = 1, …, 10. However, as in the table, we will typically suppress the dependence on s for ease of exposition.

Table 4
Proportion of participants in sessions who opt-out of reporting a cheating notion in sessions with explicit opt-out opportunities

	Send Amount										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Obs
	Proportion who selected "this has nothing to do with cheating"										
Mean	0.20	0.18	0.17	0.15	0.13	0.13	0.13	0.13	0.14	0.13	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
	Proportion who did not report a cheating notion for any reason										
Mean	0.23	0.21	0.21	0.17	0.15	0.15	0.15	0.16	0.17	0.17	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	

Notes: [1] In sessions with an explicit "opt-out" possibility participants could refrain from specifying an explicit personal cheating notion and instead respond either "I don't know" or "this has nothing to do with cheating." [2] The top row of Table 4 presents the proportion of participants who chose "this has nothing to do with cheating," while the lower row presents the proportion of participants who chose either of these two "opt-outs" or left the question entirely blank.

Table 5
Determinants of cheating notions

Dependent variable = *Cheat_notion*										
Cooperative values	Competitive values	€ sent	(€ sent)^2	Male	Age	Math score	Risk aversion	Cons	Obs	Individuals
-2.55**	1.63**	1.07***	-0.02***	-0.47	0.00	-0.02	-0.11	3.55***	3496	354
(1.09)	(0.64)	(0.07)	(0.01)	(0.43)	(0.03)	(0.11)	(0.08)	(1.31)		

Notes: [1] Estimates are from an individual-level random effects regression model. [2] Variables present in the regression, but omitted for readability: full set of income dummies; dummy for sessions with no investment fee; dummy for sessions comprising the initial study. None of these variables had significant coefficients. [3] Robust standard errors, clustered by session, appear in parentheses.

Table 6
Receivers' decision to intentionally cheat, by send amount

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	-0.08**	-0.13***	-0.08*	-0.07***	-0.04	-0.04*	-0.04	-0.05**	-0.03*	-0.04*
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
B_Cheat_notion	0.22***	0.21***	0.23***	0.22***	0.17***	0.15***	0.16***	0.17***	0.12***	0.14***
	(0.04)	(0.06)	(0.05)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
Male	0.07	-0.02	0.18	0.06	0.03	-0.05	-0.16	-0.07	0.01	-0.06
	(0.07)	(0.16)	(0.12)	(0.14)	(0.20)	(0.15)	(0.14)	(0.13)	(0.13)	(0.18)
Age	-0.03	-0.04***	-0.03***	-0.02**	-0.02**	-0.03**	-0.01	0.01	-0.01	-0.01
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Math score	-0.04	-0.03	0.04	0.05	-0.04	-0.03	-0.06***	-0.06	-0.08**	-0.03
	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)	(0.09)	(0.02)	(0.06)	(0.04)	(0.04)
Risk aversion	-0.00	0.03	0.01	-0.00	-0.02	-0.02	0.00	-0.01	-0.02	-0.07***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)
30≤ Inc <45	-0.02	0.18	0.24**	0.22	0.09	0.25*	0.50*	0.06	0.10	0.16
	(0.19)	(0.16)	(0.12)	(0.21)	(0.23)	(0.15)	(0.26)	(0.19)	(0.12)	(0.21)
45≤ Inc<70	0.12	0.01	0.06	0.10	0.29*	0.23***	0.43	0.24**	0.12	0.15
	(0.16)	(0.08)	(0.13)	(0.17)	(0.17)	(0.07)	(0.28)	(0.12)	(0.14)	(0.20)
70≤ Inc <120	0.17	0.17	0.07	-0.05	0.33*	0.41*	0.58**	0.70***	0.04	0.16
	(0.33)	(0.18)	(0.21)	(0.18)	(0.19)	(0.22)	(0.24)	(0.16)	(0.20)	(0.33)
Inc ≥120	0.00	-0.21	-0.07	0.01	-0.51	0.02	-0.21	-0.44	-0.04	-0.56*
	(0.35)	(0.28)	(0.21)	(0.20)	(0.32)	(0.28)	(0.40)	(0.31)	(0.29)	(0.33)
Constant	0.45	0.85	-0.69	-1.02*	-0.07	-0.10	-0.76*	-0.97	-0.05	-0.42
	(0.76)	(0.52)	(0.52)	(0.60)	(0.41)	(0.79)	(0.41)	(0.81)	(0.70)	(0.63)
Obs	369	366	366	369	371	370	371	369	366	366

Notes: [1] Each column presents estimates from a Probit model. Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated. [2] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%. [3] Math score is individual's self-reported score on required math exams taken during the final year of high school in Italy. [4] Income variables refer to self-reported annual family income from all sources, in thousands of euros, net of taxes. The excluded category is "below 30 thousand euros annually". [5] Observations vary over columns because not all participants reported a cheating notion for every send amount. This is discussed in the text. Additionally, we do not have demographics for all participants.

Table 7
Senders' decisions, Heckman estimates

	Main equation (1)	Selection equation (2)
Pr(NotCheated)	2.76**	0.57
	(1.38)	(0.65)
B_return_proportion	1.34***	0.28**
	(0.45)	(0.12)
Pr(NotCheated)x B_return_proportion	-1.57*	-0.07
	(0.85)	(0.46)
Low fee (dummy)	--	0.68***
		(0.09)
Age	0.11***	0.00
	(0.03)	(0.02)
Male	0.36	0.35**
	(0.32)	(0.14)
Math score	-0.00	0.12***
	(0.09)	(0.04)
Risk aversion	-0.14***	0.04
	(0.05)	(0.03)
Altruism	0.03	0.04
	(0.12)	(0.04)
30≤ Income <45	-0.29	0.13
	(0.42)	(0.25)
45≤ Income <70	-0.22	-0.04
	(0.59)	(0.23)
70≤ Income <120	-0.62**	-0.08
	(0.29)	(0.13)
Income ≥120	-0.63	0.74*
	(0.70)	(0.40)
Constant	1.45	-1.62***
	(2.16)	(0.61)
Obs	350	350
Mills Ratio	0.33	
	(0.18)	

Notes: [1] Robust standard errors, clustered by session, appear in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] For the Heckman model (cols 1-2): the dependent variable in the main equation is *how much* the sender sends; the dependent variable in the selection equation takes the value of 1 if the sender sends a positive amount and 0 otherwise; [4] The exclusion restriction for the selection equation consists of a dummy for "Low fee" sessions, a dummy taking the value of one if the observation came from a session where senders were charged nothing to send a positive amount, and 0 if the observation came from a session where senders were charged € 0.50 to send a positive amount [5] "Pr(NotCheated)" is our measure of probability about not being cheated, described in the text; "B_return_proportion" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send

amounts; "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse); "*Altruism*" is how much emphasis participants' parents placed on the value "help others" during their upbringing. [6] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table 8
Predicting receiver behavior: second-order beliefs or cheating notion beliefs?

Dependent variable = return amount conditional on send amount in column heading										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.25*	0.33***	0.31***	0.18***	0.32***	0.21**	0.19*	0.25***	0.17**	0.24***
	(0.11)	(0.09)	(0.07)	(0.04)	(0.06)	(0.07)	(0.08)	(0.05)	(0.06)	(0.06)
B_B_receivers_actions	0.31**	0.07	0.11	0.15*	0.11	0.13	0.16	0.11	0.21**	0.08
	(0.12)	(0.12)	(0.07)	(0.07)	(0.12)	(0.10)	(0.11)	(0.07)	(0.07)	(0.10)
No Personal Cheat Notion (NPCN)	-0.17	-0.76	-1.12**	-2.01***	-1.62	-1.20	-3.23**	-1.45	-4.15***	-3.83**
	(0.41)	(0.43)	(0.39)	(0.54)	(0.98)	(1.79)	(1.13)	(1.18)	(0.97)	(1.59)
NPCN X B_Cheat_notion	-0.03	-0.25	-0.23	-0.09	-0.33**	0.09	0.00	0.24	0.41**	0.08
	(0.19)	(0.19)	(0.15)	(0.10)	(0.12)	(0.16)	(0.35)	(0.13)	(0.12)	(0.13)
NPCN X B_B_receivers_actions	0.16	0.41	0.49**	0.35**	0.66**	0.11	0.37	0.01	0.06	0.33
	(0.27)	(0.29)	(0.14)	(0.10)	(0.24)	(0.41)	(0.38)	(0.17)	(0.21)	(0.23)
Constant	-0.02	0.70	2.82**	4.23***	2.59*	2.43*	4.01***	4.80**	3.78**	4.16*
	(1.51)	(0.90)	(0.90)	(0.82)	(1.30)	(1.05)	(0.95)	(1.62)	(1.56)	(1.78)
Demographics?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	375	375	375	375	375	375	375	375	375	375
R-squared	0.20	0.15	0.16	0.13	0.19	0.12	0.14	0.15	0.17	0.14

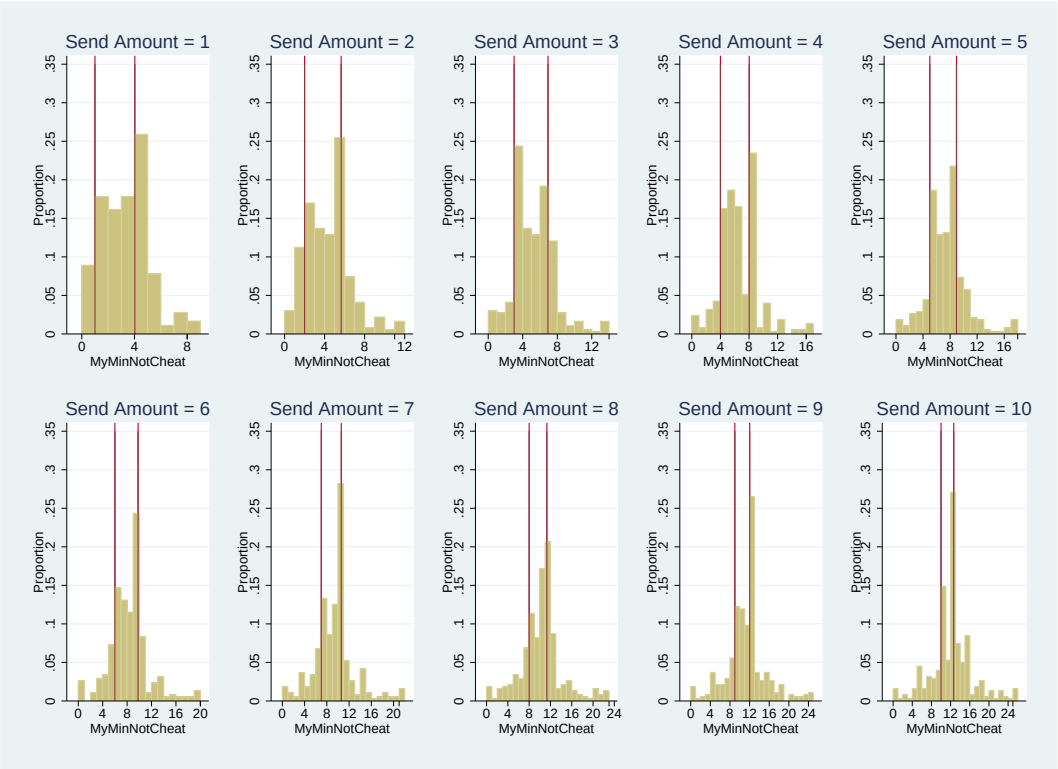
Notes: [1] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents an OLS estimate using the dependent variable r(s), where s is specified in the column heading. [3] The reported independent variables in column *i* are: "B_Cheat_notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, *i*=1,…,10; "B_B_receivers_actions" is each participant's belief about the average amount of money the sender believes the receiver will send back when the sender sends *i* euros, *i*=1,…,10. [4] Each estimate includes demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion.

Table 9
Sensitivity of amounts returned to beliefs about senders' cheating notions by decision to cheat, Heckman models

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (r ≥ B_Cheat_notion)										
B_Cheat_notion	1.17***	1.02***	0.97***	1.19***	1.07***	0.95***	0.88***	0.89***	1.09***	1.02***
	(0.17)	(0.13)	(0.14)	(0.25)	(0.15)	(0.11)	(0.16)	(0.13)	(0.22)	(0.13)
Constant	3.83**	4.98**	3.54**	4.68*	5.06**	4.88***	5.96***	4.97**	8.15**	9.26***
	(1.95)	(2.25)	(1.73)	(2.78)	(2.39)	(1.85)	(2.10)	(2.00)	(4.06)	(3.02)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
	Wald test: *B_Cheat_notions* coefficient = 1 (p-value)									
	0.32	0.86	0.84	0.43	0.63	0.66	0.43	0.39	0.67	0.84
Conditional on cheating (r < B_Cheat_notion)										
B_Cheat_notion	0.42***	0.37***	0.57***	0.38***	0.44***	0.43***	0.49***	0.58***	0.63***	0.53***
	(0.07)	(0.06)	(0.10)	(0.10)	(0.10)	(0.09)	(0.13)	(0.12)	(0.14)	(0.12)
Constant	-0.16	0.93	0.18	0.95	1.68	0.03	1.09	0.09	-0.38	-0.01
	(0.92)	(1.02)	(1.51)	(1.92)	(1.91)	(2.01)	(2.87)	(3.02)	(3.36)	(3.31)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
	Wald test: *B_Cheat_notions* coefficient = 0.5 (p-value)									
	0.23	0.04	0.47	0.24	0.52	0.43	0.95	0.52	0.34	0.79

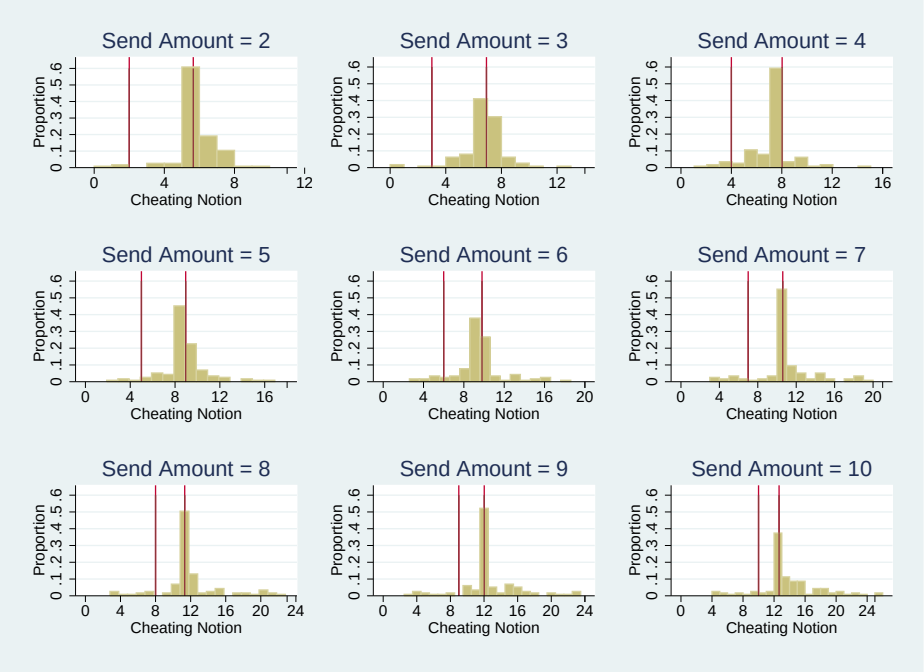
Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a Heckman model estimate using as its exclusion restriction participants' own cheating notions. [3] The dependent variable in column i is the amount a participant will send back if the sender sends i euros, i=1,…,10. [4] The reported independent variables in column *i* are: "B_Cheat_notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, i=1,…,10. [5] Each estimate includes our standard set of demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion.

Figure 1
Own Cheating Notions (Cheat_notion)



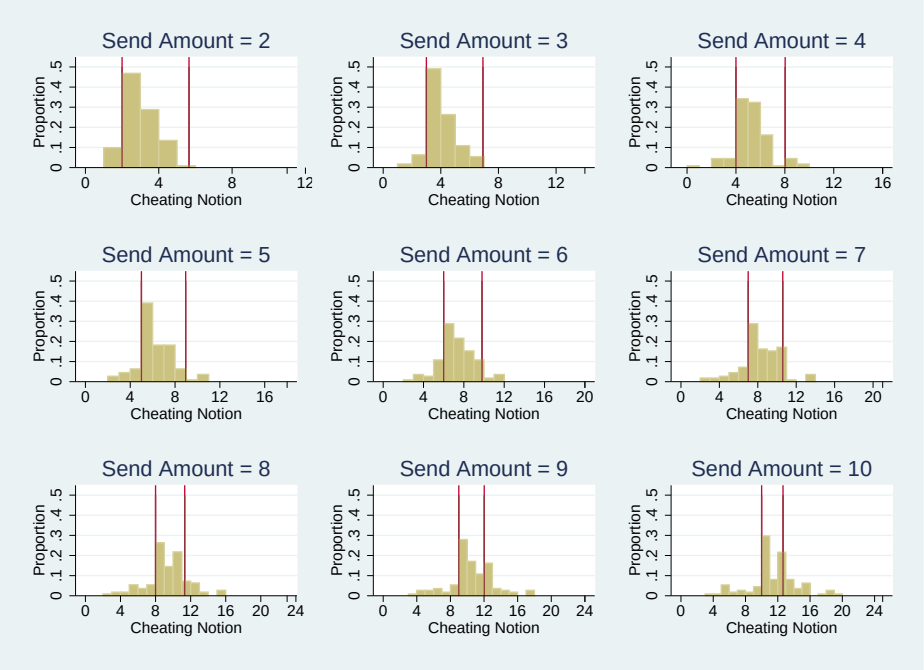
Notes: [1] The figure reports histograms of participants' personal cheating notions for each send amount s=1,….,10. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a *weakly positive return on investment* cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an *equal split* cheating definition.

Figure 2A: Within-individual Consistency of Cheating Notions across Send Amounts, equal split



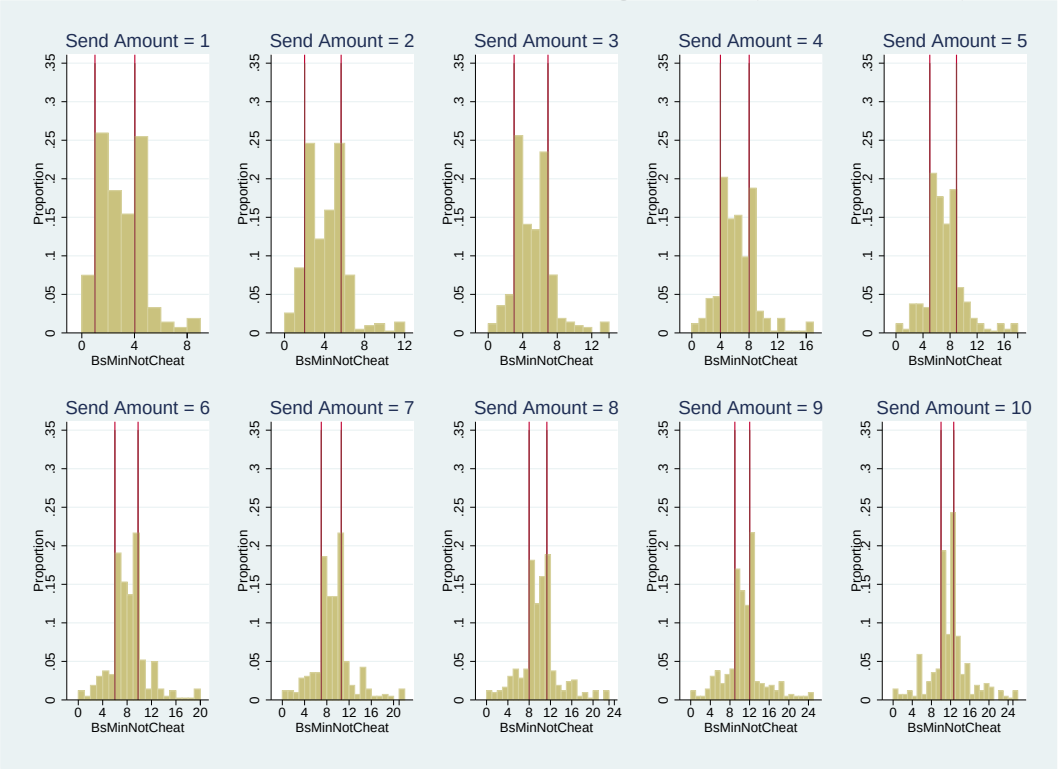
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 2B: Individual-level Consistency of Cheating Notions across Send Amounts, strictly positive return on investment



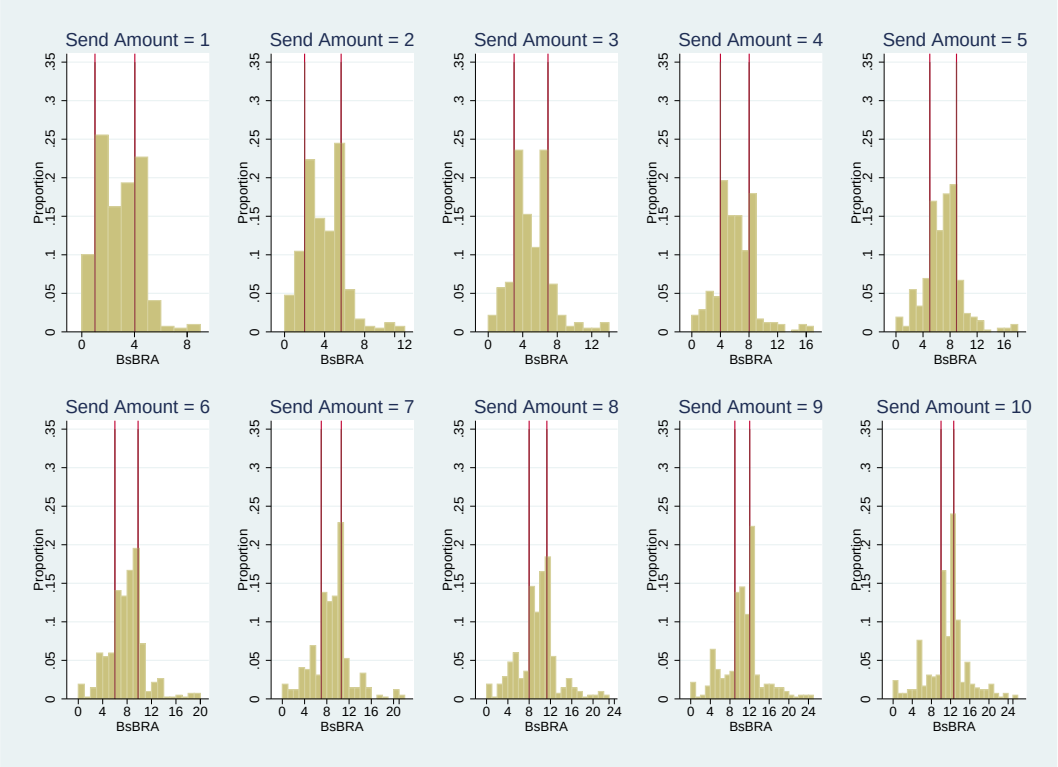
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 3
Participants' Beliefs about Others' Cheating Notions (B_Cheat_notion)



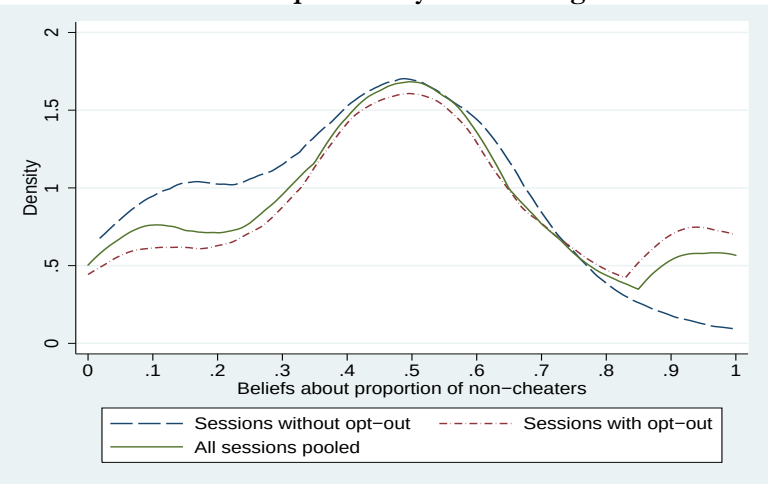
Notes: [1] The figure reports histograms of participants' beliefs about other participants' cheating notions (B_Cheat_notion). [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 4
Second-order beliefs (B_B_receivers_actions)



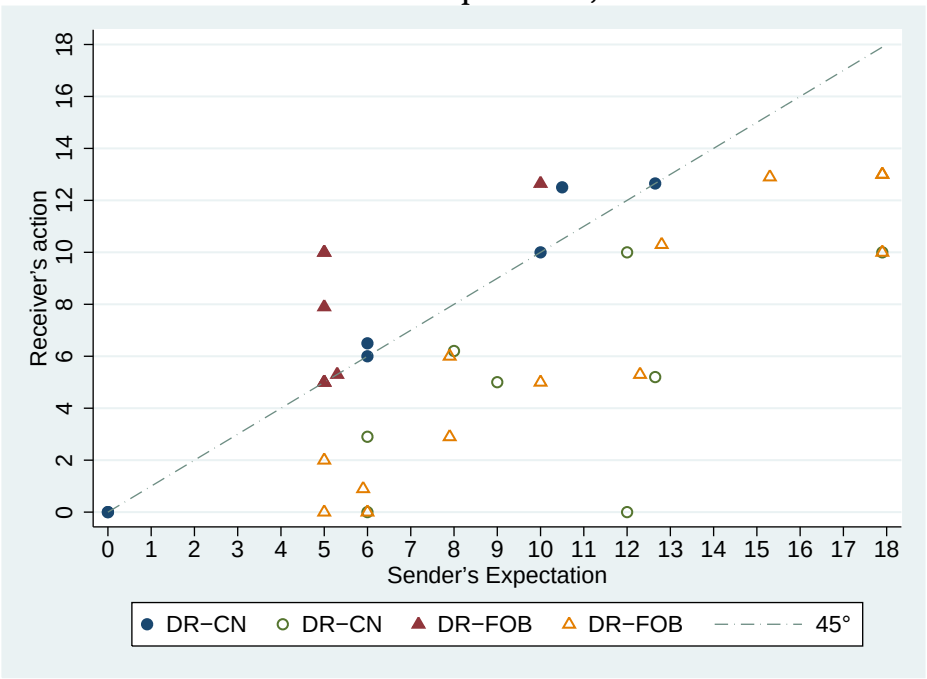
Notes: [1] The figure plots participants' beliefs about senders' beliefs about receivers' actions (B_B_receivers_actions). [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure 5
Beliefs about the probability of not being cheated



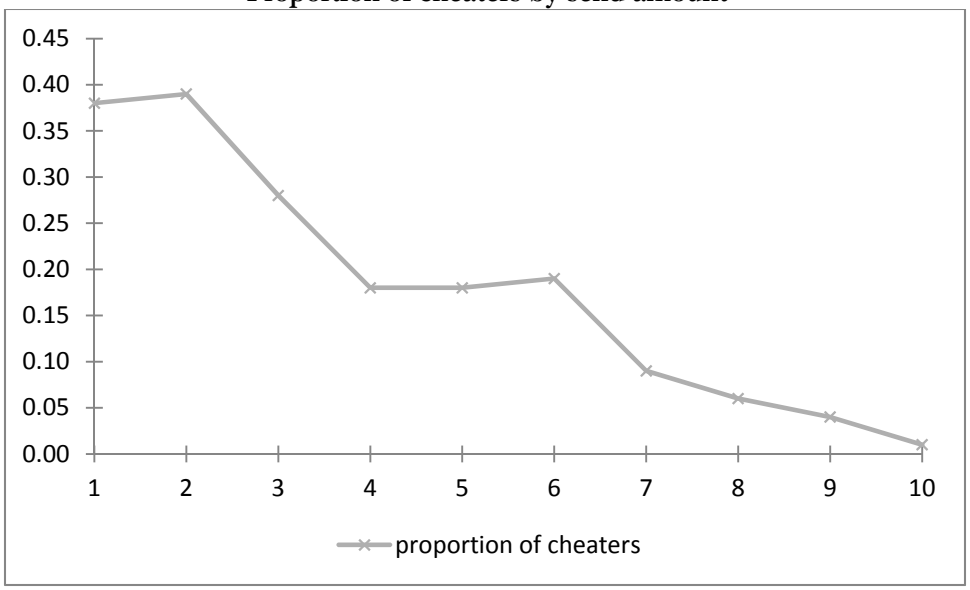
Notes: Observations in the sessions with opt-out (short-dash line) are restricted to individuals who have a cheating notion for every possible amount a sender could send. This is to ensure our summary measure of beliefs about the probability of being cheated is well-defined. Thus the density plot for the additional sessions is based on 207 (out of 306) observations.

Figure 6
Receiver's actions vs. sender's expectations, DR-CN and DR-SOB



Notes: [1] The figure restricts attention to observations in the direct-response experiment where s > 0 and plots each receiver's action against his or her sender's moral (DR-CN) or mathematical (DR_FOB) expectation. [2] Solid markers correspond to observations where the receiver did not cheat – i.e., returned at least as much as their sender's expectation – while hollow markers correspond to observations where the receiver cheated. [3] The dashed line is a 45-degree line along which a receiver's action exactly matches his or her sender's expectation.

Figure 7
Proportion of cheaters by send amount



Notes: The figure reports the proportion of cheaters (y-axis), after partialling out the effect of expectations of others' cheating notions, for each possible send amount (x-axis).

Not for publication

Appendix I: Robustness Checks

A Additional Robustness check treatments

In addition to our main experiment described in Appendix II, two further treatments were conducted for robustness. First of all, to check whether there is something peculiar about the on-line environment driving our results or whether paying only 10 percent of participants provides incentives that are too weak, we ran two sessions in the laboratory where 100 percent of participants were paid. As a second robustness exercise, we conducted sessions in which our direct cheating notion question was omitted and replaced with a series of questions asking participants how they would feel about various possible outcomes in the trust game from the point of view of the sender. The purpose of this latter treatment is to address the concern that our direct cheating notion question might prime participants to associate cheating with the trust game.

A.1 In-lab sessions

In total, 36 individuals took part in two sessions conducted in the experimental laboratory at the Einaudi Institute for Economics and Finance in Rome, Italy. Participants were recruited from the same subject pool as were the on-line sessions. There was no overlap in actual participants—i.e., no participant took part in both an on-line session and an in-lab session. All in-lab participants were paid based on their choices in the experiment and the accuracy of the their reported beliefs.

Apart from taking place in the laboratory, the design of this treatment and the materials used were exactly the same as the on-line treatments. Participants simply completed the on-line experiment in the laboratory. All sessions of the in-lab experiment allowed participants to opt out of specifying a cheating notion by selecting one of two responses: "I don't know" or "this has nothing to do with cheating." Neither session featured a fee to send a positive amount.

In Table A1 we report summary statistics for both the in-lab and most comparable on-line sessions. Receivers' behavior does not change much across these two environments: average return proportions and the propensity to intentionally cheat are all quite similar. Beliefs about these return proportions (*B_return_proportion*) and the likelihood of being cheated are also quite similar across the two environments. On the other hand, in-lab senders were slightly more likely to send a positive amount than their on-line counterparts, raising the average amount sent by in-lab senders. However, conditional on sending a positive amount average send amounts were again quite similar: 5.36 in on-line low fee sessions; 5.43 in the laboratory; with standard errors 0.25 and 0.44, respectively.

In terms of cheating notions (*Cheat_notion*), the picture is also quite similar in the lab and on-line experiments: the vast majority of participants have a cheating notion for all possible send amounts (Table A2); the vast majority have a cheating notion *at least as demanding as* the weakly positive return on investment (TableA3). Considering the proportion of participants whose cheating notions are consistent with various definitions

(Table A4), we again see that the weakly positive return on investment describes a small minority of participants, while a similar but relaxed notion, a strictly positive return on investment, describes a substantial minority of participants for most send amounts, as does an equal split rule: over all send amounts, these two rules each account for about 27%-29% of participants' reported cheating notion. We also, again, find that literal inequality aversion fits very few participants' definitions of cheating. We find the same patterns when considering beliefs about others' cheating notions (Table A5), which is also consistent with our on-line findings.

Considering next the relationship between second-order beliefs (*B_ B_ Cheat_ notion*) and cheating notions and related beliefs, the in-lab environment delivers similar patterns as those found in the on-line environment. Own cheating notions are again highly predictive of beliefs about how much receivers will return (*B_ Receivers_ actions*) (Table A6). In-lab beliefs about others' cheating notions (*B_ Cheat_ notion*) are highly predictive of in-lab second-order beliefs (*B_ B_ Receivers_ actions*) (Table A7). As in the on-line data, *Cheat_ notion* is typically negatively related to intentional cheating while *B_ Cheat_ notion* is usually positively related to intentional cheating (Table A8).

In Table A9, we replicate the pattern suggesting that beliefs about others' cheating notions (*B_ Cheat_ notion*) function as thresholds for those who refrain from cheating. Because we have many fewer observations here, to show this we take a more straightforward approach and do not model selection explicitly. Instead, we simply split the data into those who refrain from intentional cheating (top panel) and those who intentionally cheat (bottom panel) and run simple univariate OLS regressions of return amounts on beliefs about others' cheating notions. We find that, just as in the main data, for those who refrain from intentionally cheat, return amounts vary essentially one-to-one with *B_ Cheat_ notion* for most send amounts. For those who intentionally cheat, return amounts are consistently much less sensitive to *B_ Cheat_ notion* which is, again, consistent what we find in the on-line data.

Considering the sender's side of the exchange, next we consider how send amounts vary with cheating and monetary return beliefs (Table A10). Because we have few observations and lack the exogenous variation in senders' incentives which we exploited in the analysis of our on-line data, we account for selection into sending a positive amount here by estimating a Tobit model rather than a Heckman model. The results paint a picture qualitatively similar to the on-line data: amounts sent vary positively and significantly with both expected (lack of) cheating (*Pr(NotCheated)*) and expected return (*B_ return_ proportion*).

A.2 Treatments without cheating notion question

We also conducted (on-line) sessions of a treatment in which we dropped our direct cheating notion question and replaced it with a section where participants were asked to indicate how they would feel, as a sender, about various send/return amount scenarios. In total, 170 participants took part in this treatment. As with the main study, ten percent of participants were randomly chosen to be paid their experimental earnings.

To keep the number of individual questions reasonable, we selected three common send amounts—$S = 1, 5$ and 10—and, for each of these, asked participants how they would "feel" if the receiver returned four specific amounts: $0, \frac{S}{2}$, S and $\frac{f(S)}{2}$. These send/return

scenarios were chosen to line up with the cheating notions common in the data from our main study. In terms of feelings, for each send/return amount scenario participants were asked to select exactly two options from a list of several options that best described how they would feel if the scenario were realized. The list of options included positive evaluations ("[the receiver] was generous," "[the receiver] treated me fairly"), neutral evaluations ("[the receiver] was intelligent," "I have no particular opinion of [the receiver's] behavior") and negative evaluations ("[the receiver] cheated me," "[the receiver] disappointed me"). A free-form response option was also available.

To compare the qualitative data we have in this treatment with data from our main sessions, for each send/return scenario investigated in this treatment we calculate the proportion of participants in our main treatment who would feel cheated according to their own reported cheating notions. We compare this proportion to the proportion of respondents in the "feelings" treatment reporting feeling "disappointed" or "cheated." To maximize comparability, from our main treatment data we use only sessions where participants were allowed to opt out of specifying a cheating notion. We find a strong positive relationship between the proportion of participants expressing negative feelings in particular scenarios and the implied proportion of participants feeling cheated in those scenarios in the data from the main treatment (Figure A1). We interpret this as support for the view that trust game participants have well-defined cheating notions and evidence against the view that the cheating notions they report can be mainly attributed to priming.

A.2.1 Evidence on receivers' motivations

In sessions without a direct cheating notion question, at the end of the experiment we added a section in which participants were asked to descibe the rationale they used, if any, for deciding how much to return in the role of receiver. Participants were asked:

> *Describe, in general, how you arrived at your decisions concerning how much to return when you played role B [receiver] for each amount A could have sent you*

Participants could select among four pre-programmed options, or, if none on the list suited them they could select "other" and specify their own rationale. Three of the four pre-programmed responses were meant to capture positive reciprocity, ("the more A [the sender] sent, the more I returned in order to reward nice behavior"); negative reciprocity ("the less A [the sender] sent, the less I returned, in order to punish bad behavior"); vulnerability ("the more A [the sender] sent, the more I returned in order to compensate A [the sender] for being at the mercy of my actions"). The fourth pre-programmed option was essentially a decline to state option ("I did not have any particular rationale in mind.").

Table A11 presents the results. Overall, 83 percent of participants selected one of the four pre-programmed option. The modal response, selected by 42 percent of participants, was that receivers return more when senders send more to compensate senders for their vulnerability. The second most common response reflected positive reciprocity. Almost nobody (6 percent) selected negative reciprocity as their primary rationale, while a similarly low percentage selected the pre-programmed decline to state option (6 percent).

B Robustness checks on beliefs

A common concern whenever beliefs are elicited is the extent to which the elicitation mechanism itself colors reported beliefs. Monetary incentives meant to ensure that participants report beliefs truthfully may give rise to other potential confounds, such as hedging motives: by shading reported beliefs toward bad outcomes, individuals may reduce the variance of their experimental earnings. On the other hand, monetary incentives that are too weak can allow reported beliefs to be non-truthful for various reasons. In particular, one may worry that the significant correlation between *B_Cheat_notion* and receivers' return amounts arises because of a tendency for participants to ex-post rationalize their receiver strategies: by reporting believing that whatever they return is enough to not cheat others, participants can maintain a positive moral self-image.

First we consider ex-post rationalization. If ex-post rationalization is driving beliefs about others' cheating notions (*B_Cheat_notion*), then quadrupling the incentives for belief accuracy in the additional sessions should make this motive less relevant. Evidence of ex-post rationalization would be a consistently smaller correlation between return amounts and *B_Cheat_notion* in the "high belief pay" sessions.

As a simple test for ex-post rationalization, Table A12 (panel A) presents panel regressions of *B_Cheat_notion* as a function of return amounts incorporating a dummy for high belief pay and an interaction with return amounts. The coefficient of interest is on the interaction between high belief pay and return amount: if ex-post rationalization is important when belief pay is low, and diminished for high belief pay, we would expect this coefficient to be consistently negative and significant. Instead, the estimated coefficient on the interaction term is positive and marginally significant providing evidence against ex-post rationalization. Adding our standard set of demographics does not change the results. Moreover, restricting to the subset of observations where the receiver does not intentionally cheat—where the ex-post rationalization argument has the most bite—changes nothing qualitatively. We omit these last two robustness checks to save space, but they are available on request. It should also be noted that variation in belief pay could not have directly affected receivers' actions, since participants did not know there would be a belief elicitation section until after they had submitted their strategies.

Next, consider hedging motives. As a concrete example, consider a sender who has chosen to send 10 euros. If the sender believes the receiver is trustworthy and reports this belief, then in the good state of the world where the receiver *is* trustworthy, the sender earns a lot—both beliefs and actions pay off. However, in the bad state of the world, say, where the receiver returns nothing, the sender loses quite a lot—neither actions nor beliefs pay off. By shading reported beliefs downward—towards a higher likelihood of an untrustworthy sender—the sender can shift some earnings out of the good state of the world into the bad state of the world, reducing earnings variance, i.e., risk.

To test for hedging motives in beliefs, we estimate participants' stated beliefs about the amount of money receivers will return (*B_Receivers_actions*) for each possible send amount. We present panel regressions, where we control for whether a sender actually chose to send a particular amount, risk aversion and an interaction between these two variables. Since hedging motives can only (literally) apply to the send amount a sender actually chooses, one measure of the hedging motive is the coefficient on the dummy for

actually-chosen send amounts. A secondary prediction is that more risk averse individuals care about hedging more, so the interaction term should be negative. Table A12 (panel B) presents our estimates, which provide no support for the importance of hedging. In fact, contrary to hedging motives, reported beliefs about return amounts are marginally significantly *higher* for the amount a sender actually chose to send as evidenced by the coefficient on "Chosen send amount." Risk aversion plays no significant role. Controlling for demographics and/or the level of belief pay does not change anything qualitatively, so we omit these specifications.

C Additional Robustness checks on cheating notions

One additional concern with cheating notions is that they may be (reverse) caused by beliefs. Although priming is not an issue here, as we elicited beliefs after cheating notions, one explanation for the strong correlation between *Cheat_ notion* and *B_ Receivers_ actions* could be that that individuals simply report how much they expect back from receivers as their cheating notion. One reason this could happen is through an individual's desire to maintain a positive self-image and to avoid appearing, to themselves or to the experimenters, as "foolish" for allowing themselves to be cheated. To be clear, if senders expect not to be cheated and hence their cheating notion affects their reported beliefs, that is fine for our purposes. However, if participants first form beliefs about how much receivers will return and then report this belief as their cheating notion because of, e.g., a desire to not appear like a "sucker," then this calls into question the informativeness of the reported cheating notion.

In the latter case, it seems likely that such processes would affect reported cheating notions much more strongly for situations which could *actually* occur—i.e., for the one send amount an individual actually chooses. For concreteness, suppose an individual chooses to send $s = 3$ in the role of sender. Since this is an event that may actually occur, when asked about his or her cheating notion for $s = 3$ an individual may report his or her belief about how much the receiver will return instead of his or her cheating notion in order to avoid looking like a sucker if the event actually occurs. This might be particularly likely if *B_ Receivers_ actions* is less than *Cheat_ notion*. Such a process would tend to inflate reported cheating notions and, at the same time, overstate the correlation between *Cheat_ notion* and *B_ Receivers_ actions*. However, for all other send amounts ($s = 1, 2, 4, \ldots, 10$), since they cannot actually occur, such processes should have little effect on *Cheat_ notion* or its relationship with *B_ Receivers_ actions*.

To test for this effect, we report in Table A13 the results of ten separate regressions—one for each send amount—using *Cheat_ notion* as the dependent variable. On the right hand side, we include an individual's beliefs about the amount the receiver will return (*B_ Receivers_ actions*), a dummy indicating whether the individual chose to send the amount listed in the column heading and an interaction between these two variables. We control for our usual set of demographics, but as they have little explanatory power here we do not report them for ease of exposition.

We find that whether an individual actually chooses a particular send amount has no consistent effect on his or her reported cheating notion: half of the estimated coefficients on

Chosen send amount are positive, half are negative, and only one out of the ten coefficents is significant at conventional levels. Similarly, whether an amount was actually chosen has no consistent effect on the relationship between *B_ Receivers_ actions* and *Cheat_ notion*: five of the ten coefficients on the interaction between *B_ Receivers_ actions* and *Cheat_ notion* are positive, the other five are negative and only one out of the ten is statistically significant. Considered together, our results provide little evidence for cheating notions being reverse-caused by beliefs because, e.g., participants want to avoid looking like a sucker.

D Cheating notions and guilt aversion theory

In this section we test for the conjectured correlations between: i) *Cheat_ notion* and beliefs about receivers' actions (*B_ Receivers_ actions*); and ii) beliefs about others' cheating notions (*B_ Cheat_ notion*) and second-order beliefs (*B_ B_ Receivers_ actions*).

In Table A14 we report ten separate regressions—one for each send amount—using *B_ Receivers_ actions* as the dependent variable and, as the main explanatory variable, an individual's own personal cheating notion (*Cheat_ notion*). We control for available demographics and relevant experimental design features. In this latter category, we include a dummy for whether there was a sending fee in the session as this might factor into a sender's definition of return on investment. As a simple check on whether the reported beliefs are true beliefs, or rather whether the relationship between beliefs and cheating notions is driven by nuisance factors (e.g., ex-post rationalization), we include a dummy indicating sessions where we *quadrupled* belief elicitation incentives as well as an interaction term between this dummy and own cheating notions. The main lesson from this exercise is that one's own cheating notion is consistently a highly significant predictor of senders' first-order beliefs (*B_ Receivers_ actions*). The strength of the relationship is large in magnitude as well: a one euro increase in *Cheat_ notion* translates into a roughly 50 cent increase in *B_ Receivers_ actions*. Examining the coefficient on the interaction between cheating notions and belief elicitation incentives, we find that much stronger incentives have no consistent impact on this relationship and that, moreover, the impact is almost never significant. These patterns suggest that reported beliefs are true beliefs. Finally, it is worth noting that demographics have little explanatory power with one exception: gender. Male participants consistently expect about 40 to 50 cents less back from receivers than female participants.

In Table A15, we estimate receiver's second-order beliefs (*B_ B_ Receivers_ actions*) as a function of their beliefs about others' cheating notions (*B_ Cheat_ notion*). As before, we control for available demographics, relevant experimental design features, beliefs incentives and an interaction between beliefs incentives and reported beliefs about others' cheating notions. We find that beliefs about others' cheating notions are always highly significant predictors of second-order beliefs and that this relationship is also large in magnitude: a one-euro increase in *B_ Cheat_ notion* translates into a 34 to 83 cent increase in second-order beliefs with an average increase, over all ten send amounts, of about 60 cents. Strengthened belief incentives, again, have no consistent impact on this relationship and, moreover, their effect is almost never significant at conventional levels. Demographics play a slightly larger role here: being male or having more mathematical ability tends to lower second-order

beliefs; being older tends to raise them. The main lesson from Table 6, however, is that beliefs about others' cheating notions exhibit a strong positive relationship with second-order beliefs.

Table A1: Comparison of behavior in the lab and on-line, summary statistics

	Send > 0	Send amount	Return proportion	B_return_proportion	Proportion of non-cheaters	Pr(NotCheated)
			In-lab sessions			
	0.97	5.28	1.25	1.36	0.43	0.56
	(0.03)	(0.45)	(0.10)	(0.10)	(0.06)	(0.03)
Obs	36	36	36	36	36	36
			On-line low fee sessions			
	0.90	4.83	1.28	1.22	0.53	0.53
	(0.03)	(0.26)	(0.06)	(0.06)	(0.03)	(0.02)
Obs	150	150	149	148	150	135

Table A2: Proportion of participants with a cheating notion (Cheat_notion), in-lab sessions

	Send amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ cheating notion	0.72	0.86	0.83	0.92	0.94	0.94	0.97	0.97	0.97	0.97
	(0.08)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Obs	36	36	36	36	36	36	36	36	36	36

Notes: [1] Raw proportions reported. [2] Standard errors appear in parentheses

Table A3: Proportion of participants who would feel cheated by (return amount) < (send amount), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ (cheating notion) ≥ (send amt)	0.88	0.84	0.97	0.94	0.94	0.97	0.89	0.83	0.83	0.86
	(0.06)	(0.07)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.06)	(0.06)	(0.06)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. [2] Standard errors appear in parentheses

Table A4: Proportion of participants for whom Cheat_notion is consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Weakly positive return on investment	0.08 (0.05)	0.10 (0.05)	0.30 (0.09)	0.18 (0.07)	0.18 (0.07)	0.12 (0.06)	0.20 (0.07)	0.14 (0.06)	0.11 (0.05)	0.23 (0.07)
Strictly positive return on investment	0.15 (0.07)	0.16 (0.07)	0.50 (0.09)	0.39 (0.08)	0.29 (0.08)	0.32 (0.08)	0.29 (0.08)	0.29 (0.08)	0.23 (0.07)	0.31 (0.08)
Inequality Aversion	0 --	0.03 (0.03)	0 --	0.03 (0.03)	0.06 (0.04)	0.03 (0.03)	0 --	0.03 (0.03)	0.23 (0.07)	0.37 (0.08)
Equal split	0.35 (0.10)	0.23 (0.08)	0.20 (0.07)	0.18 (0.07)	0.32 (0.08)	0.32 (0.08)	0.34 (0.08)	0.23 (0.07)	0.23 (0.07)	0.37 (0.08)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. Classifications are not mutually exclusive so that, e.g., the same cheating notion can be labeled as consistent with both SPROI and Inequality aversion. [2] Standard errors are in parentheses. [3] A weakly positive return on investment (WPROI) cheating notion entails reporting exactly the send amount (s) as one's cheating threshold in sessions without a sending fee. [4] "SPROI" (strictly positive return on investment) is a more generous definition of WPROI taking into account a reasonable interest rate, r = 10%. We multiply the send amount by 1+r to get an "exact SPROI" definition. To be as generous as possible to this notion, and to account for the fact that experimental participants typically have a well-known predilection to state whole-number values, we then calculate the least integer greater than this exact value, denoted by ceiling("exact SPROI"). For each send amount, *s*, We label as SPROI all cheating thresholds falling within the interval with integer end-points: [s, ceiling("exact SPROI")]. [5] "Inequality Aversion" refers to a cheating notion which requires equal monetary outcomes, and we label a cheating notion as consistent with inequality aversion if it lies within the smallest closed interval with integer endpoints containing this outcome. As an example, consider s = 1. The total surplus in this case is 10.50 – 1 + 8.05 = 17.55, and half of this surplus is 8.775. Any cheating notion in the interval [8, 9] would therefore be labeled as consistent with inequality aversion. [6] An "Equal-split" (ES) cheating notion entails a cheating threshold of half of the entire amount allocated to the receiver. As with SPROI and Inequality Aversion above, to account for participants' predilection for whole numbers, the definition of ES for each send amount, s, includes all cheating thresholds falling within the smallest interval with whole-number end-points containing a precisely-equal split of the receivers' total earnings: i.e., $\frac{f(s)}{2} \in$ [n, n+1]. For example, if a sender sends s = 3, a receiver receives f(s) = 11.30, and $\frac{f(s)}{2}$ = 5.65. Consequently, ES for s = 3 would include all cheating thresholds within the interval [5, 6].

Table A5: Proportion of participants whose beliefs about others' cheating notions (B_Cheat_notion) are consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Weakly positive return on investment	0.28	0.22	0.17	0.17	0.25	0.17	0.14	0.17	0.14	0.19
	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
Strictly positive return on investment	0.39	0.36	0.39	0.28	0.31	0.28	0.28	0.28	0.22	0.33
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)
Inequality Aversion	0	0	0	0	0.06	0.08	0	0.14	0.22	0.31
	--	--	--	--	(0.04)	(0.05)	--	(0.06)	(0.07)	(0.08)
Equal split	0.25	0.31	0.33	0.14	0.33	0.36	0.36	0.25	0.22	0.31
	(0.07)	(0.08)	(0.08)	(0.06)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)
Obs	36	36	36	36	36	36	36	36	36	36

Table A6: Beliefs about the amount receivers will return (B_Receivers_actions) as a function of own cheating notions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	0.75***	0.75***	0.79***	0.53***	0.57***	0.51***	0.62***	0.84***	0.53***	0.64***
	(0.10)	(0.10)	(0.08)	(0.11)	(0.09)	(0.11)	(0.19)	(0.24)	(0.18)	(0.18)
Constant	0.17	0.48	0.74	2.01**	2.21***	2.74**	2.47	0.99	3.94*	3.04
	(0.30)	(0.46)	(0.50)	(0.78)	(0.78)	(1.04)	(1.79)	(2.26)	(2.02)	(2.13)
Observations	26	31	30	33	34	34	35	35	35	35
R-squared	0.72	0.65	0.76	0.42	0.53	0.42	0.24	0.28	0.21	0.27

Table A7: Beliefs about senders' beliefs about amount receivers will return (B_B_Receivers_actions), as a function of beliefs about others' cheating notions (B_Cheat_notion), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.62***	0.61***	0.62***	0.57**	0.59***	0.63***	0.68***	0.64***	0.65***	0.67***
	(0.18)	(0.20)	(0.21)	(0.24)	(0.21)	(0.20)	(0.20)	(0.17)	(0.17)	(0.17)
Constant	0.85	1.40*	1.67	2.42	2.89*	3.12*	2.81	3.59*	3.64*	3.80*
	(0.54)	(0.81)	(1.12)	(1.51)	(1.50)	(1.66)	(1.83)	(1.80)	(1.95)	(2.09)
Observations	36	36	36	36	36	36	36	36	36	36
R-squared	0.25	0.22	0.20	0.14	0.19	0.23	0.26	0.28	0.29	0.30

Table A8: Intentional cheating (reduced form), in-lab sessions

	Sent Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	-0.78*	-0.35**	-0.08	-0.05	0.02	-0.08	-0.05	-0.02	-0.25**	-0.13*
	(0.43)	(0.17)	(0.12)	(0.11)	(0.07)	(0.09)	(0.11)	(0.11)	(0.12)	(0.07)
B_Cheat_notion	1.08**	0.32	0.15	0.16	0.18	0.05	0.25**	0.11	0.28*	0.34***
	(0.50)	(0.20)	(0.17)	(0.14)	(0.13)	(0.13)	(0.11)	(0.13)	(0.16)	(0.11)
Constant	-0.68	0.27	-0.36	-0.73	-1.13	0.30	-1.49	-0.41	-0.11	-1.98*
	(0.64)	(0.62)	(0.74)	(0.87)	(0.94)	(0.85)	(1.09)	(1.06)	(1.01)	(1.11)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Each column presents estimates from a Probit model, with the (binary) dependent variable being "receiver intentionally cheats if sent relevant amount." Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated, i.e., by the event r < B_Cheat_notion. This threshold amount is also inserted as a control in each estimate by the variable "B_Cheat_notion." [3] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%.

Table A9: Intentional cheating (reduced form), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (r ≥ B_Cheat_notion)										
B_Cheat_notion	1.24***	1.09***	0.90***	1.08***	1.02**	1.27***	0.44	1.00***	0.93***	0.76
	(0.32)	(0.32)	(0.26)	(0.18)	(0.34)	(0.32)	(0.60)	(0.27)	(0.24)	(0.48)
Constant	0.30	0.51	1.37	1.05	1.36	-0.42	6.27	1.95	2.12	4.74
	(0.77)	(1.16)	(1.32)	(1.07)	(2.28)	(2.61)	(4.90)	(2.58)	(2.53)	(5.05)
Obs	15	16	17	19	15	18	14	11	15	14
R-squared	0.53	0.46	0.44	0.68	0.40	0.50	0.04	0.60	0.55	0.17
Conditional on cheating (r < B_Cheat_notion)										
B_Cheat_notion	0.44**	0.43**	0.20	0.37	0.45	0.25	0.44*	0.71***	0.53**	0.40
	(0.19)	(0.18)	(0.21)	(0.22)	(0.28)	(0.22)	(0.21)	(0.21)	(0.23)	(0.28)
Constant	-0.35	0.10	1.84	1.47	0.85	2.84	1.96	-0.19	1.61	2.63
	(0.61)	(0.81)	(1.13)	(1.42)	(2.16)	(1.86)	(2.13)	(2.28)	(2.70)	(3.53)
Obs	21	20	19	17	21	18	22	25	21	22
	0.22	0.24	0.05	0.16	0.12	0.07	0.17	0.32	0.22	0.09

Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a simple OLS regression of return amount conditional on beliefs about others' cheating notion for the send amount listed in the column heading. [3] The top panel is restricted to observations not involving intentional cheating, while the bottom panel is restricted to observations involving intentional cheating.

Table A10: Send amount (Tobit), in-lab sessions

	Dependent variable = send amount		
	(1)	(2)	(3)
Pr(NotCheated)	4.29*	4.94**	6.97***
	(2.19)	(2.25)	(1.90)
B_return_proportion	1.54*	1.57**	1.41*
	(0.81)	(0.75)	(0.77)
Male		1.53*	0.93
		(0.81)	(0.75)
Age		-0.16*	-0.30**
		(0.09)	(0.11)
Math score		-0.34	0.07
		(0.42)	(0.37)
Risk aversion			-0.47**
			(0.17)
Altruism			0.04
			(0.21)
30≤ Income <45			-1.48
			(0.98)
45≤ Income <70			0.03
			(1.10)
45≤ Income <70			1.76
			(1.47)
Income ≥120			-2.99**
			(1.26)
Constant	0.86	5.85	8.66*
	(1.52)	(4.60)	(5.01)
Obs	36	34	32

Notes: [1] Robust standard errors in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] Each column presents a Tobit model estimate where the dependent variable is *how much* the sender sends and censoring below 0 is taken into account. [5] "Pr(NotCheated)" is our measure of participants' subjective beliefs about not being cheated, described in the text. [6] "B_return_proportion" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send amounts. [7] "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse). [8] Altruism is how much emphasis participants' parents placed on the value "help others" during their upbringing. [9] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table A11: Proportion of receivers specifying a particular rationale

	Overall	High fee sessions	Low fee sessions
Sender vulnerability	0.42	0.40	0.45
	(0.04)	(0.05)	(0.06)
Positive reciprocity	0.29	0.31	0.27
	(0.04)	(0.05)	(0.05)
Negative reciprocity	0.06	0.06	0.05
	(0.02)	(0.03)	(0.03)
No motive	0.06	0.05	0.08
	(0.02)	(0.02)	(0.03)
Obs	170	93	77

Notes: [1] Raw proportions reported; [2] Standard errors in parentheses; [3] Proportions in each column sum to less than one, with the unaccounted for observations being participants who elected to supply their own rationale rather than one of the four pre-programmed rationale; these self-supplied rationale varied widely and are not easily classifiable.

Table A12: Robustness checks on beliefs, main study data

Panel A: checking for ex-post rationalization

Dependent variable = *B_Cheat_notion*

Return amount	Amount sent	High belief pay	(High belief pay) X (Return amt)	Cons	Obs	Individuals	R^2
0.11***	0.85***	0.12	0.05*	1.58***	4254	428	0.5
(0.02)	(0.03)	(0.24)	(0.03)	(0.20)			

Panel B: checking for hedging motives in beliefs

Dependent variable = *B_Receivers_actions*

Amount sent	Chosen send amount	Risk aversion	(Chosen send amt) X (Risk aversion)	Cons	Obs	Individuals	R^2
0.82***	0.29*	-0.00	-0.02	1.61***	4146	417	0.34
(0.02)	(0.17)	(0.04)	(0.03)	(0.23)			

Notes: [1] Both the top and bottom panel report individual random effects regressions pooling observations across all send amounts. [2] Robust standard errors, clustered by session, appear in parentheses. [3] "High belief pay" is a dummy taking the value of one if the session involved a 20 euro maximum belief pay, and 0 if the maximum possible belief pay was 5 euros; "Chosen send amount" is a dummy variable indicating the amount a participant actually chose to send in the role of sender; "Risk aversion" is an incentive-compatible index of risk aversion obtained from a previous experiment. [4] We drop observations for which we have no measure of risk aversion.

Table A13: Robustness check on own cheating notion, main study data

Dependent variable = *Cheat_notion*										
	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Receivers_actions	0.66[A]	0.66[A]	0.68[A]	0.58[A]	0.48[A]	0.54[A]	0.57[A]	0.50[A]	0.54[A]	0.52[A]
	(0.06)	(0.09)	(0.08)	(0.10)	(0.08)	(0.11)	(0.07)	(0.09)	(0.08)	(0.07)
Chosen send amount	-0.18	0.42	-0.30	1.30	-2.22[B]	-1.16	0.73	1.02	-1.62	1.04
	(0.75)	(0.79)	(0.88)	(1.07)	(0.69)	(1.21)	(0.97)	(2.46)	(1.52)	(2.29)
Chosen send amount X B_Receivers_actions	0.11	-0.32	0.18	-0.29	0.31[B]	0.07	0.00	-0.02	-0.03	-0.08
	(0.29)	(0.32)	(0.25)	(0.23)	(0.09)	(0.18)	(0.08)	(0.35)	(0.11)	(0.18)
Low Fee	-0.08	0.00	-0.11	-0.26	-0.10	-0.06	0.41	0.23	0.43[C]	0.05
	(0.15)	(0.18)	(0.11)	(0.20)	(0.24)	(0.13)	(0.27)	(0.27)	(0.22)	(0.26)
Constant	1.99C	2.70[B]	2.19[C]	3.89[B]	6.36[B]	4.92[B]	2.55	6.74[B]	4.92[C]	5.68[B]
	(0.96)	(0.91)	(0.93)	(1.48)	(2.09)	(1.81)	(2.01)	(2.17)	(2.23)	(1.97)
Demographic controls?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.33	0.39	0.30	0.25	0.30	0.32	0.25	0.31	0.29

Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' personal cheating notions (*Cheat_notion*). [2] Robust standard errors, clustered by session, appear in parentheses. [3] Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "B_Receivers_actions" is a participant's belief about how much a receiver will return for the send amount indicated in the column heading; "Chosen send amount" is a dummy variable indicating the participant actually chose to send the amount in the column heading in the role of sender. [5] Demographic controls are included but not reported for readability. The set of demographic controls is identical to the set reported in Table 6 in the manuscript. "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros. [6] Observations vary over columns because we do not have demographics for all participants and because not all participants reported a cheating notion for all send amounts.

Table A14: Beliefs about the amount receivers will return as a function of own cheating notions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cheat_notion	0.61^{A}	0.58^{A}	0.52^{A}	0.46^{A}	0.36^{A}	0.46^{A}	0.57^{A}	0.50^{A}	0.51^{A}	0.52^{A}
	(0.06)	(0.02)	(0.02)	(0.06)	(0.06)	(0.04)	(0.03)	(0.10)	(0.10)	(0.10)
Male	-0.30^{C}	-0.49^{B}	-0.34^{C}	-0.53^{A}	-0.43^{A}	-0.37^{C}	-0.23	-0.22	-0.49	-0.28
	(0.15)	(0.17)	(0.17)	(0.13)	(0.10)	(0.19)	(0.22)	(0.32)	(0.31)	(0.25)
Age	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.02	0.04	-0.00	-0.00
	(0.01)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)
Math score	-0.04	-0.08^{C}	-0.06	0.01	0.10	0.02	0.01	0.05	-0.01	-0.04
	(0.05)	(0.03)	(0.05)	(0.07)	(0.09)	(0.10)	(0.07)	(0.15)	(0.13)	(0.17)
Risk aversion	0.03	0.02	0.04	0.02	0.02	0.07	0.01	0.11	0.06	0.17
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	(0.09)	(0.10)	(0.11)	(0.10)	(0.15)
30≤ Inc <45	0.18	0.47^{B}	0.34	0.62^{B}	0.12	0.07	-0.06	-0.31	-0.08	-0.17
	(0.19)	(0.15)	(0.24)	(0.25)	(0.30)	(0.37)	(0.32)	(0.48)	(0.54)	(0.66)
45≤ Inc<70	0.24	0.31	0.32	0.57^{B}	0.29	0.38^{C}	-0.04	-0.25	-0.15	-0.19
	(0.13)	(0.25)	(0.29)	(0.24)	(0.26)	(0.19)	(0.43)	(0.32)	(0.34)	(0.37)
70≤ Inc <120	-0.03	0.17	0.39	0.66^{B}	0.34	0.52	-0.14	0.06	0.01	-0.26
	(0.19)	(0.18)	(0.25)	(0.27)	(0.31)	(0.46)	(0.68)	(0.68)	(0.74)	(0.91)
Inc ≥120	0.26	0.19	0.14	0.39	0.01	-0.36	-0.08	-0.18	-0.42	-0.83
	(0.26)	(0.21)	(0.21)	(0.30)	(0.38)	(0.53)	(0.67)	(0.59)	(0.64)	(0.85)
Low Fee	-0.13	-0.23	-0.30^{B}	-0.14	-0.20	-0.30^{B}	-0.52^{B}	-0.48^{C}	-0.61^{B}	-0.25
	(0.11)	(0.21)	(0.12)	(0.22)	(0.14)	(0.12)	(0.20)	(0.24)	(0.23)	(0.26)
High belief Incentives	0.22	0.71^{A}	0.21	0.09	-0.49	-0.35	0.62	0.08	-0.41	-0.71
	(0.15)	(0.16)	(0.32)	(0.74)	(0.81)	(0.87)	(0.46)	(1.31)	(1.31)	(1.48)
Own cheating notion X High belief Incentives	-0.12	-0.16^{B}	0.00	0.00	0.08	0.05	-0.06	-0.02	0.04	0.05
	(0.06)	(0.06)	(0.05)	(0.11)	(0.09)	(0.09)	(0.06)	(0.12)	(0.12)	(0.11)
Constant	1.00	1.27	1.90^{C}	2.45^{C}	2.24^{C}	2.96^{C}	3.00^{B}	1.70	3.72^{C}	3.44
	(0.81)	(0.74)	(0.84)	(1.05)	(1.06)	(1.30)	(1.16)	(1.36)	(1.71)	(1.84)
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.34	0.38	0.28	0.23	0.28	0.31	0.25	0.30	0.29

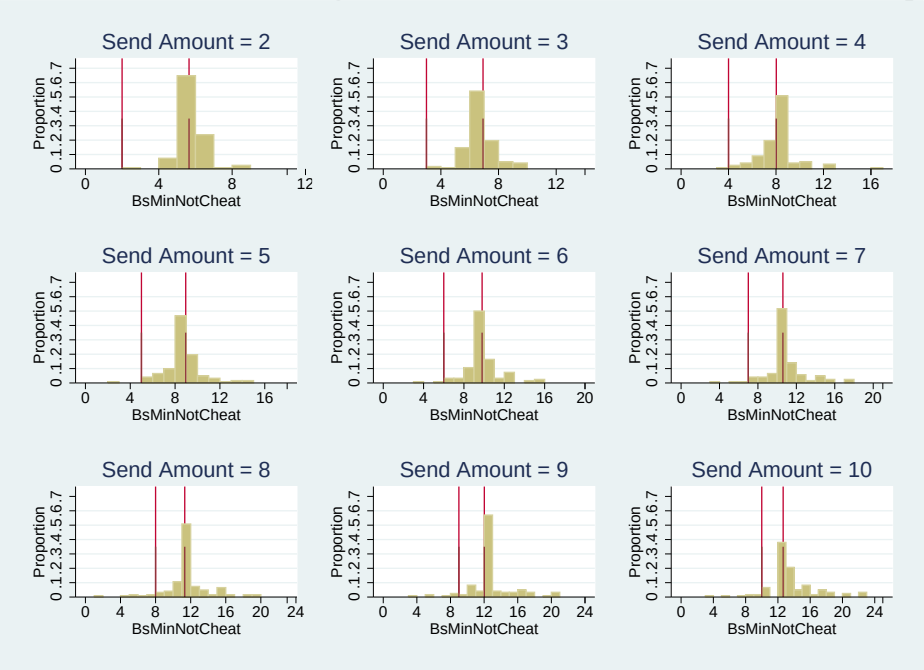
Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' beliefs about the amount receivers will return (*B_Receivers_actions*). [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable is a participant's own cheating notion. Additional demographic controls include: "Math score" = self-reported score on required math exams taken during the final year of high school in Italy; "Risk aversion" = an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism from a prior, unrelated, experiment, which takes values from 1 (risk loving) to 10 (very risk averse); "Inc" = self-reported annual family income from all sources, in thousands of euros, net of taxes. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because not all participants reported a cheating notion for every send amount and because we do not have demographics for all participants. [7] The coefficients and significance levels on the main explanatory variable, "Own cheating notion," are virtually identical if demographics are omitted. From s = 1, …, 10, the coefficients and significance levels are: 0.59^{A}, 0.59^{A}, 0.54^{A}, 0.46^{A}, 0.37^{A}, 0.45^{A}, 0.58^{A}, 0.49^{A}, 0.50^{A}, 0.51^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 2.

Table A15: Second-order beliefs (B_B_Receivers_actions) as a function of beliefs about others' cheating notions (B_Cheat_notion)

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B_Cheat_notion	0.83^{A}	0.66^{A}	0.69^{B}	0.84^{A}	0.58^{A}	0.65^{A}	0.51^{B}	0.34^{A}	0.46^{A}	0.45^{A}
	(0.12)	(0.14)	(0.21)	(0.07)	(0.12)	(0.08)	(0.18)	(0.08)	(0.08)	(0.06)
Male	-0.26^{B}	-0.54^{A}	-0.57^{A}	-0.55^{B}	-0.56^{B}	-0.64^{A}	-0.77^{A}	-0.73^{A}	-0.92^{B}	-0.96^{B}
	(0.09)	(0.08)	(0.12)	(0.16)	(0.18)	(0.13)	(0.17)	(0.16)	(0.37)	(0.31)
Age	0.05^{C}	0.05^{B}	0.07^{B}	0.07^{B}	0.07^{B}	0.08^{B}	0.10^{C}	0.09^{C}	0.08	0.06
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)
Math score	-0.10^{B}	-0.13^{B}	-0.09^{C}	-0.18^{C}	-0.06	-0.03	-0.11	-0.19C	-0.07	-0.04
	(0.04)	(0.04)	(0.04)	(0.08)	(0.07)	(0.08)	(0.10)	(0.08)	(0.13)	(0.12)
Risk aversion	-0.01	-0.02	-0.00	-0.00	-0.02	-0.03	-0.04	0.03	-0.00	0.03
	(0.02)	(0.03)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.08)
30≤ Inc <45	-0.28	-0.36^{B}	-0.16	-0.39^{C}	-0.11	-0.28	-0.19	-0.14	-0.14	-0.61
	(0.17)	(0.14)	(0.17)	(0.20)	(0.20)	(0.15)	(0.32)	(0.30)	(0.36)	(0.33)
45≤ Inc<70	0.06	0.04	0.26	0.35	0.52C	0.31	0.23	0.25	0.31	0.16
	(0.09)	(0.29)	(0.26)	(0.26)	(0.26)	(0.37)	(0.23)	(0.29)	(0.38)	(0.46)
70≤ Inc <120	-0.15	-0.30	-0.04	-0.02	0.05	0.08	0.02	-0.14	0.14	0.09
	(0.09)	(0.31)	(0.25)	(0.31)	(0.29)	(0.37)	(0.33)	(0.45)	(0.48)	(0.65)
Inc ≥120	-0.17	-0.65^{C}	-0.83	-0.72	-0.81	-0.57	-0.45	-1.08	-0.55	-0.64
	(0.18)	(0.30)	(0.62)	(0.62)	(0.82)	(0.89)	(0.87)	(0.93)	(0.99)	(0.99)
Low Fee	0.03	-0.14	-0.24	-0.23	-0.18	0.02	-0.14	-0.21	-0.14	-0.24
	(0.07)	(0.16)	(0.20)	(0.19)	(0.23)	(0.09)	(0.21)	(0.25)	(0.28)	(0.31)
High Belief Incentives	0.21	-0.20	-0.14	0.43	-0.88	-0.45	-1.57	-4.01^{A}	-2.53	-2.54^{C}
	(0.46)	(0.58)	(1.08)	(0.54)	(1.06)	(0.78)	(1.73)	(0.95)	(1.37)	(1.12)
Est. others' cheating notion X High Belief Incentives	-0.20	0.00	-0.07	-0.20^{C}	0.05	-0.02	0.11	0.34^{A}	0.19	0.19^{C}
	(0.14)	(0.14)	(0.21)	(0.09)	(0.15)	(0.09)	(0.19)	(0.09)	(0.11)	(0.08)
Constant	0.59	1.67	1.28	1.63	2.35	1.76	3.71	6.34^{A}	4.91^{C}	5.57^{B}
	(0.86)	(1.15)	(1.66)	(1.45)	(1.78)	(1.71)	(2.55)	(1.68)	(2.11)	(1.99)
Observations	375	375	375	375	375	375	375	375	375	375
R-squared	0.45	0.40	0.34	0.39	0.33	0.34	0.32	0.32	0.32	0.34

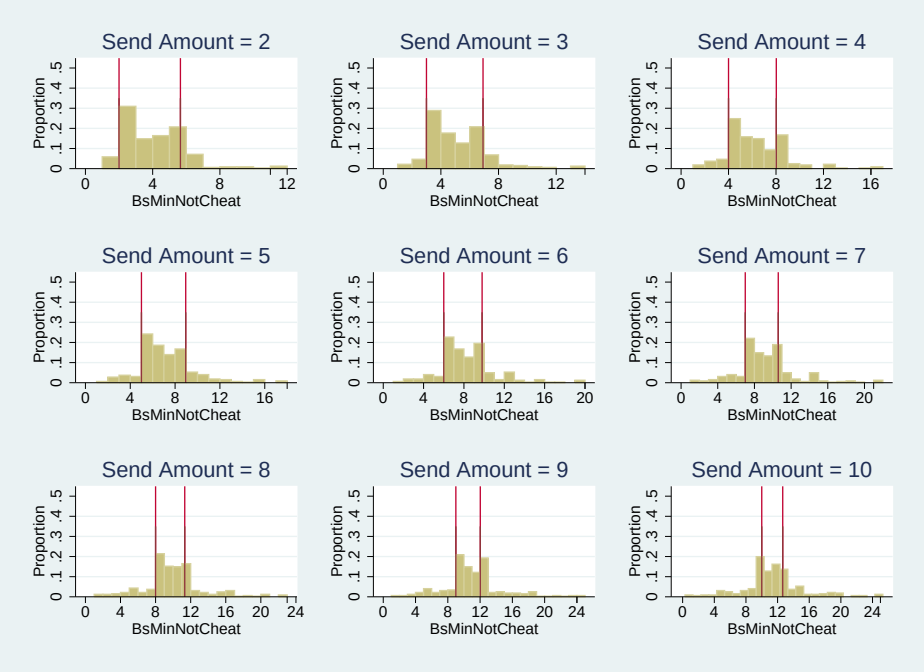
Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' second-order beliefs *B_B_Receivers_actions*. [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "B_Cheat_notion" is a participant's belief about others' cheating notions. Other demographic controls are identical to those in Table 6, above. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because we do not have demographics for all participants. [7] If demographics are omitted, the coefficients and significance levels on the main explanatory variable, "B_Cheat_notion," are virtually identical. From s = 1, …, 10, the coefficients and significance levels are: 0.84^{A}, 0.69^{A}, 0.73^{A}, 0.83^{A}, 0.63^{A}, 0.68^{A}, 0.55^{A}, 0.40^{A}, 0.49^{A}, 0.48^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 8.

Figure A1a: Individual-level Consistency of B_Cheat_notion across Send Amounts, equal split



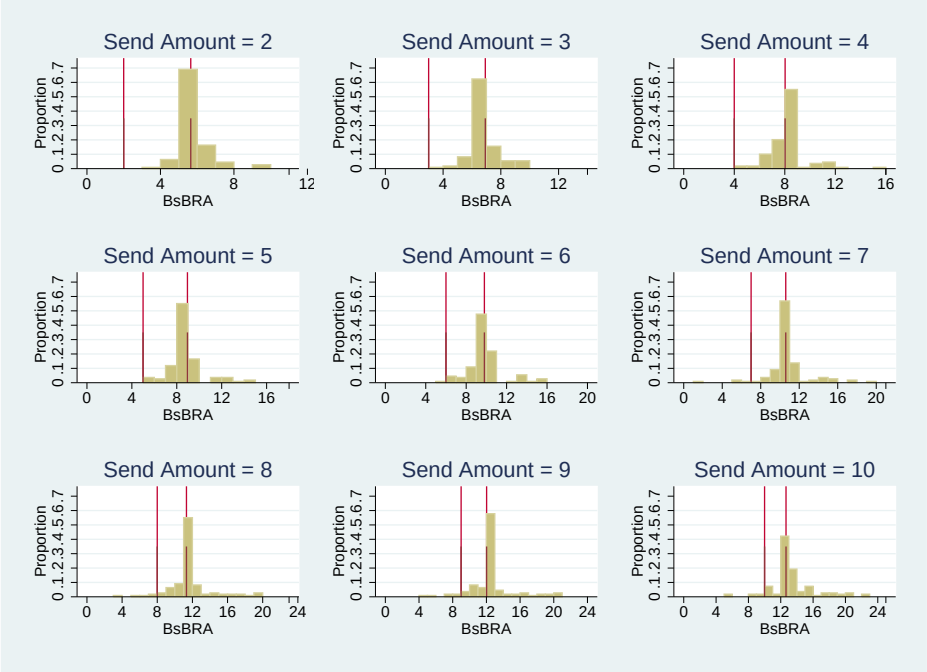
Notes: [1] The figure restricts attention to participants whose beliefs about others' cheating notions (*B_Cheat_notion*) were consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' beliefs about others' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A1b: Individual-level Consistency of B_Cheat_notion across Send Amounts, strictly positive return on investment



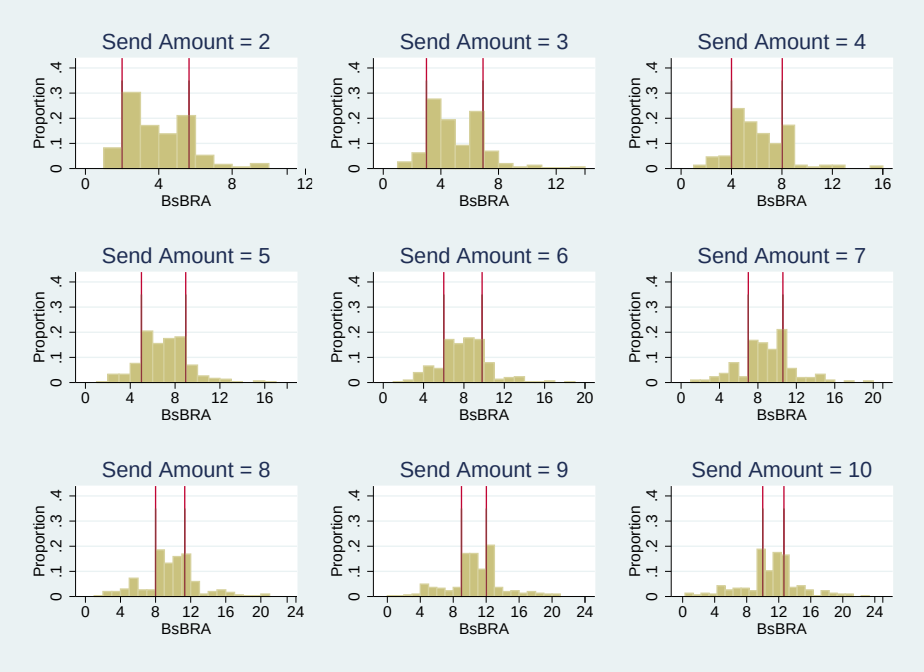
Notes: [1] The figure restricts attention to participants whose beliefs about others' cheating notions (*B_Cheat_notion*) were consistent with strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' beliefs about others' cheating notions for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A2a: Individual-level Consistency of B_B_Receivers_actions across Send Amounts, equal split



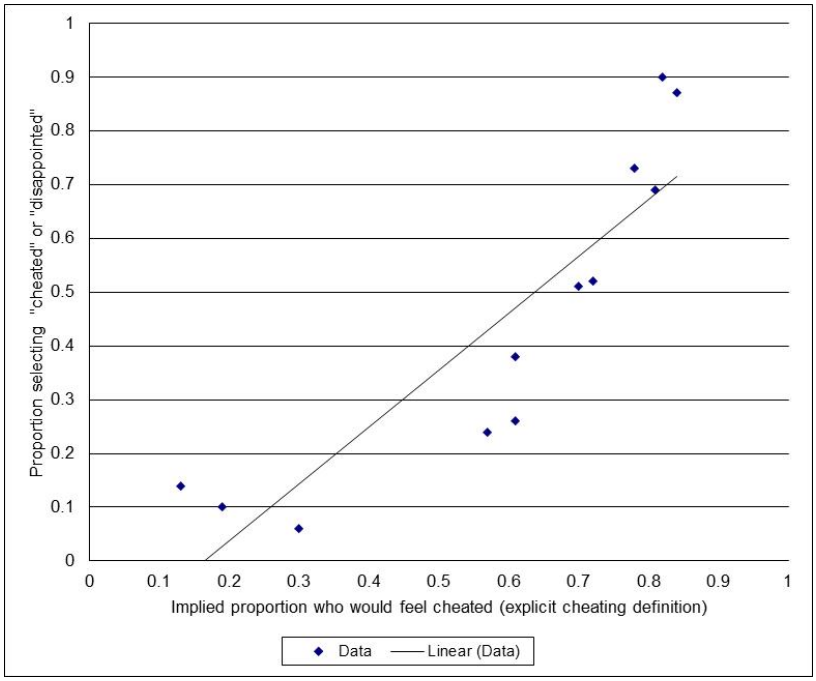
Notes: [1] The figure restricts attention to participants whose second-order belief (*B_B_Receivers_actions*) was consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' *B_B_Receivers_actions* for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A2b: Individual-level Consistency of B_B_Receivers_actions across Send Amounts, strictly positive return on investment



Notes: [1] The figure restricts attention to participants whose second-order belief (*B_B_Receivers_actions*) was consistent with a strictly positive return on investment conditional on a send amount of 1, and presents histograms of these participants' *B_B_Receivers_actions* for all other send amounts. [2] Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

Figure A3: Comparison of proportion feeling cheated by elicitation method



Appendix II: Experiment Instructions

In this experiment, you will be randomly paired with another participant and assigned randomly one of two roles: A or B. This pairing will be anonymous. Neither the person in the role of A nor the person in the role of B wil know with whom they have been paired.

The role of A

The player in the role of A is given 10.50 euros and must decide whether to send some all or none of this money to the player in the role of B, the person with whom A has been paired. [If A decides to send some of the this money, A will be charged a fee of 0.50 euros.] For every euro that A sends, B will receive more than 1 euro according to the table below.

If A sends €	1	2	3	4	5	6	7	8	9	10
B receives €	8.05	11.3	13.85	16.05	17.9	19.6	21.2	22.65	24.05	25.3

The role of B

After A makes his or her decision about how much to send to B, B decides how much of the money he or she receives—the amounts in the table above (8.05 euros, 11.30 euros, etc.)—to return to A. The player in the role of B will specify an amount to return for each possible amount they could receive. For example, if A sends 4 euros and B therefore receives 16.05 euros, B must decide how much of this 16.05 euros to return to A; and a decision must be made for every amount A could send (1,2,3,. . . ,10 euros).

Your earnings

For every pair of participants, one in the role of A and one in the role of B, the decisions that both A and B make determine the pairs earnings. Both A and B will be informed of the outcome determined by their choice.

In general:

- If A sends a positive amount to B:
    1. A's earnings will be: € 10.50 – (euros sent to B) + (euros returned by B) – (€ 0.50 fee)
    2. B's earnings will be: (euros received by B according to the table above) – (euros returned to A)
- If A sends nothing to B:
    1. A's earnings will be € 10.50
    2. B's earnings will be € 0.

Specifically, for every pair of players the result of this situation will be determined as follows:

i Every participant specifies their decision for each possible role (A and B).

ii The computer will randomly assign a role to each participant and randomly and anonymously pair each participant assigned the role of A with a participant assigned the role of B.

iii Within each pair, A's decisions will be combined with B's decision to determine the outcome for both A and B.

A Experiment Screens

A.1 Sender decision screen 1

If you are assigned the role of A, do you want to send money to B? If you send money, you will be charged a € 0.50 fee.

Choose "send" or "don't send" on this screen. If you choose "send", you will specify the amount to send on the next screen.

__ Send money
__Don't send money

A.2 Sender decision screen 2

How much money do you want to send if you are assigned the role of A?

__ € 1
__ € 2
...
__ € 10

A.3 Receiver decision screens

[There are 10 separate screens. A representative question is below.]

Imagine that you have been assigned the role of B ...
How much will you send back to A if A sends € 7 and you therefore receive € 21.20?

A.4 Cheating definition screen

If you are assigned the role of A, what is the minimum amount you would need to receive back from B in order to not feel cheated?

If you send €1 and therefore B receives €8.05, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating

__ I do not know

...
If you send €10 and therefore B receives €25.30, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating
__ I do not know

A.5 Belief elicitation

A.5.1 Instructions, screen 1

Now, we begin a new section. In this section as in the previous section, each question can contribute to your potential earnings.

Specifically, in this section you will be asked to estimate the choices other participants made in the previous section. Every question is about the choices of other participants, so please exclude your own actions from your estimations. The accuracy of your estimates will be calculated excluding your own actions as well.

Your earnings from this section will be determined by choosing one of your estimations at random and paying you according to the accuracy of this randomly chosen estimation. Every estimate has the same chance of being chosen by the computer. Your potential earnings from this experiment will be the sum of your earnings in this section and in the previous section.

The formula used to calculate your earnings from the randomly-chosen estimate is detailed on the next page.

A.5.2 Belief compensation formula screen

The method used to calculate your earnings from your estimates is detailed below. The most important thing to notice is that more accurate estimates have higher chances of earning money.

- Your estimate, R, is inserted into the following formula where "r" stands for the true value of the thing being estimated and "r_{max}" is the maximum value this true value can attain.

$1 - \left(\frac{R-r}{r_{\max}}\right)$

- This produces a number between 0 and 1. Call this number "z".
- The computer chooses a number between 0 and 1 with each number in between 0 and 1 being equally likely. Call this number "y".

- If $y \leq z$, you will earn €5.00 [€20.00] for your estimate.
- If $y > z$, you will earn €0.00 for your estimate.

An example

Suppose you are asked to estimate the average amount participants in the role of A send in the previous section of this experiment. And, imagine that this average turns out to actually be €4.00. The maximum value this average could have taken is €10. Therefore "r_{max}" in the equation above is 10 and r is 4. The equation therefore becomes:
$1 - \left(\frac{R-4}{10}\right)$
Notice that the closer your estimate, R, is to the actual value of 4 in our hypothetical example, the larger is z and therefore the larger is the probability of earning €5 [€20.00] for your estimate rather than €0.

- If your estimate is exactly correct, then (R-4)/10 = 0 and therefore z=1. Because the number chosen by the computer is at most one, an exactly correct estimate always pays €5 [€20.00].
- On the other hand, the probability with which your estimate earns you €5 [€20.00] diminishes the farther away from the true value your estimate is: z becomes smaller and so does the chances that $y < z$.

Click continue to begin start the estimation section

A.5.3 Beliefs elicitation screen 1

How much, on average, will players in the role of A send to B's? Insert a number between 0.00 and 10.00 : _ _ _

A.5.4 Beliefs elicitation screen 2

How much, on average, will B's return to A's?

If A sends €1 and B therefore receives €8.05, B's will return on average: _ _ _
...
If A sends €10 and B therefore receives €25.30, B's will return on average: _ _ _

A.5.5 Beliefs elicitation screen 3

What is the minimum amount (on average) that A's will need back from B's in order to not feel cheated?

If A sends €1 and B therefore receives €8.05, to not feel cheated A will need back from B at least: _ _ _

...

If A sends €10 and B therefore receives €25.30, to not feel cheated A will need back from B at least: ___

A.5.6 Beliefs elicitation screen 4

What percent of participants in the role of B will return enough money to you (if you are assigned the role of A) so that you don't feel cheated?

If you send €1 and B therefore receives €8.05, what percent of B's will return enough so that you don't feel cheated?: ___

...

If you send €10 and B therefore receives €25.30, what percent of B's will return enough so that you don't feel cheated?: ___

A.5.7 Beliefs elicitation screen 5

How much money (on average) do other participants in the role of A believe will be returned to them by B's?

If A sends €1 and B therefore receives €8.05, how much money does A believe B will return? ____

...

If A sends €10 and B therefore receives €25.30, how much money does A believe B will return? ____

Appendix III: Direct Response Experiment

Section 1: Experimental design and procedures

This appendix describes the procedures and provides instructions for the direct-response experiment.

The experiment was conducted in the laboratory at the Einaudi Institute for Economics and Finance using pen and paper. It consisted of two treatments: DR-CN and DR-FOB. The sole difference between the two treatments was what we elicited from senders and subsequently transmitted to receivers. In DR-CN we elicited and transmitted senders' cheating notions; in DR-FOB we elicited and transmitted senders' first-order beliefs about their receivers' actions.

Both treatments proceeded as follows. After arriving at the lab but before being seated all participants were presented instructions for our simplified trust game. Participants were told that the experiment they would participate in would involve this game. They were then publicly randomly assigned either the sender role or the receiver role.[1] Receivers were escorted to a separate waiting room where they were instructed to wait quietly for senders to make their decisions. Once all receivers had left the room, senders were assigned experiment codes in a transparently random fashion—by drawing numbered chips from an opaque bag. Each code corresponded to a seat in the lab. Seats were separated from each other by opaque dividers, essentially creating private cubicles.

After drawing a code, each sender was handed a decision sheet and instructed to go to their cubicle to fill out their sheet. Each decision sheet asked for only two pieces of information: i) the participant's experiment code; and ii) whether they would send 0, 5 or 10 euros to their co-player. The latter piece of information was supplied by ticking a box next to one of the three options. When all senders were finished making their decisions, decision sheets were collected and another sheet of paper was handed out. This sheet asked for three pieces of information: i) their experiment code; ii) their chosen send amount;[2] and iii) either their cheating notion (DR-CN) or how much money they believed their co-player would return to them (DR-FOB).

Both the cheating notion question and the (first-order) belief question were similar to the questions used in our main experiment, but adapted to refer only to the sender's chosen send

[1] For a session with N participants, (N/2) red poker chips and (N/2) blue poker chips were placed in an opaque bag and then each participant, without looking, drew one poker chip from the bag. Those who drew a red (blue) poker chip were assigned the role of sender (receiver). As in all of our experiments for this paper, more neutral wording was used. The sender role was always referred to as "Role A" while the receiver role was "Role B." If an odd number of participants showed up, one was randomly selected to be sent home and paid a 5 euro show-up fee.

[2] If a participant asked, they were instructed to simply check the same box they had checked before. Very few participants asked.

amount and the sender's specific co-player. The cheating notion question was: "How much money would you need back from player B [the receiver] in order to not feel cheated?" As in our main experiment, participants could specify a number or select either "I don't know" or "this has nothing to do with cheating." The first-order belief question was "How much money will player B [the receiver] send back to you?" Participants could insert a number or select "I don't know." As in our main experiment, proper incentives were provided for truthful belief reporting.[3] To enhance the credibility of our beliefs elicitation mechanism, we used a physical randomizing device to resolve uncertainty.[4]

When all senders had completed this final sheet they were escorted to the waiting room. At the same time, the receivers who had been waiting there were escorted to the laboratory. Upon entering the lab, receivers were randomly assigned an experiment code by drawing a chip from among the remaining chips in the opaque bag, which insures there was no duplication in code numbers. Each receiver was handed their own blank decision sheet as well as a decision sheet from one randomly selected sender and instructed to sit in their assigned cubicle. Each receiver's decision sheet asked for five pieces of information: i) the receiver's experiment code; ii) the experiment code of the sender with whom the receiver had been paired; iii) how much money their sender chose to send to them; iv) their sender's cheating notion (DR-CN) or first-order belief (DR-FOB); and, finally, v) the receiver's decision about how much money to return. Receivers could return any amount € 0.00 ≤ r ≤ € f(s).

Once all receivers had completed their decision sheet, they were escorted back to the waiting room. In the waiting room, experimental earnings were calculated. After each participant was paid individually in cash he or she was instructed to leave the premises before the next person would be paid. This design implements a nearly double blind procedure and ensures that each participant's decision is as consequential as possible. In addition to their experimental earnings, all participants were paid a 5 euro show-up fee.

[3] Differently from our main experiment, to ameliorate hedging motives senders were instructed that either the belief question or their trust game outcome would determine their earnings. Senders were informed that we would randomly draw a number from 1 to 100, with a number larger than 75 dictating that senders' earnings would be determined by the accuracy of their beliefs. As in our main experiment we used a randomized quadratic scoring to determine senders' potential earnings from their reported belief. Senders were provided with details of this scoring rule as well as a numerical example.

[4] At the front of the room was a miniature bingo blower containing balls numbered from 1 to 100. To decide whether beliefs would be remunerated we extracted a number from this bingo blower in front of all senders. This number was extracted after all senders had submitted their beliefs but before they left the room.

Section 2: Experimental materials

Sheet 1: General game description provided to all participants before role assignment

The Game

Your experiment code is _____

General Instructions

In this experiment, you will be paired randomly with one other participant and randomly assigned one of two roles: **A** or **B**. This pairing will be anonymous. Neither the person assigned the role **A** nor the person assigned the role **B** will discover with whom they have been paired.

*The role of **A***:
The player assigned the role **A** is given €10.50 and must decide whether to send some, all or none of this money to the player assigned the role **B**, the player with whom **A** has been paired. For every euro that **A** sends, **B** receives more than one euro as reported in the table below.

If **A** sends:	€ 0	€ 5	€ 10
B receives:	€ 0	€ 17.90	€ 25.30

*The role of **B***:
After **A** makes his or her decision about how much to send to the player assigned the role of **B**, **B** must decide how much of the money he or she receives to send back to **A**. The possible amounts **B** can receive are reported in the table above. For example, if **A** sends € 5 and **B** therefore receives € 17.90, **B** must decide how much of this € 17.90 to send back to **A**.

Your earnings:
For every pair of participants, one assigned the role of **A** and one assigned the role of **B**, the decision of **A** together with the decision of **B** will determine both **A**'s and **B**'s earnings. Both **A** and **B** will be informed of the outcome determined by their decisions. However, you will not discover who your co-player was and your co-player will not discover who you are.

In general:

- if **A** sends a positive amount,
  - **A's** earnings will be: (€ 10.50) - (euro sent to **B**) + (euro sent back by **B**);
  - **B's** earnings will be: (the euro value associated with **A**'s send amount reported in the table above) - (the amount returned to **A**).

- If **A** sends zero euros to **B**,
  - **A's** earnings are € 10.50;
  - **B's** earnings are € 0.

Sheet 2: Sender's initial decision sheet (DR-CN and DR-FOB)

ROLE A

Your experiment code is _____

After you have read the game instructions on the previous page carefully, please respond to the key question below.

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

Thank you for participating in this role.

Sheet 2: Sender's final decision sheet (DR-CN)

ROLE A

Your experiment code is _____

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

QUESTION: *What is the minumum amount you would need to receive back from player B in order to not feel cheated? [leave the space blank if you chose to send € 0]*

YOUR RESPONSE:

Insert a number: € _ _ . _ _

… or choose one of the following two options:

[] I don't know
[] this has nothing to do with cheating

Sheet 2: Sender's final decision sheet (DR-FOB)

ROLE A

Your experiment code is _____

KEY QUESTION: ***how much will you send to B?***

YOUR RESPONSE:

[] I will send € 0 so that B receives € 0.00

[] I will send € 5 so that B receives € 17.90

[] I will send € 10 so that B receives € 25.30

QUESTION: *How much money will player B will return to you? [Leave blank if you chose to send € 0]*

YOUR RESPONSE:

Insert a number: € _ _ . _ _

… or choose the following option:

[] I don't know

NB:

- Your earnings from this latter question will depend on how accurate your guess is (for details, see the next page).
- You will be paid <u>either</u> your earnings from this question <u>or</u> your earnings from the game.
- To determine whether this question determines your earnings, before you leave this room we will extract a number from 1 to 100 using the randomizing device at the front of the room. If the extracted number is greater larger than 75, your earnings will be determined by this question.

Back of Sheet 2 (DR-FOB)

How we will calculate your earnings from this question:

We use the following method to calculate your earnings from the latter question in euros. The most important feature to notice is that more accurate estimates yield higher a probability of earning money.

- Your estimate, call this "R", is inserted into the following formula where "r" denotes the true value of the number being estimated and "r_{max}" denotes the maximum value the number being estimated can attain.

$$1 - \left(\frac{R - r}{r_{max}}\right)^2$$

- This produces a number between 0 and 1. We will multiply the number produced by 100 to obtain a number between 0 and 100. Call this number "z".
- At the same time, we will choose randomly a number between 0 and 100. Call this number we randomly select "y".

If y ≤ z, you will earn € 15 for your estimate,

If y > z, you will earn € 0 for your estimate.

If this question is chosen to determine your earnings, "y" will be chosen by extracting a second number using the randomizing device at the front of the room.

An example:

Imagine you are estimating the average amount that participants in the role of A will send in this game. To be concrete, suppose this average actually turns out to be 4. The maximum value this average could attain is 10, so that "r_{max}" = 10. Plugging both of these facts into the equation above yields:

$$\frac{z}{100} = 1 - \left(\frac{R - 4}{10}\right)^2$$

Now, notice that the closer your estimate, R, comes to the actual value, 4, the higher "z" will become and, consequently, the larger will be the probability that you will earn € 15 for your estimate instead of nothing.

For example, if your estimate is exactly correct, i.e., R = 4, then $\left(\frac{R-4}{10}\right)^2$ = 0 and therefore z = 100. Since the number we will randomly draw, "y," is always less than 100, your exactly correct estimate would earn you € 15 with certainty.

On the other hand, the farther away your estimate R is from the the true value, the larger z will become. Since this means that the probability that y ≤ z also increases, your chances of earning € 0 instead of € 15 from your estimate also increase.

Sheet 3: Receiver's decision sheet

Role B

Your experiment code is _____

Please read the instructions for the game. Then, read through the additional materials provided to discover: i) your co-player's code; ii) how much money your co-player in Role A decided to send to you; and [Treatment CN: *iii) how much your co-player needs back in order to not feel cheated*.] [Treatment FOB*: iii) how much your co-player believes you will send back*.] *Please write these facts in the spaces below.*

My co-player's code is: _____

My co-player sent € _ _ . _ _ , so that I received € _ _ . _ _.

[*Treatment CN:* My co-player needs back in order to not feel cheated: € _ _ . _ _]
[*Treatment FOB:* My co-player believes I will send back: € _ _ . _ _]

Next, if your co-player sent you some money please choose how much you will return.

KEY QUESTION: How much will you return to A?

YOUR RESPONSE:

I will send back to A € _ _ . _ _

Thank you for participating in this role.