The Economic Journal, **126** (*September*), 1703–1738. Doi: 10.1111/ecoj.12271 © 2015 Royal Economic Society. Published by John Wiley & Sons, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

TRUST AND CHEATING*

Jeff Butler, Paola Giuliano and Luigi Guiso

When we take a taxi we may feel cheated if the driver takes an unnecessarily long route despite the lack of a contract to take the shortest possible path. Is the behaviour of the driver affected by beliefs about our cheating notions? We address this question in the context of a trust game. We find that both parties to a trust exchange have personal notions of cheating and that these notions have a bimodal distribution. We conceptualise cheating notions as moral expectations, which provide a micro-foundation for guilt. Cheating notions substantially affect decisions on both sides of the trust exchange.

When taking a taxi 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 adviser 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 holiday 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 how the travel agent, broker or mechanic ought to behave, perhaps feeling cheated or let down when behaviour 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 (http://en.wikipedia.org/wiki/Transportation_in_New_York_City). And New York is not alone: about one million people use taxis every day in Hong Kong (http://www.gov.hk/en/about/abouthk/factsheets/docs/transport.pdf.), 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% of investors consult a financial adviser. Overall, in the UK 91% of intermediary mortgage sales are 'with advice' (Chater *et al.*, 2010). In the US, 73% of all retail investors consult a financial adviser before purchasing shares (Hung *et al.*, 2008).[1] Given their ubiquity, understanding precisely what drives behaviour in such trust-based exchange opportunities is an important undertaking.

* Corresponding author: Paola Giuliano, UCLA Anderson School of Management, 110 Westwood Plaza, C517 Entrepreneurs Hall, Los Angeles, CA 90095, USA. Email: paola.giuliano@anderson.ucla.edu.

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 study.

[1] See also Inderst and Ottaviani (2012) for a general review on financial advice.

In this article, we focus on one intuitively plausible yet under-explored determinant of behaviour on both sides of such exchange opportunities: individuals' personal subjective notions of what constitutes cheating. These notions can be thought of as the threshold of the behaviour of the counterpart that, according to the individual, separates cheating from non-cheating behaviour. While individuals may hold widely divergent views on what constitutes cheating and this heterogeneity in cheating notions may, in turn, translate into heterogeneous behaviour, economists know virtually nothing about the individual-level relationship between cheating notions and behaviour 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.[2] While this methodology has proven useful for showing that pecuniary concerns alone fail to account for a significant portion of exchange behaviour, its ability to provide a detailed understanding of how idiosyncratic cheating notions translate into behaviour is obviously limited.

We investigate the role of cheating notions in the context of a trust game (Berg *et al.*, 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 stylised, 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 behaviour 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; finally, we elicit participants' beliefs about others' cheating notions and behaviour, as well as their first and second-order beliefs about others' behaviour 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.[3]

We test several hypotheses. At the most basic level, we test whether trust-based exchanges do indeed engender personal cheating notions and whether counterparties anticipate these notions. Whether or not this will be the case is not *a priori* obvious: taxi drivers, mechanics and financial advisers may very well choose to ignore or downplay the possibility that their customers could ever feel cheated in order to reconcile

[2] For example, in a seminal work in this vein Berg *et al.* (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 (Bohnet and Zeckhauser, 2004).

[3] The between-subjects experiment is described in subsection 3.5. Details on all of our other experiments are provided in the online Appendix.

opportunistic behaviour with a positive self-image.[4] Conditional on an affirmative answer to our first question, we test the hypothesis that these implicit cheating notions have an impact in determining behaviour 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 *et al.* (1995); while, contrary to the assumptions of much of the trust game literature, a sizeable 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.[5] 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 behaviour on both sides of the potential exchange.

An important question for our study is the relationship between our results and guilt aversion (Battigalli and Dufwenberg, 2007). Building on the insights from the theory of guilt aversion, we could expect that receivers' behaviour 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 distinct from beliefs about actions. Second, we derive three different testable hypotheses which allow us to disentangle the relevance of cheating notions from the role of first and second-order beliefs in the determination of guilt.

On the first point, 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 behaviour, whereas senders' first-order beliefs are mathematical assessments about the probability of particular events. Guilt aversion theory is based on mathematical and not moral expectations (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007).

On the second point, we show that receivers' beliefs about senders' cheating notions are empirically a more fundamental determinant of guilt than second-order beliefs. One way we show this is to document that second-order beliefs have little explanatory power for receivers' behaviour beyond what is captured by beliefs about senders' moral expectations. A second way we show this is to run an additional experiment using a between-subjects design where in one treatment we provide receivers with their senders' mathematical expectations and in another treatment we provide receivers with their senders' moral expectations (see subsection 3.5 for details). We test whether senders' moral expectations constrain cheating more than senders' mathematical expectations as well as whether it is moral expectations or mathematical expectations that act more like a guilt threshold. In line with moral expectations being a fundamental determinant of guilt, we find that cheating is less likely when receivers

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

[5] 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.

know their senders' moral expectations than when they are informed of their senders' mathematical expectations and that, conditional on not cheating, the amount a receiver returns more closely mirrors his sender's moral expectations than his sender's mathematical expectations. All together, we find strong support in our data for the notion that beliefs about cheating notions provide a micro-foundation for guilt.

The remainder of the article proceeds as follows: Section 1 links our results to various strands of the existing literature. 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, concludes and suggests avenues for future research. Additional experimental treatments, analyses conducted to address the robustness of elicited cheating notions and a comparison between behaviour in our main experiment – conducted online – and a smaller study conducted in a more traditional laboratory environment can be found in the online Appendix (Appendix A). The online Appendix also provides instructions for our main experiment (Appendix B) and details the design of, and provides instructions for, our direct-response experiment (Appendix C).

1. Contribution to the Existing Literature

Our article contributes three new sets of results to the literature. First and foremost, we provide the first direct evidence on the relationship between personal cheating notions and individual behaviour in trust-based exchange opportunities. Taking into account individuals' own personal notions of what constitutes cheating, we document a significant relationship between expected cheating and how much individuals trust.

The second contribution of our study is the investigation of how beliefs about others' cheating notions constrain the behaviour of the entrusted. We find that the behaviour of individuals who refrain from intentional cheating moves in one-to-one correspondence with their beliefs about others' cheating notions.

The third contribution of our study is showing that cheating notions provide a micro-foundation for guilt which has strong promise of lending empirical content to theoretical models of guilt aversion.

Our findings are relevant for several strands of the literature. Our evidence speaks to the ongoing debate about the relevance of guilt aversion for moral behaviour. Vanberg (2008) starts from the observation that the exchange of promises has a pronounced effect on subsequent levels of co-operation in experimental games. While this observation has been used as primary evidence in favour of the relevance of guilt aversion with the interpretation that promises are kept because, once given, they cause people to attribute certain pay-off expectations to others, another plausible interpretation is that people have a taste for keeping their word (Ellingsen and Johannesson, 2004). Vanberg (2008) provides evidence suggesting that the effects of promises cannot be accounted for by changes in pay-off expectations, ostensibly leaving little room for guilt aversion theory to matter. Our study shows that guilt aversion still plays a role in co-operation even when promises are not possible so that even if a preference for promise-keeping explains the previous evidence there is still scope for guilt to affect behaviour.

More generally, our results contribute to the debate over how non-pecuniary preferences affect behaviour and where these preferences come from. Our finding that receivers' behaviour 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.[6] 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 article is also related to a nascent literature examining directly the relationship between behaviour 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 behaviour by directly eliciting these norms and relating the elicited social norms to observed behaviour. 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).

Our results are 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 behaviour, while moral norms may sway behaviour through internal sanctions even without external observability. 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 (2013).

Finally, our article relates to the huge literature investigating behaviour in the trust game.[7] The bulk of this literature focuses on what drives senders' behaviour – 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 *et al.* (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 document and study the roles that participants' moral expectations and related beliefs play in determining the behaviour of both receivers and senders. In doing so, we shed empirical light on the unresolved question of what it is that senders are trusting

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

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

receivers to do, what receivers believe senders are expecting of them, and the determinants of receivers' behaviour.

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 online.

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 behaviour 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.[8]

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 behaviour we observe.[9] However, we can partially address concerns about role uncertainty with a subsequent experiment featuring no role uncertainty (described in subsection 3.5, 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 analysing 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 randomisation 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 (Charness *et al.*,

[8] 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, p. 387) find that 'there are significantly more studies that find no difference across elicitation methods than studies that find a difference'. Moreover, behaviour in games where players have large action sets – as is ours – are more robust to the elicitation method than games where players make, for example, 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.

[9] For example, Iriberri and Rey-Biel (2011) show that such 'role uncertainty' tends to increase costly surplus creation and decrease spiteful behaviour in a simple game similar to the trust game we study here. It is not clear which behaviour 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 for example, our participants – to pay them for their work.

2012). We attempt to address the most obvious threats and confounds directly (e.g. *ex post* rationalisation). 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' behaviour 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 euro (0 euro). Second, 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 euro, the receiver receives $8\sqrt{s}$ euro.[10] 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.[11]

An added benefit of using a concave production function is to provide a relatively smooth relationship between behaviour and beliefs at the individual level, a feature which will prove useful when we examine the 'intensive margin' of trust: how much to

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

[11] 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 $f(s)/2$, which is 1.50 euro when $f(s) = 3s$ but $8.05/2 = 4.025$ euros using our concave trust production function. Consequently, these two notions would differ by only 0.50 euro using the standard trust production function, while, in stark contrast, our concave function separates these two cheating notions by just over 3 euro.

send conditional on sending something.[12] 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 some sessions: in 'high fee' sessions, senders who choose to send a strictly positive amount incur the fee, whereas senders who choose 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 model formally 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 $f(s) > 0$ euro, receivers' feasible actions were $\{0.00, 0.01, \ldots, f(s)\}$. Before discovering whether they would play the role of sender or receiver, participants submitted a complete contingent strategy for each role. The order in which participants submitted their strategies – whether first for sender, then for receiver or first for receiver and then for sender – was randomised. 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 10 separate screens. Each of these 10 screens asked only one question: 'if the sender sends s euro and you therefore receive $f(s)$ euro, how much will you return?'.[13]

2.2. *The Cheating Notion Question*

Before describing the measurement of our cheating notion, it will prove useful to provide a more formal definition of what we mean by a cheating notion. The feeling of being cheated is an emotional response to counterparty behaviour that an individual considers untoward. In situations like the trust game where a more positive response by the second-mover, that is returning more money, is less likely to give rise to negative affect, a cheating notion is a threshold separating counterparty behaviour that will definitely engender a negative emotional response from behaviour that will not. As cheating notions are defined according to subjective emotional responses, they are distinct from more familiar notions like social norms which rely on a common agreement about what constitutes socially acceptable behaviour. Being based on an anticipated emotional response also helps to separate cheating notions from the closely related concept of disappointment as it exists in the literature today, which is based on a comparison between mathematical expectations of counterparty behaviour

[12] 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 behaviour outside of the laboratory 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).

[13] The order in which receivers faced their 10 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, for example facing a monotonically increasing or decreasing sequence of send amounts.

and actual counterparty behaviour. Cheating notions are most closely related to the concept of 'betrayal', which is also an emotional response. Here, our contribution is to elicit from individuals the threshold on counterparty behaviour that will give rise to an emotional response directly, allowing these thresholds to be completely idiosyncratic, rather than indirectly inferring the existence of such thresholds from more aggregated data as is the practice in the betrayal literature (Bohnet and Zeckhauser, 2004).

In the experiment, we proceed as follows: 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:[14]

> '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 euro and *B* were to therefore receive $f(s)$ euro, you would need back how many euro?'

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.[15]

Some may argue that by asking participants about cheating so directly we may prime them to associate behaviour 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 online Appendix A, subsection 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

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

[15] Our design initially did not include the two explicit opt-out responses mentioned above. Although responding to the question was always completely voluntary, we realised that not providing pre-programmed opt-out responses could make some participants feel obliged 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 utilise all 428 observations. In online Appendix A, we show that our results are robust to restricting attention only to sessions with explicit opt-out.

(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.

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 online Appendix B for exact wording).[16]

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.[17]

Participants were informed that one estimate from this subsection would be chosen to count towards their potential earnings. This chosen belief was remunerated according to a randomised 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 euro in most sessions while, in the remaining sessions, an exactly correct belief paid 20 euro. 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* rationalisations of beliefs about others' cheating notions or others' expected returns. For example, participants could *ex post* rationalise 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

[16] Items (*ii*)–(*v*) were asked for each possible send amount.
[17] This was done to avoid mechanical correlations between reported beliefs and participants' own strategies or cheating notions.

not feel cheated. We treat these concerns extensively using several different robustness check exercises. Full details are reported in the online Appendix A, subsection A.2.

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 < 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 euro 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-laboratory sessions. We had participants come to the laboratory and complete the online experiment. In these in-laboratory sessions, 100% of participants were paid their experimental earnings. Participants' behaviour in our in-laboratory trust game was remarkably similar to behaviour in our online 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 online Appendix A, subsection A.1).

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

Table 1

Experimental Design

	Number of sessions	Explicit cheating notion question opt-out	Investment fee	Max belief pay	Observations
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	SD	Min	Max	*N*
Male	0.46	0.499	0	1	420
Age	23.73	4.171	18	58	420
Maths 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
Pr(*NotCheated*)	0.42	0.232	0	1	427
Average proportion of non-cheaters	0.49	0.376	0	1	428

2.5. *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, introduced spurious correlations among our measures which may be driving our results (see the discussion in Charness *et al.*, 2012). One could also be concerned about the unintended effects of eliciting individuals' beliefs about the behaviour of the experimental population rather than about the behaviour of their specific co-players.[18]

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 experiment, however, we retained the quadratic trust production function so that receivers could receive three possible amounts: {0,17.90,25.30}. This

[18] 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.

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.[19]

3. Results

We establish three main results:

(*i*) there is substantial heterogeneity in cheating notions and beliefs about others' cheating notions;
(*ii*) cheating notions affect decisions on both sides of the trust exchange; and
(*iii*) cheating notions beliefs have more explanatory power than second-order beliefs in explaining receivers' behaviour.

We, therefore, show that cheating notions are a fundamental determinant of guilt and that understanding them may provide a micro-foundation for guilt and guilt aversion theory.

Before starting our analysis, we establish short acronyms for some of our variables. We typically denote the sender's action by s and the receiver's action by $r(s)$. In addition to behaviour 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 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:

[19] 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 the online Appendix C.

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 euro?'
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 percentage 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 percentage of *B*s will return enough so that you will not feel cheated?'.

Notes. Each variable listed in this table is actually a set of 10 variables, one for each possible send amount $s = 1, \ldots, 10$. However, as in the Table, we will typically suppress the dependence on s for ease of exposition.

(*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 behaviour.

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 footnote 15). 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% when considering sending 10 euro, to a high of 20% 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

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)	Observations
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
SE	(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
SE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	

Notes. 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'. 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.

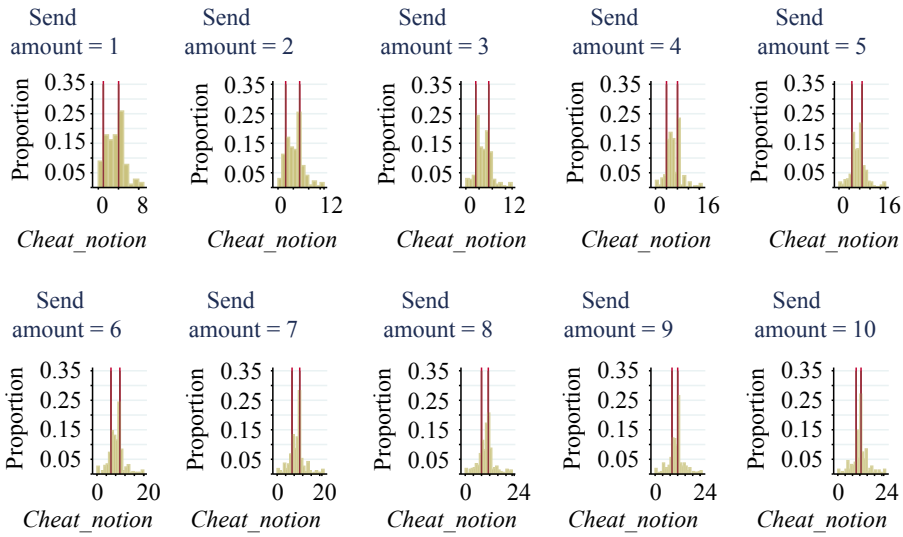

Fig. 1. *Own Cheating Notions (Cheat_notion)*

Notes. The Figure reports histograms of participants' personal cheating notions for each send amount $s = 1, \ldots, 10$. 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.

'this has nothing to do with cheating', or just leaving the question blank – in these same sessions is also low, ranging from 17% to 23%. 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 summarised 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.[20] An individual whose cheating notions are consistent with this rule would for each send amount, $s \in \{1, \ldots, 10\}$, 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 that is for each s, the line is placed at $f(s)/2$. Accordingly, we call this an 'equal split' cheating notion.[21] 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 behaviour 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 rule 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 2), 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.[22] We find that, again, cheating notions are consistent with a positive return on investment rule for about one-third

[20] This is the cheating standard explicitly assumed in Berg *et al.* (1995) and incorporated into much of the subsequent literature on trust (Bohnet and Zeckhauser, 2004).

[21] 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 euro. An individual with an equal split cheating notion would feel cheated by receiving < 4.02 euro back which would correspond to (sender earnings, receiver earnings) = (13.02, 4.03).

[22] 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 fee.

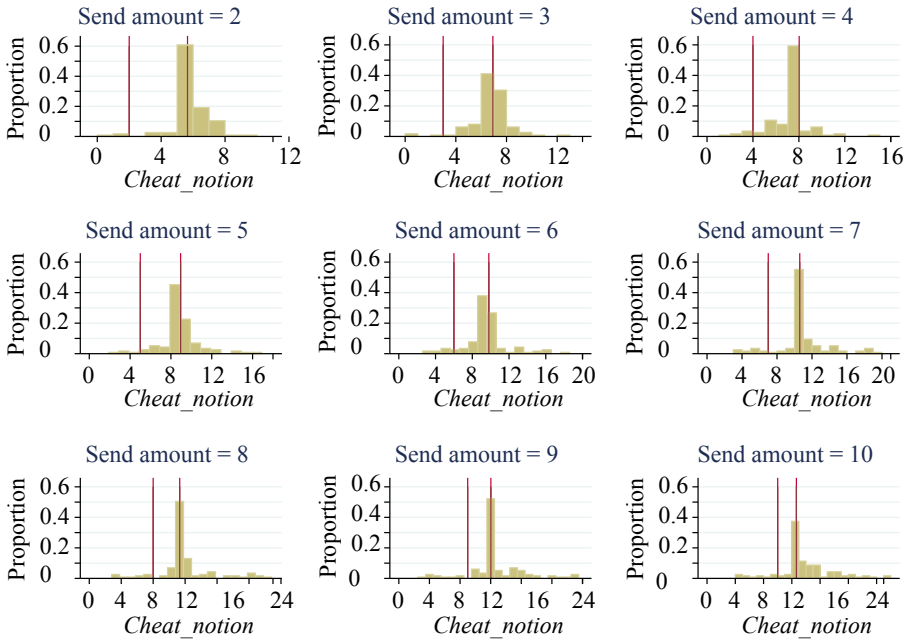

Fig. 2. *Within-individual Consistency of Cheating Notions Across Send Amounts, Equal Split*
Notes. The Figure restricts attention to participants whose cheating notions were consistent with equal split conditional on a send amount of 1, and presents histograms of these participants' cheating notions for all other send amounts. Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

(31%) of our participants when $s = 1$. For this approximately one-third of participants, we report in Figure 3 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, that is 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 4 and 5). In the online Appendix, we also report analogous within-individual consistency exercises for both first-order and second order beliefs, finding a remarkable amount of consistency.[23]

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. A plausible conjecture for this link is

[23] See Figures A1 and A2 for first-order beliefs and Figures A3 and A4 for second-order beliefs (all in the online Appendix).

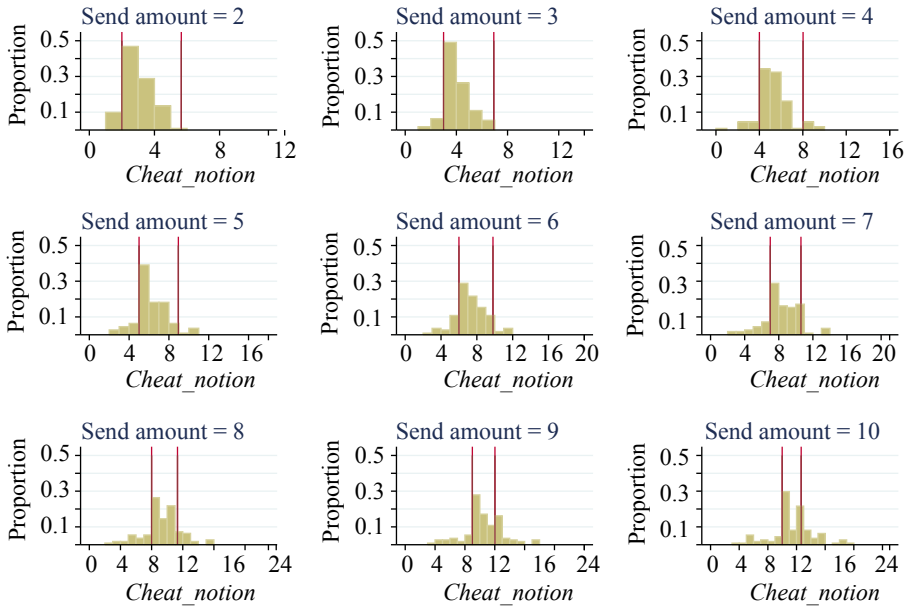

Fig. 3. *Individual-level Consistency of Cheating Notions Across Send Amounts, Strictly Positive Return on Investment*

Notes. 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. Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

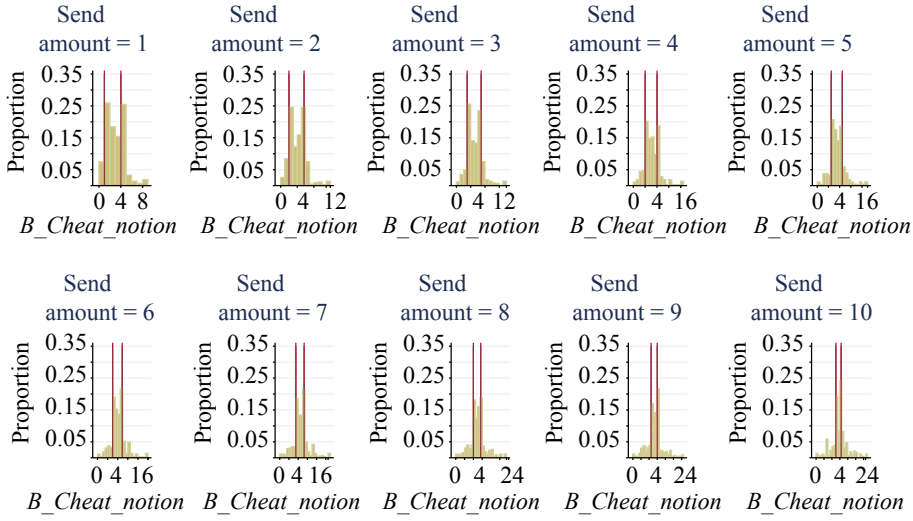

Fig. 4. *Participants' Beliefs About Others' Cheating Notions (B_Cheat_notion)*

Notes. The Figure reports histograms of participants' beliefs about other participants' cheating notions (*B_Cheat_notion*). Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

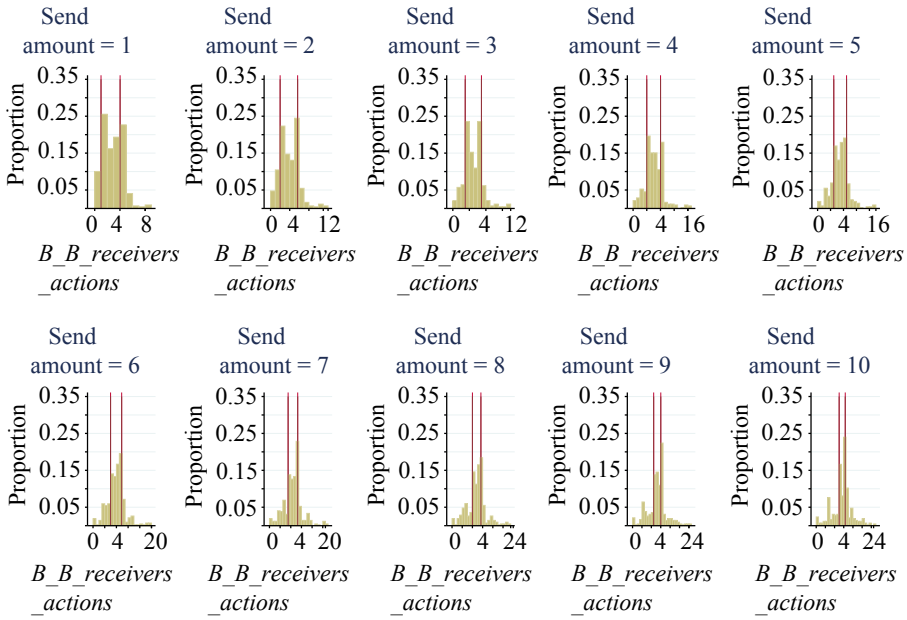

Fig. 5. *Second-order Beliefs (B_B_receivers_actions)*
Notes. The Figure plots participants' beliefs about senders' beliefs about receivers' actions (*B_B_receivers_actions*). Vertical lines are placed at the weakly positive return on investment and equal split cheating definitions.

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 (Ross *et al.*, 1977; in a trust game context, see Butler *et al.*, forthcoming). If own cheating notions themselves are persistent – perhaps being based on moral values which tend to be culturally transmitted from parents to children (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 cheating notion beliefs.

We provide evidence on both links.

We focus our attention on the second link, the most relevant for our article. In our data, 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 ($p < 0.01$).[24] In the online Appendix, we investigate the first link and test directly for a relationship between

[24] 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.

our participants' cheating notions and the values their parents emphasised during their upbringing.

Overall, 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 behaviour. Consequently, in the next subsection, we focus on the relationship between cheating notion beliefs and behaviour.

3.2. *The Relationship Between Cheating Notions Beliefs and Behaviour*

In this subsection, we look at the effect of cheating notion beliefs on the behaviour of both receivers and senders.

3.2.1. *Receivers' decision to cheat intentionally*

One advantage of focusing on cheating notion beliefs directly is that we can study what drives receivers' decision to cheat intentionally. 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 5 presents our estimates of receivers' propensities to cheat intentionally 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 mathematics 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 that is, 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 receivers' own cheating notions (*Cheat_notion*) is stronger for lower levels of s: the reported probit coefficients imply that the marginal effect of

Table 5

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)
Maths 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)
Observations	369	366	366	369	371	370	371	369	366	366

Notes. 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. Robust standard errors, clustered by session, in parentheses. ***Significant at 1%, **significant ay 5%, *significant at 10%. Maths score is individual's self-reported score on required maths exams taken during the final year of high school in Italy. Income variables refer to self-reported annual family income from all sources, in thousands of euros, net of taxes. The excluded category is 'below 30,000 euro annually'. 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.

an increase in *Cheat_notion* at send amount 10 is half that at send amount 1 (1.6 percentage points *versus* 3.6 percentage points, respectively).[25]

3.2.2. *The effects of cheating beliefs on senders' behavioural trust*

As a second step, we consider whether and how the spectre of being cheated affects senders' behaviour. While previous research suggests that expected cheating or betrayal may affect the likelihood of trusting behaviour (Bohnet and Zeckhauser, 2004), it is an open question whether the likelihood of being cheated affects the intensive margin of trust – that is, 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 10 *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).[26] The measure can therefore be interpreted as subjective probability of not being cheated. We call this measure Pr*(NotCheated)*.

Figure 6 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 centres around a 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*,

[25] In addition to running regressions for each send amount separately, we also ran a probit model pooling across all send amounts and using participant random fixed effects. We control for all the same variables as in Table 5 and insert an additional control for send amount. Results are qualitatively similar: cheating notions are negatively and significantly correlated to intentional cheating and *B_cheat_notion* is positively and significantly related to intentional cheating. Additionally, the coefficient on send amount in this specification is negative and significant, which provides further evidence that cheating significantly decreases in sender vulnerability or 'niceness'.

[26] Individuals who did not report a cheating notion conditional on sending s euro 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 euro '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 euro as 1 before constructing their summary measure.

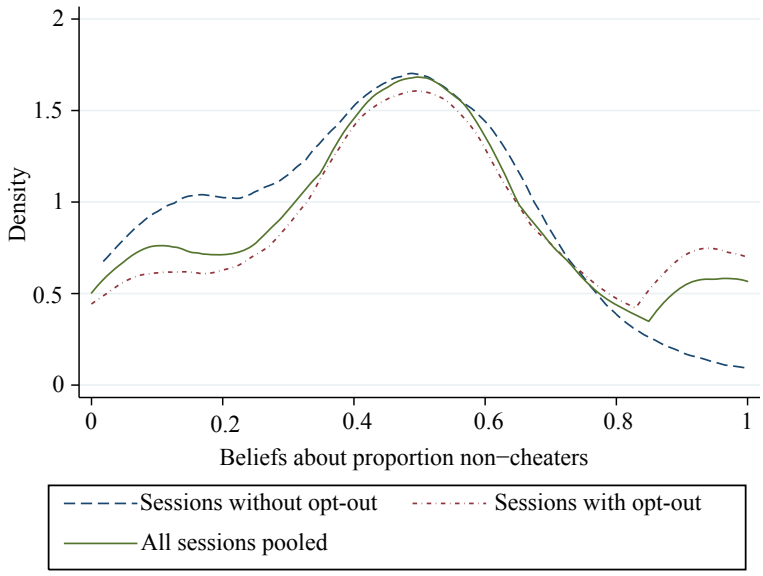

Fig. 6. *Beliefs About the Probability of not Being Cheated*
Notes. Observations in the sessions with opt-out (short-dash line) are restricted to individuals who have a cheating notion for every possible amount a sender could send. This is to ensure our summary measure of beliefs about the probability of being cheated is well-defined. Thus, the density plot for the additional sessions is based on 207 (out of 306) observations.

B_receivers_actions, conditional on sending *s* euro 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.[27]

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' behaviour 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 as well as a proxy for altruism obtained from that same survey.[28]

Table 6 presents the estimates. The second column presents the selection equation, which is a probit model estimate of the decision to send something *versus*

[27] 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).
[28] 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.

nothing, that is 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 spectre of being cheated plays a significant role in the intensive margin of trust: the positive

Table 6

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*) × *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)
Maths 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)
Observations	350	350
Mills's ratio	0.33	
	(0.18)	

Notes. Robust standard errors, clustered by session, appear in parentheses ***Significant at 1%, **significant at 5%, *significant at 10%. For the Heckman model (columns. 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. 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. '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. Income variables refer to (self-reported) annual family income from all sources, in thousands of euro, net of taxes. The lowest category is excluded: 'below 30,000 euro'.

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 behavioural 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 10 (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' behaviour. 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.[29]

3.3. *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 counterparty behaviour: 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 theories of guilt aversion rely on beliefs about others' mathematical expectations to define the threshold of behaviour engendering guilt. We, on the other hand, hypothesise that beliefs about others' moral expectations

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

are a more fundamental determinant of guilt. Consequently, we argue that moral expectations may provide a micro-foundation for guilt and guilt aversion theory.[30]

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 expect mathematically to not be cheated.[31] 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 subsection, we show that not only are senders' mathematical and moral expectations correlated but that this correlation is also reflected in receivers' beliefs – *B_Cheat_notion* and *B_B_Receivers_actions*. 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 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.

[30] A methodological caveat in our definition of first and second-order beliefs is that while cheating notions are elicited at the individual level first and second-order beliefs are elicited with reference to an 'average other' rather than an individual's specific co-player. We think that it is reasonable to assume that, since the individual knows nothing about his or her specific co-player, asking about an average other should not result in a different answer than asking about one's specific co-player. This view is supported by our between-subjects experiment which produces consistent behavioural patterns when the first-order beliefs of each second mover's specific co-player are transmitted to the second mover directly.

[31] We provide evidence in the online Appendix A, subsection A.3, that reported cheating notions are not 'reverse-caused' in this respect: that is, that participants do not form beliefs about the amounts participants return and then simply report this belief as their cheating notion as to avoid, for example 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.

For the sake of brevity, we report details and results of these tests in the online Appendix.[32] 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' behaviour beyond what is contained in cheating notion beliefs? We turn to this question next.

3.3.1. *What constrains receivers behaviour more?*

In this subsection, 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. We do this in three ways. First, we estimate receivers' behaviour, $r(s)$, as a function of both *B_Cheat_notion* and *B_B_receivers_actions* simultaneously for each $s = 1, \ldots, 10$, separately.[33] Our estimates reveal that receivers' beliefs about senders' moral expectations (*B_Cheat_notion*) are almost always highly significant predictors of receivers' behaviour while their beliefs about senders' mathematical expectations (*B_B_receivers_actions*) almost never are (Table 7).[34,35]

Second, we 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. We do so by providing some receivers with their sender's mathematical expectations and other receivers with their sender's moral expectations and showing that moral expectations behave more like the type of threshold we would expect from guilt aversion models.

[32] For details about the empirical strategy to test for these correlations and the corresponding results, see part A.4. of the online Appendix and Tables A14–A15.

[33] In each of these 10 regressions we also control for a host of demographics to isolate the impact of the beliefs in question on behaviour. 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.

[34] Obviously, one may be worried that the lack of significance of *B_B_Receivers_Actions* is due to collinearity 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.

[35] The results in Table 7 could be consistent with Vanberg (2008). The author provides evidence that only a specific moral norm (promise-keeping) matters in determining cooperation, suggesting that beliefs do not matter at all. We show in a trust context that with a larger sample size and more precise measures of moral norms and beliefs, both moral norms and beliefs matter. The lack of significance of second-order beliefs in our context can be explained by the fact that the direct effect of second-order beliefs is simply difficult to detect since moral norms shape these beliefs.

Table 7

Predicting Receiver Behaviour: 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 × *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 × *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^2	0.20	0.15	0.16	0.13	0.19	0.12	0.14	0.15	0.17	0.14

Notes. Robust standard errors, clustered by session, in parentheses. ***Significant at 1%, **significant at 5%, *significant at 10%. Each column presents an OLS estimate using the dependent variable $r(s)$, where s is specified in the column heading. 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 euro, $i = 1, \ldots, 10$; '*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 euro, $i = 1, \ldots, 10$. Each estimate includes demographic controls, omitted for readability from the Table. These controls are: gender, age, mathematics score, family income and risk aversion.

We can perform this exercise using the data from our direct-response experiment and testing whether the event $[r < Cheat\_notion]$ in the cheating notion treatment (DR-CN) is less likely than the event $[r < B\_receivers\_actions]$ in the first-order beliefs treatment (DR-FOB).[36] 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 behaviour 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 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).

As an additional way of asking which type of expectations acts more like a guilt threshold, we look at whether, conditional on satisfying the sender's expectation, receivers exactly satisfy the expectation. If the expectation in question is a true guilt threshold, then returning more would not reduce guilt 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 that conditional on returning at least as much as their sender's mathematical or moral expectation, receivers' behaviour will more closely mirror moral expectations than mathematical expectations: $r - B\_Cheat\_notion < r - B\_B\_receivers\_actions$.

To test this hypothesis, we pool all send amounts and 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. 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 (SE = 0.35) in DR-CN, while in DR-FOB this distance is almost three times as large (1.43, SE = 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 7 we plot the raw data from the direct-response experiment, this time restricting attention to observations where $s > 0$. We

[36] 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*).

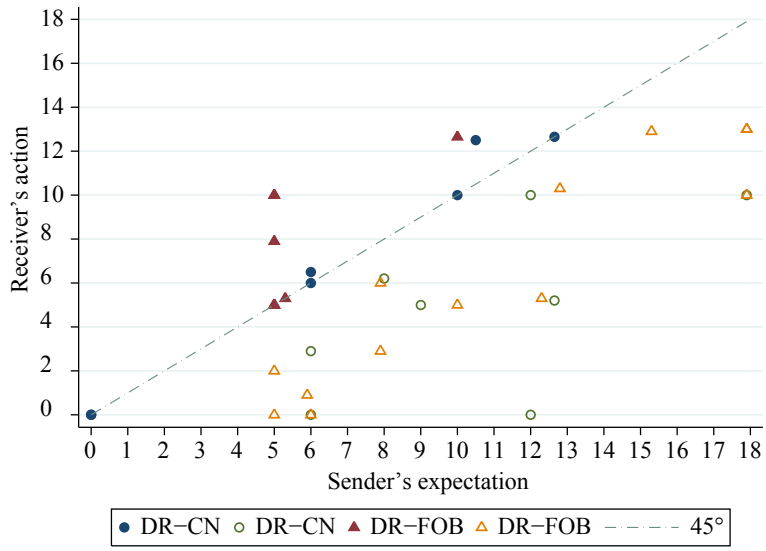

Fig. 7. *Receiver's Actions Versus* Sender's Expectations, DR-CN and DR-SOB

Notes. 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. Solid markers correspond to observations where the receiver did not cheat – that is, returned at least as much as their sender's expectation – while hollow markers correspond to observations where the receiver cheated. The dashed line is a 45-degree line along which a receiver's action exactly matches his or her sender's expectation.

overlay the plot with a 45° line. We use solid markers to indicate observations above the 45° 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 match them, exactly as observations in DR-CN not involving cheating typically lie quite close to the 45° 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 their 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.[37] Overall we find that receivers who are given their sender's cheating

[37] 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 euro to a maximum of 12 euro with an average of 4.99 (SE = 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.

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 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 account formally 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 8 are broadly consistent with intentional cheating giving rise to guilt.[38] 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' behaviour than their beliefs about senders' mathematical expectations. The interpretation we favour is that violating senders' moral expectations is a primary determinant of guilt in trust-based exchanges.

Wrapping up, in this subsection 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 behaviour, 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

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

Table 8

Sensitivity of Amounts Returned to Beliefs About Senders' Cheating Notions by Decision to Cheat, Heckman Model

	Send amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating ($r \geq B\_Cheat\_notion$)										
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
Observations	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
Observations	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. Standard errors in parentheses. ***Significant at 1%, **significant at 5%, *significant at 10%. Each column presents a Heckman model estimate using as its exclusion restriction participants' own cheating notions. The dependent variable in column i is the amount a participant will send back if the sender sends i euro, $i = 1, \ldots, 10$. 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 euro, $i = 1, \ldots, 10$. Each estimate includes our standard set of demographic controls, omitted for readability from the table. These controls are: gender, age, mathematics score, family income and risk aversion.

elicitation 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. Conclusion, Discussion and Interpretation

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 article investigates what individuals' personal, subjective notions of what constitutes cheating – their moral expectations – can tell us about behaviour 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. Altogether, our results suggest that studying cheating notions and related beliefs can help us understand and predict behaviour in trust-based exchange.

An interesting question to ask is whether it is possible to provide a more general view on the type of preferences that could explain receivers' cheating decisions. Informing this exercise is the fact that receivers show a declining propensity to cheat as they receive larger sums from senders (Figure 8).[39] This 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.

[39] Figure 8 plots 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. 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 the Figure.

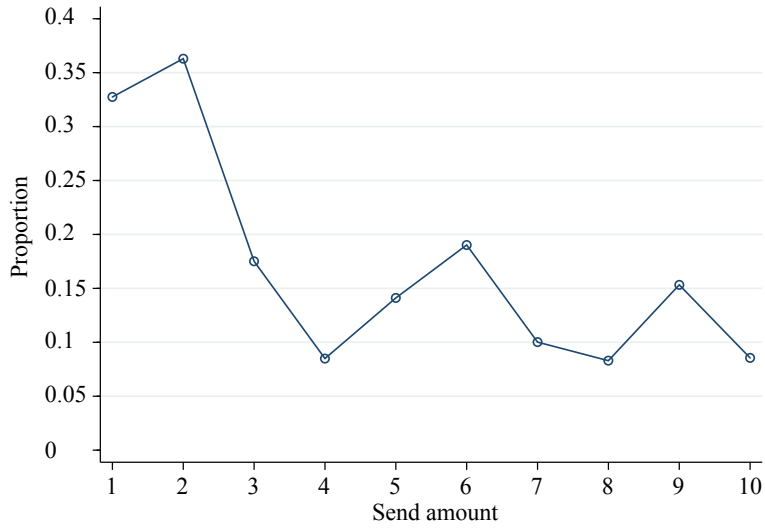

Fig. 8. *Proportion of Cheaters by Send Amount*

Notes. The Figure plots the fraction of receivers who intentionally cheat at each send amount after partialling out the effect of *B_Cheat_notion*. 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 the Figure.

Patterns in our data also appear to be inconsistent with literal interpretations of many influential social preferences models.[40] In the online Appendix, we provide a unified way to model both senders' and receivers' preferences that is consistent with our data. The model adds a moral cost function to standard pecuniary preferences. 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, for example they might suffer guilt. Senders lose utility when receivers cheat them. The model helps in explaining:

(*i*) why the decision to cheat depends on others' expected cheating notions;
(*ii*) why cheating depends on the intensity of moral preferences as proxied for by receivers' own cheating notions; and
(*iii*) why the probability of cheating decreases in amounts sent as shown in Figure 8.

[40] 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 behind when sent 4 euro. The patterns suggest that receivers' behaviour is unlikely to be explained by purely distributional concerns.

It also helps rationalising 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.

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 behaviour, 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.

EIEF and University of Nevada - Las Vegas
UCLA Anderson School of Management
EIEF

Accepted: 2 February 2015

Additional Supporting Information may be found in the online version of this article:

Appendix A. Robustness Checks.
Appendix B. Experiment Instructions.
Appendix C. Direct Response Experiment.
Data S1.

References

Akerlof, G.A. and Dickens, W.T. (1982). 'The economic consequences of cognitive dissonance', *American Economic Review*, vol. 72(3), pp. 307–19.
Battigalli, P. and Dufwenberg, M. (2007). 'Guilt in games', *American Economic Review*, vol. 97(2), pp. 170–6.
Berg, J., Dickhaut, J. and McCabe, K. (1995). 'Trust, reciprocity and social history', *Games and Economic Behavior*, vol. 10(1), pp. 122–42.
Bisin, A. and Verdier, T. (2010). 'The economics of cultural transmission and socialization', in (J. Benhabib, A. Bisin and M.O. Jackson, eds.), *Handbook of Social Economics*, vol. 1A, pp. 339–416, Amsterdam: North-Holland.
Bohnet, I. and Zeckhauser, R. (2004). 'Trust, risk and betrayal', *Journal of Economic Behavior and Organization*, vol. 55(4), pp. 467–84.
Brandts, J. and Charness, G. (2000). 'Hot vs. cold: sequential responses and preference stability in experimental games', *Experimental Economics*, vol. 2(3), pp. 227–38.
Brandts, J. and Charness, G. (2011). 'The strategy method vs. the direct-response methodL a first survey of experimental comparisons', *Experimental Economics*, vol. 14(3), pp. 375–98.
Butler, J.V., Giuliano, P. and Guiso, L. (forthcoming). 'Trust, values and false consensus', *International Economic Review*.
Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton NJ: Princeton University Press.
Castillo, M., Petrie, R., Torero, M. and Vesterlund, L. (2012). 'Gender differences in bargaining outcomes: a field experiment on discrimination', NBER Working Paper No. 18093.
Charness, G. and Dufwenberg, M. (2006). 'Promises and partnership', *Econometrica*, vol. 74(6), pp. 1579–601.
Charness, G., Gneezy, U. and Kuhn, M.A. (2012). 'Experimental methods: between-subject and within-subject design', *Journal of Economic Behavior & Organization*, vol. 81(1), pp. 1–8.
Charness, G. and Rabin, M. (2002). 'Understanding social preferences with simple tests', *Quarterly Journal of Economics*, vol. 117(3), pp. 817–69.

Charness, G. and Schram, A. (2013). 'Social and moral norms in allocation choices in the laboratory', Economics Working Paper Series, Department of Economics, UC Santa Barbara.
Chater, N., Huck, S. and Inderst, R. (2010). 'Consumer decision-making in retail investment services: a behavioral economics perspective', Report to the European Commission/SANCO.
Dufwenberg, M. and Gneezy, U. (2000). 'Measuring beliefs in an experimental lost wallet game', *Games and Economic Behavior*, vol. 30(2), pp. 163–82.
Ellingsen, T. and Johannesson, M. (2004). 'Promises, threats and fairness', Economic Journal, vol. 114(495), pp. 397–420.
Fehr, E. and Schmidt, K.M. (1999). 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics*, vol. 114(3), pp. 817–68.
Gneezy, U. (2005). 'Deception: the role of consequences', *American Economic Review*, vol. 95(1), pp. 384–94.
Hung, A.A., Clancy, N., Dominitiz, J., Talley, E., Berrebi, C. and Suvankulov, F. (2008). 'Investor and industry perspectives on investment advisers and broker-dealers', Technical Report, Rand Institute for Civil Justice.
Inderst, R. and Ottaviani, M. (2012). 'Financial advice', *Journal of Economic Literature*, vol. 50(2), pp. 494–512.
Iriberri, N. and Rey-Biel, P. (2011). 'The role of role uncertainty in modified dictator games', *Experimental Economics*, vol. 14(2), pp. 160–80.
Krupka, E.L. and Weber, R.A. (2013). 'Identifying social norms using coordination games: why does dictator game sharing vary?', *Journal of the European Economic Association*, vol. 11(3), pp. 495–524.
Lundquist, T., Ellingsen, T., Gribbe, E. and Johannesson, M. (2009). 'The aversion to lying', *Journal of Economic Behavior & Organization*, vol. 70(1-2), pp. 81–92.
Reuben, E. and Riedl, A. (2013). 'Enforcement of contribution norms in public good games with heterogeneous populations', *Games and Economic Behavior*, vol. 77(1), pp. 122–37.
Ross, L., 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*, vol. 13(3), pp. 279–301.
Rustichini, A. and Villeval, M. (2012). 'Moral hypocrisy, power and social preferences', GATE Working Paper No. 1216.
Schlag, K. and van der Weele, J.J. (2013). 'Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality', *Theoretical Economics Letters*, vol. 3(1), pp. 38–42.
Selten, R. (1967). 'Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments', in (H. Sauermann, Ed.), *Beiträge zur experimentellen Wirtschaftsforschung*, pp. 136–68, Tübingen: Mohr.
Smith, V.L. (1976). 'Experimental economics: induced value theory', *American Economic Review*, vol. 66(2), pp. 274–9.
Vanberg, C. (2008). 'Why do people keep their promises? An experimental test of two explanations', *Econometrica*, vol. 76(6), pp. 1467–80.