



Munich Personal RePEc Archive

# Conformist Preferences in Mixed-Motive Games

Naef, Michael and Sontuoso, Alessandro

Economics Dept., Royal Holloway, University of London, Philosophy,  
Politics and Economics, University of Pennsylvania

13 September 2015

Online at <https://mpra.ub.uni-muenchen.de/66965/>  
MPRA Paper No. 66965, posted 29 Sep 2015 06:54 UTC

# *Conformist Preferences in Mixed-Motive Games*

Michael Naef<sup>a</sup> and Alessandro Sontuoso<sup>b,\*</sup>

<sup>a</sup> *Economics Dept., Royal Holloway, University of London, Egham, Surrey, TW20 0EX*

<sup>b</sup> *Philosophy, Politics and Economics, University of Pennsylvania, 249 S 36th St., Philadelphia, PA 19104*

September 13, 2015

We examine a novel class of conformist preferences which falls within the realm of belief-dependent motivations in that the peers' expectations about others' behavior may affect every group-member's welfare. Similar other-regarding motivations, like guilt-aversion, have been inferred from evidence of a belief-behavior correlation but the issue of causality has been disputed. In examining conformism we propose a design that verifies the presence of the relevant causality direction while ruling out alternative other-regarding motivations. Our data reveal "self-servingly conformist" behavior in that subjects choose to match their strategy to the peers' expectations *when it is in their interest to do so*.

KEYWORDS: conformist preferences, consensus effects, belief-dependent utility, guilt aversion, social norms, trust.

## **I. Introduction**

Conformism is an element of major importance to economic outcomes as information about peer behavior has been shown to influence a diverse range of choices, including employees' retirement savings decisions and executives' decisions (Beshears *et al.* [2015], Shue [2013]). The approach of economic theory to conformism has generally fitted into two broad research streams. The first of such streams assumes that peers' behavior is copied since it reflects private information that is relevant to the individual's own decision (Banerjee [1992], Bikhchandani *et al.* [1992]):<sup>1</sup> in this case the economists' interest is motivated by the fact that herding leads to actions that are informed in ways that might make the individual worse off. The second stream, sometimes referred to as the esteem-based account of conformism, aims to capture the individual's intrinsic inclination to identify with a certain class of people (*e.g.* by modeling one's utility from being perceived as an individual of high status or as a principled person; Bernheim [1994], Akerlof [1980]): in this case the focus is not on information-acquisition but rather on status-seeking. While the former approach can be used

---

<sup>1</sup> An informational cascade (otherwise referred to as "herding behavior") occurs when individuals observe the predecessors' actions, and then make the same choice that others have made irrespective of their own private information signals. See Anderson and Holt [1997] for an experimental account of herding behavior.

\* Corresponding author.

*E-mail addresses:* michael.naef@rhul.ac.uk; sontuoso@sas.upenn.edu.

to describe the patchy adoption of innovations (Kapur [1995]), the latter can explain why performance in real effort tasks may be influenced by social pressures (Gaviria and Raphael [2001]). In short, both accounts of conformity have informed several applications, yet both generally imply that whether a conformist copies or not the others has no *direct* impact on the others' welfare.<sup>2</sup>

In this paper we set out to study conformism in strategic situations such that one's actions affect – not only one's own but – also the others' outcomes.<sup>3</sup> Specifically, we set out to study the case of individual preferences bending towards collective behaviors, so that whether one conforms to what others think is “normal” may affect every group-member's welfare.<sup>4</sup> We believe this case to be of great importance to economic outcomes in that it closely characterizes some of the complex strategic interactions that take place within small socio-economic groups. Yet the analysis of conformist preferences in non-controlled environments has traditionally presented some challenges (Manski [1993]). In fact, when regularities about collective behaviors are observed, it is often difficult to find out whether group-members behaved similarly because they had similar individual characteristics; or else if it was each individual's perception of the others' behavior, preferences and beliefs that

---

<sup>2</sup> While *herding* models imply that the predecessors' choices influence one's choice, the payoff structure is usually defined in such a way that each individual who chooses the right option will get some fixed reward (irrespective of the others' choices). In this respect whether a conformist copies or not the others has no direct impact on the others' welfare. Similarly, *esteem-based* models assume an individual to care about the others' perception of her own status, but the others' actions do not typically enter each individual's utility function.

<sup>3</sup> Note that our goal is not to identify the psychological determinants of conformist preferences, but just to understand what we should see in the data generated by our experiment if such belief-dependent preferences are at work.

<sup>4</sup> The importance of social comparison and information on relative performance has been studied within the realm of labor economics. Kandel and Lazear [1992] and Huck *et al.* [2012] address the question of whether peer pressure improves performance in situations where payoffs are based on team incentives (our approach substantially differs because, there, a subject's effort task was explicitly defined so as to create a *positive externality* on other team members' payoffs). Note that the importance of information transmission has been studied also with reference to “nudges” (*i.e.* persuasion tactics that make the individual aware of the actions of others who have been in a similar, usually non-strategic, situation); see Costa and Kahn [2013] for a field experiment involving energy conservation nudges.

determined group behavior (e.g. peer norms).<sup>5</sup> With reference to the latter case, the social psychology literature has provided overwhelming evidence that the way people behave is linked to their beliefs about others, but such belief-behavior relationships remain multifaceted and intricate. For instance, a line of psychological research has focused on a *tendency for individuals to adjust their behavior to what is commonly perceived to be normal in their group* (which is often termed “social conformity”; Cialdini and Goldstein [2004]).<sup>6</sup> In contrast, another line of research has shown that our own behavior influences what we expect of others, where the “(false) consensus effect” denotes a *bias to overestimate the extent to which others are like us* (Ross *et al.* [1977]).

Economists have only recently turned their attention to belief-behavior relationships, especially with regard to belief-dependent motivations: like in the case of social conformity, such motivations are characterized by a causal direction running from beliefs to own behavior, whereas the opposite direction would hold for the consensus effect. But while the experimental economics literature has provided substantial evidence for a correlation between beliefs and behavior in social dilemmas, the question of causal direction remains open. Put it differently, is the observed correlation due to the individuals basing *their own behavior* on what they or others think is normal to do (*i.e.* the causality runs from beliefs to own behavior)? Or else is it due to the individuals basing *what they think is normal to do* on their own behavior (*i.e.* the causality runs from own behavior to beliefs)? In this paper we examine

---

<sup>5</sup> Note that such perceptions may be right or wrong, regardless of the observed behavior. For example, *pluralistic ignorance* involves every individual to privately hold the belief that her own preferences are different from those of all other group-members, even though the observed behavior is identical in that each individual refrains from publicly revealing her preferences (Allport [1933]). This should not be confused with the case in which a minority of the individuals have correct perceptions about most other group-members, and *intrinsically like* to adapt to the majority’s preferences and beliefs.

<sup>6</sup> The first tests for conformity were conducted by Solomon E. Asch [1956], a pioneer in social psychology who undertook a series of small-group studies on the social pressures to conform. His experimental subjects were asked to determine the relative lengths of lines: all but one of the participants in each session were confederates of the experimenter and had beforehand been instructed to give wrong answers in unanimity at certain points; as a result, approximately 35% of subjects gave the same incorrect answer as the misleading majority. Note that, while in Asch’s experiments subjects responded to others’ *observed* behavior, further research has extended the notion of conformism to include a tendency to adapt to the others’ *presumed* behavior as well as to the others’ expectations, values, and norms (Berkowitz and Perkins [1986], Bicchieri [2006], *etc.*).

a class of belief-dependent motivations – “conformist preferences” – for which we provide formal definitions and further propose a design so as to unequivocally verify the presence of the relevant causality direction.

Unlike in the herding and esteem-based accounts of conformity, here each player’s preferences are represented as a function of both beliefs *and* strategy profiles. Our notions of conformist preferences also differ from previously investigated belief-dependent motivations, including *intention-based inequity aversion* and *reciprocal kindness* (Falk *et al.* [2008], Dhaene and Bouckaert [2010], Fehr and Gächter [2000], *etc.*), and *guilt aversion* (Dufwenberg and Gneezy [2000], Charness and Dufwenberg [2006], Vanberg [2008]). In fact each of the previous investigations has typically assumed that players care about one predefined rule/norm of “kindness” or of “promise-keeping”. In this paper, in contrast, we depart from those previously suggested motivations as our conformity notions do not imply a (more or less) stable disposition towards a unique rule of fair behavior. All we assume is that the *peers’ expectations about other group-members* may serve the individual as a means to guiding her own actions: the interpretation we put forward is that this may occur because conformist individuals take their peers’ expectations to signal *whatever* behaviors are “normal” of the group.

It should be stressed that belief-dependent motivations are usually inferred from evidence of a belief-behavior correlation.<sup>7</sup> In particular, a correlation between actions and second-order beliefs – in trust games – has been interpreted as evidence for “guilt aversion” (Charness and Dufwenberg [2006]), sometimes referred to as “trust responsiveness” (Guerra and Zizzo [2004], Bacharach *et al.* [2007]). This notion implies that an individual *i* adapts her behavior to the opponents’ beliefs about *i*’s behavior, in order to avoid the uncomfortable feeling of guilt that ensues from having let someone down. More specifically the presence of guilt aversion has been inferred from the evidence of a correlation between own behavior and

---

<sup>7</sup> The only exception is provided by Costa-Gomes *et al.* [2014], who create an artificial instrumental variable to estimate the causal effect.

second-order beliefs (about own behavior), with such beliefs being elicited by asking subjects what they think that the opponents expect from them. However, it has been suggested that any such observed correlation may have been due to consensus effects, which involve the opposite causal direction: *people who are inclined to cooperate might infer from their own inclination that people in general are cooperative* (Ellingsen *et al.* [2010]).<sup>8</sup>

In what follows we will define the class of belief-dependent motivations we refer to as conformist preferences, which imply that individuals adapt their behavior to other individuals' expectations (thereby involving a causal direction inverse to that of consensus effects); we then propose a design that verifies the presence of the relevant causality. In short, a conformist individual is assumed to follow her peers' expectations: not to avoid letting the matched coplayer down – as in the case of guilt aversion – but rather to fit in with the purported majority (Boyd and Richerson [1985]). To that end, we believe we unequivocally establish the presence of the causality implied by conformist preferences for two reasons: (i) we provide an *exogenous variation* in beliefs which, in the event of the relevant belief-behavior correlation, would rule out consensus effects as an explanation; (ii) such an exogenous variation involves one's second-order *belief about the behavior of other players in the same role*. Hence, unlike in previous studies here we manipulate expectations about the behavior of same-role players who, as such, are not directly affected by those beliefs. For the first time we provide evidence for the causal effect of such beliefs on behavior: in particular, our data reveal “*self-serving conformism*” as subjects choose to match their strategy to the exogenous information *only when it is in their interest to do so*.

Specifically, our experimental design involves a standard two-player dichotomous trust game where we inform each subject of the behavior other *same-role participants* expect

---

<sup>8</sup> Ellingsen *et al.* [2010] tested for a causal effect of the matched coplayer's expectations on one's behavior by transmitting the respective Trustor's first-order belief to each Trustee (thereby generating an “induced” *second-order belief about own behavior* per Trustee). Having found no evidence for guilt aversion, they conjectured that the correlation (between own behavior and second-order beliefs about own behavior) observed in previous experiments might have been driven by consensus effects. In other words, in Ellingsen *et al.*'s experiment, the existence of consensus effects has been presumed from the *absence of evidence* for causal effects in the other direction.

of *same-role participants*. More precisely, first, we elicit each subject's expectation about the behavior of players in the same role (which can be termed as the subject's *first-order belief about the peers' behavior*); then, some of these beliefs are averaged and transmitted to other participants in the same role, which amounts to providing subjects with an "*induced*" *second-order belief about the peers' behavior*. Note that the consensus effects hypothesis predicts that the first-order beliefs should correlate with the subjects' own behavior in that subjects would assume that most others behave like they do. Our data can indeed be consistent with a consensus effect explanation (as we find a strong and significant correlation between *first-order* beliefs and behavior), although we are not directly able to infer the direction of causality underlying this particular correlation. However, by exogenously varying *second-order* beliefs we are able to provide a clearer picture of how consensus and conformity relate to and interact with each other. Our data reveal significant, though conditional, conformity effects in that we find that subjects adapt their behavior to the induced second-order beliefs *only when it serves their interests*.<sup>9</sup> In fact, our results show that Trustees with a first-order belief of predominant cooperation (on the part of other Trustees) were positively and significantly affected by the inducement of a second-order belief of predominant cooperation; put it differently, Trustees who had *high* expectations of their peers' cooperative behavior – but were informed that their peers did not share those expectations – cooperated significantly less often than those Trustees who had high expectations and were told that their peers too had high expectations (therefore, the information about the peers' expectations had an impact on their own behavior, leading some to keep more money for themselves). In contrast, Trustees with a first-order belief of predominant non-cooperation were not significantly affected by the inducement of a second-order belief of predominant cooperation; that is,

---

<sup>9</sup> In this connection, in a gift exchange game with a principal and two employees, Thöni and Gächter [2014] find that – when an employee learns that her coplayer has provided a lower effort than hers – the employee revises her effort downwards; however the employee hardly increases her effort when a coplayer has provided a higher effort than hers. This would be consistent with what we refer to as self-serving conformism, although our framework explicitly deals with expectations about *unmatched participants* while Thöni and Gächter's design involves a *three-player* game (where there are no direct strategic interdependencies between the two employees' decisions).

Trustees who had *low* expectations of their peers' cooperative behavior were unaffected by the information about their peers' expectations (therefore, Trustees who were informed that their peers expected high cooperativeness were just as likely to keep the money for themselves as those who were informed that their peers expected low cooperativeness).<sup>10</sup>

In summary, we find that the peers' expectations about the behavior of others do serve the individual as a means to guiding her own actions; the interpretation we suggest is that such expectations are taken to signal what behaviors are normal of the group. In what follows we will refer to an individual's tendency to conditionally "adopt" the peers' expectations (*i.e.* the induced second-order beliefs about the peers' behavior) as an instance of self-serving conformism, whereas we will refer to "*pure conformity*" when the tendency is unconditional. Before proceeding it should be noted that our findings are related to those of Xiao and Bicchieri [2010] whose study shows that, in a trust game variant where the precepts of two potentially applicable social norms conflict, the Trustees' "*normative expectations*" (*i.e.* beliefs that a certain behavior ought to be followed) are consistent with one norm/principle only when it is in the Trustees' interest, and are otherwise consistent with the other principle.<sup>11</sup>

The remainder of the essay is organized in this manner: II reviews some consensus-conformity issues in more detail; III presents the experimental design; IV introduces some theoretical definitions and results, and derives from them our experimental hypotheses; V discusses the main experimental results; VI provides some robustness checks, and VII concludes.

---

<sup>10</sup> The effect of the exogenous beliefs on Trustors' behavior is symmetrical, that is, an increase in the transmitted belief has a significant positive effect only on the group of Trustors who held an empirical expectation of predominant non-cooperation: in section V we will show that such a symmetrical impact of exogenous beliefs on Trustors' behavior is consistent with our self-serving conformism hypothesis.

<sup>11</sup> See also Dana *et al.* [2007] and Andreoni and Bernheim [2009] for a line of research highlighting an illusory preference for fairness.



## II. Overview of some conceptual issues

The focus of this paper is not on moral or social norms per se, since our design involves the elicitation of *empirical* and not of *normative* expectations; however, the experimental literature on norm-conformity does provide some relevant insights (Schram and Charness [2015], Krupka and Weber [2013]).<sup>12</sup> In particular Bicchieri and Xiao [2009] proposed a first test of conformity to a social norm in dictator games: to that end, they presented each Dictator with information about the majority of the Dictators' choices in a previous session (*i.e.* empirical information), or with information about the majority beliefs about “what ought to be done” (*i.e.* normative information), or with both empirical and normative information. After that, Dictators made their decision and, only then, they were asked what they thought other Dictators believed they would and should choose (which amounts to eliciting empirical and normative beliefs).<sup>13</sup> While Bicchieri and Xiao's [2009] results provide a first measure of norm-conformity, their design would not be apt to investigate – at the individual level – how prior expectations interact with the induced expectations: since the authors elicited subjects' beliefs *after* some information was transmitted, their design does not allow one to measure the relative strengths of conformity and consensus.<sup>14</sup>

---

<sup>12</sup> According to the theoretical literature on norm-conformity, a *social* norm is at work if behavior is consistent with empirical and normative expectations (Sugden [2000], Bicchieri [2006]). In particular, Bicchieri [2006] provides a set of conditions for conformity to a social norm, and Sontuoso [2013] integrates Bicchieri's account with psychological game theory. Specifically, Sontuoso assumes that there exists a set of default rules of behavior; players keep some of these rules stored in their minds and derive from them expectations as to which (rule-driven) strategies are appropriate for the present game. A “social norm” is followed by a population whenever players maximize their expected utilities, given their beliefs and their preferences being conditional on a *psychological-game* theoretic version of Bicchieri's [2006, p. 11] conditions, that is: *i.* players are aware of the existence of some rule of behavior; *ii.* they believe that the others will behave in keeping with some rule (*i.e.* *empirical expectations* condition); *iii.* they believe that the others expect them to behave in keeping with some rule, and the cost of a potential violation is sufficiently high to make it disadvantageous (*i.e.* *normative expectations* condition).

<sup>13</sup> Bicchieri and Xiao's results show that the percentage of fair offers was higher in the sessions where either empirical or normative information about fair behavior was *transmitted* (compared with sessions where either empirical or normative information about selfish behavior was provided). Also, they found that Dictators' *elicited* empirical beliefs were correlated with their behavior.

<sup>14</sup> Gächter *et al.* [2013] contrast social preferences with social norm-based explanations of the impact of information transmission. Their design involves a three-player gift exchange game where a principal pays a

More generally, the psychology literature conceives of conformity as implying a tendency for individuals to adjust their behavior to what is (commonly perceived to be) normal in their group. In our experimental game the actual behavior of one's *peers* (*i.e.* players who have been assigned the same role) is unobservable. Hence, our focus will indeed be on conformism as a subject's response to some collective perception of "normal behavior", with normal being intended as *most frequent* so as to capture a majority-conformity bias (Boyd and Richerson [1985]). For the purposes of this investigation we define *conformist preferences* as a tendency for subjects to – more or less conditionally – behave as "it is thought" that most players in the same role would behave. In the lab this would effectively amount to testing the hypothesis that one can predict subject *i*'s behavior from *i*'s belief about some collective *belief about the behavior of i's peers* (*i.e.* predict *i*'s behavior from *i*'s second-order belief about the peers' behavior). Yet it should be noted that, if the actual behavior of one's peers is unobservable, then it is likely that one's first- and second-order beliefs about the peers' behavior will coincide: it follows that, in this case, a correlation *between* the individual's first-/second-order belief *and* the individual's own behavior would not allow us to distinguish between conformity and consensus effects.<sup>15</sup>

Hence, our experimental design introduces an exogenous variation in second-order beliefs about the peers' behavior. In this case, our (for the moment informal) definition of conformist preferences implies that a subject would adapt her behavior to some aggregate measure of the others' beliefs. In fact, in the event of belief transmission – and in the case of

---

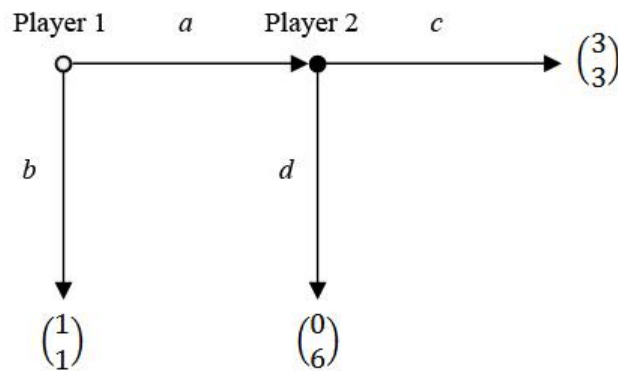
wage to each of two employees, who then make effort choices sequentially. The authors show that the availability of information about others' efforts has an influence on the second employee's decisions even though there are no direct strategic interdependencies (note that, in their specific case, this would be predicted also by inequity aversion models).

<sup>15</sup> Note that the consensus effects hypothesis makes the following prediction: if consensus effects are present, (when forming a belief about the others' behavior) a subject will estimate the other players' decisions based on her own decision. That is, consider a subject *i* who chooses a certain action with a higher probability than some other player *g* does; then, subject *i* will give a higher *estimate of a third player j choosing that action* than the estimate given by player *g* (Ross *et al.* [1977], Dawes [1989], Engelmann and Strobel [2000]). Therefore, consensus and conformism are empirically related in that both imply a correlation between own behavior and some perception of the "normal behavior".

any difference between  $i$ 's *first-* and  $i$ 's *induced second-order* beliefs – a tendency to follow the latter beliefs would reveal the presence of conformist preferences.<sup>16</sup>

### III. Experimental design and procedures

Consider the following dichotomous trust game.



**Figure 1** - The Trust Game

At the initial node Player 1 (referred to as “Participant A” in the lab) chooses either  $a$  or  $b$ : when opting for  $b$  the game terminates and material outcomes are allocated as shown in the relevant vector of payoffs, with the number on top referring to Player 1’s payoff. If Player 1 opts for  $a$  the choice passes to Player 2 (referred to as “Participant B” in the lab), who in turn can decide on  $c$  or  $d$ , the consequences of which are shown in the respective payoff vectors.

---

<sup>16</sup> Whereas that would *not* necessarily mean that the conformist subject has consciously updated her first-order belief upon receiving the exogenous information, it would certainly mean that she has updated her second-order belief.

## 1. Design

Our main treatment introduces an exogenous variation in second-order beliefs about the peers' behavior by showing subjects the average guess (about the strategy taken by same-role participants) made by a sample of other same-role participants.<sup>17</sup>

Each experimental session consisted of the following stages: Introduction Stage; Play Stages I, II; Payment Stage.

**Introduction Stage.** Subjects were randomly allocated to terminals and given the paper instructions; there, they were told that in Part I they would be assigned one of two roles, and explained the decisions involved in each role. (As it will soon be clear, given that every subject in each part was to be privately assigned the same role, the matching of subjects would be effectively implemented only at the end of Part II.) After going through the paper instructions, each subject was asked to answer a set of control questions. A summary of the instructions was finally read aloud by the experimenter.

**Play Stage, Part I.** All plays were conducted using the strategy method. The order of subsequent tasks was as reported below.

- (i) Subjects were (privately) assigned the role of Participant B.
- (ii) Each subject was asked to guess how many of the other Participants B in the same session would choose either  $c$  or  $d$ , which – in the lab – were labeled as “*share*” and “*keep*”, respectively. (For the purposes of presenting the econometric analysis, below such stated beliefs will be denoted by  $\gamma_2(\textit{share})$  and  $\gamma_2(\textit{keep})$ , respectively.) Subjects entered their guess by positioning a slider to the desired percentage.<sup>18</sup>

---

<sup>17</sup> Note that Ellingsen *et al.*'s [2010] design would not be apt to investigate whether/how conformism operates in trust games, because their design only involved the transmission of information about *one* individual's guess (*i.e.* the opponent's). Instead, social psychologists describe conformism as a subject's response to some *collective* perception of “normal behavior”.

<sup>18</sup> Note that  $\gamma_2(\textit{keep}) \equiv 1 - \gamma_2(\textit{share})$ . Also note that the slider was initially positioned at a value of 50%. Subjects had to enter a guess by moving the slider towards a higher *share* rate (*i.e.* towards a value of 100%) or towards a higher *keep* rate (*i.e.* towards a value of 0%). Subjects could not leave the slider in the initial position: this was intended to encourage participants to take a stance and express a belief about the modal behavior.

- (iii) Each subject was invited to wait until all participants had entered their guesses, after which each subject was given feedback about other Participant B's guesses. (Using our formal notation  $\bar{y}_2(\textit{share})$  and  $\bar{y}_2(\textit{keep})$ , respectively.)
- (iv) Each subject was asked to choose either  $c$  or  $d$ , namely “*share*” and “*keep*”, respectively.

**Play Stage, Part II.** Subjects were told that Part II involved exactly the same steps as in Part I, although they would be assigned a different role and matched with a different participant than in Part I. After they had been given a brief reminder of the instructions, both on-screen and orally, subjects were privately assigned the role of Participant A. Steps (i)-(iv) of Part II had the same structure as above, except that each subject's decision, guess, and transmitted information were about  $a$  or  $b$ , which – in the lab – were labeled as “*in*” and “*out*”, respectively.

**Payment Stage.** The payment mechanism consisted of two parts:

- each subject received a £3 show-up fee;
- each subject was paid according to the outcome, as shown in the vector of payoffs, at the end-node realized in Part I as well as at the end-node realized in Part II (payoffs were in pound sterling).

A few comments are now due. First, it should be noticed that the above order of the decisions (which is reversed with respect to the natural sequence Participant A — Participant B) was made possible by the adoption of the strategy method. Also, it should be stressed that in Part II each subject was matched with a participant other than that she was matched with in Part I. Besides, subjects did not know about the tasks to be undertaken in Part II until the end of Part I.<sup>19</sup>

---

<sup>19</sup> Obviously no one would know how much subjects had earned in each part, until the end of the experiment (because every subject in each part was privately assigned the same role, and the matching of subjects was effectively implemented at the end of Part II).

## 2. *The belief transmission mechanism*

The information provided at step (iii) consisted of the average guess made by a sample of other participants in the same role, in the same session. Note that, when entering their guesses (at step (ii)), subjects did not know that those guesses would be pooled and transmitted to other participants (at step (iii)). Such feedback was shown in the lower part of the same screen in which subjects were asked to enter their guesses: the message was phrased in such a way as to look like the outcome of an opinion survey; the font style and size were the same as those of the other messages, in order not to make the information too prominent. In Part I the message read: «A sample of other participants B in this session expects on average that  $\langle x \rangle\%$  will transfer half the money, whereas  $\langle 100-x \rangle\%$  will keep all the money». Similarly, in Part II the message read: «A sample of other participants A in this session expects on average that  $\langle x \rangle\%$  will OPT IN, whereas  $\langle 100-x \rangle\%$  will OPT OUT». (See Appendix B for additional information on the experimental procedure and instructions.)

In order to collect enough data so as to conveniently test for our key hypotheses, a computerized sampling method was used for selecting the guesses to be pooled and passed on to participants. To ensure that in Part I some subjects received information about an average belief of low cooperation (*i.e.*  $\bar{\gamma}_i(\text{share}) < 0.5$ ) and some others received information about an average belief of high cooperation (*i.e.*  $\bar{\gamma}_i(\text{share}) > 0.5$ ), each subject  $i$  was shown the average guess made by a specifically selected sample of participants. More precisely, the guesses were selected in a way such that all  $\bar{\gamma}_i(\text{share})$  converged to the values of 0.25 and 0.75. Similarly, in Part II the guesses were selected in a way such that all  $\bar{\gamma}_i(\text{in})$  converged to the values of 0.25 and 0.75. It should be noted that, for a given subject  $i$ , the pieces of information transmitted in Part I and II formed one of the following combinations:  $\bar{\gamma}_i(\text{share}) \sim \bar{\gamma}_i(\text{in}) \sim 0.25$ , or  $\bar{\gamma}_i(\text{share}) \sim \bar{\gamma}_i(\text{in}) \sim 0.75$ , or  $\bar{\gamma}_i(\text{share}) \sim 0.25$  and  $\bar{\gamma}_i(\text{in}) \sim 0.75$ , or  $\bar{\gamma}_i(\text{share}) \sim 0.75$  and  $\bar{\gamma}_i(\text{in}) \sim 0.25$ . That is, some subjects received information about an average belief of low (or high) cooperation in both Part I and Part II, whereas some subjects received information about an average belief of low cooperation in Part I and high cooperation in Part II or *vice versa*. Note that the reason why the sampling algorithm was devised in a way to select samples of subjects (*i.e.* guesses) such that all  $\bar{\gamma}_i(\cdot)$  converged to 0.25 and 0.75 was just to obtain two distributions of transmitted beliefs for each part of a session, that is, the “low transmitted belief” and the “high transmitted belief” distributions. In

this respect it should be noticed that one could have chosen any other value; on the other hand, 0.25 and 0.75 have been preferred for the only reason that they are unique in that each is the central value of a range of beliefs about low cooperation [0.00-0.49] and high cooperation [0.51-1.00], respectively.

It should be stressed that we phrased the on-screen message reporting the average of others' stated beliefs in such a way as to minimize deception. As mentioned above, the message explicitly stated that the reported information referred to *a sample of other participants* (see Appendix B). While such on-screen message never referred to a "random" or "representative" sample, we acknowledge that some subjects might have interpreted it as being random. On the other hand, we believe that our design complies with the norm of experimenter honesty in that we did not lie to the subjects: firstly, most English dictionaries simply define a sample as "a subset of a population"; secondly, we note that – when a subset is intended to be representative of the whole – especially in the context of opinion survey results, the word *sample* is always accompanied by some *qualifier* (e.g. representative, random, unbiased, fair, *etc.*). Finally, it should be stressed that the reason why we chose not to transmit mean beliefs from random samples was clearly not an attempt to deceive subjects, but just a means to have some subjects receive information about an average belief of low cooperation and have some others receive information about high cooperation, in a symmetric fashion. As a matter of fact, should we have used random samples, we might have needed hundreds of observations for us to obtain the symmetric distribution of transmitted beliefs necessary to conduct the analysis below.

#### **IV. Theoretical insights, predictions and hypotheses**

Consider the dichotomous trust game of Figure 1 above. Recall that subjects play both roles, first *as a Trustee* and then *as a Trustor* – each time with a different counterpart – and assume that they maximize an expected utility function exhibiting some other-regarding preferences, as follows.

$$u_1(s_1 = a) = 3 \cdot \alpha_1(c) + 0 \cdot (1 - \alpha_1(c)) + f(\beta_1(s_1), s_1) \quad (1)$$

$$u_1(s_1 = b) = 1 + f(\beta_1(s_1), s_1) \quad (2)$$

$$u_2(s_2 = c) = 3 + f(\beta_2(s_2), s_2) \quad (3)$$

$$u_2(s_2 = d) = 6 + f(\beta_2(s_2), s_2) \quad (4)$$

Note that:

- $\alpha_1(s_2)$  denotes *Player 1's first-order belief about Player 2 choosing  $s_2$* , with  $s_2 \in \{c, d\}$ ;
- $\beta_i(s_i)$  denotes *Player  $i$ 's second-order belief about herself choosing  $s_i$* , with  $i \in \{1, 2\}$  and  $s_1 \in \{a, b\}$ ,  $s_2 \in \{c, d\}$ ;
- $f(\beta_i(s_i), s_i)$  is a function (of Player  $i$ 's second-order belief and move) that captures the *psychological utility one gets from a certain behavior*.

It should be highlighted that the above beliefs are first- and second-order beliefs *about the behavior of a subject* (as opposed to the first- and second-order beliefs *about the peers' behavior* we referred to in the previous section). For the moment we will simply assume that  $f(\beta_1(\cdot), a)$  and  $f(\beta_2(\cdot), c)$  are *weakly positive*, and that  $f(\beta_1(\cdot), b)$  and  $f(\beta_2(\cdot), d)$  are *weakly negative*, for any second-order belief  $\beta_i(s_i)$ .<sup>20</sup> By looking at equations (1) to (4), it is clear that Player 1 will find it relatively more convenient to cooperate, the larger is the absolute value of  $f(\beta_1(\cdot), a)$  or of  $f(\beta_1(\cdot), b)$ . Similarly, it is clear that Player 2 will find it relatively more convenient to cooperate, the larger is the absolute value of  $f(\beta_2(\cdot), c)$  or of  $f(\beta_2(\cdot), d)$ . Yet what is unclear is how, say, Player 2 derives her second-order belief  $\beta_2(s_2)$ , or how that affects the magnitude of the psychological term, or even how Player 2's strategy relates to Player 1's strategy (given that each subject plays both roles, with different coplayers).

In what follows we shall briefly explore some alternative motivations/theories, which are characterized by different belief-behavior relationships and, hence, different psychological terms. In order to rule out the alternatives to conformism as explanations for

---

<sup>20</sup> We are being intentionally vague about the specific functional form of  $f(\beta_i(s_i), s_i)$ . Yet the reader can anticipate that – as it will soon be clear – most other-regarding preferences can be captured by some variation of  $f(\beta_i(s_i), s_i)$ .



our data patterns, we will start by reviewing the implications – to our sequential trust game – of models with inequity averse or reciprocal players.

**Proposition 1.** (Blanco *et al.* [2014], p. 127) Models with inequity averse players (Fehr and Schmidt [1999], Bolton and Ockenfels [2000]) or reciprocal players (Dufwenberg and Kirchsteiger [2004]) predict a negative correlation of moves.

The above claim is proved by Blanco *et al.* [2014] who, like us, find experimental evidence for a positive correlation of Trustees' and Trustors' strategies and, as a result, rule out inequity aversion and reciprocity as possible motivations in their sequential trust games.<sup>21</sup> In short, the intuition is that both inequity averse and reciprocal players (as modeled in the aforementioned theories) are more likely to defect as Trustors but more likely to cooperate as Trustees than selfish players: so the aforementioned theories would only be consistent with a negative correlation of moves. Blanco *et al.* [2014] provide a thorough discussion and proofs of Proposition 1 with reference to each of the above theories, so we refrain from discussing this result in detail.

Before considering some explanations that could be consistent with the observed correlations we are going to need some notation. Let  $\gamma_i(s_i)$  denote *Player i's belief about the strategy taken by other same-role players*. Also, let  $\bar{\gamma}_i(s_i)$  denote the *information transmitted to Player i* about the mean belief (about the strategy taken by same-role players) that some other same-role players hold. In this regard it should be stressed that, say, for some given Player 2,  $\gamma_2(\cdot)$  and  $\bar{\gamma}_2(\cdot)$  are both probability distributions over the strategies of other Trustees: yet they differ in that  $\gamma_2(s_2)$  is the individual's estimate (*i.e.* the estimate directly made by one player), whereas  $\bar{\gamma}_2(s_2)$  is the aggregate measure of the estimates made by multiple players and then transmitted to that individual. So, once again, one can view  $\gamma_2(s_2)$  as Player 2's first-order belief about the peers' behavior, and  $\bar{\gamma}_2(s_2)$  as Player 2's induced second-order belief about the peers' behavior. (In what follows we will often refer to the

---

<sup>21</sup> In our case the correlation coefficient is 0.3705,  $p = 0.0001$ .

former as the “prior” or “stated” belief, and to the latter as the “exogenous” or “transmitted” belief.)

Now, one recurring explanation for behavioral regularities in experimental trust games is given by guilt aversion. To establish a direct parallel with empirical studies of guilt averse preferences in sequential trust games, our main focus will be on Trustees’ rather than on Trusters’ behavior. (So, for the sake of simplicity our next propositions and hypotheses will be stated with reference to Trustees’ behavior.) If guilt aversion is present, then there will be a simple causal relationship from *second-order beliefs about own behavior* to *own behavior*. In that case the Trustee’s psychological term  $f(\beta_2(s_2), s_2)$  is defined to be the additive inverse of any (positive) difference between the expected value of the Truster’s payoff *derived from*  $\beta_2$ ,<sup>22</sup> and the payoff  $m_1(s_2)$  the Truster actually ends up with, that is:

$$f(\beta_2(s_2), s_2) = -k_2[\max\{0, E_{\beta_2}[m_1|h^0] - m_1(s_2)\}], \quad (5)$$

with  $k_2$  representing the (non-negative) guilt sensitivity of each individual (Battigalli and Dufwenberg [2007]). Assuming that guilt aversion is the only motivation at play, then it follows that providing Trustees with exogenous information  $\bar{\gamma}_2(s_2)$  should have no effect on their behavior, as per Proposition 2 below.

**Proposition 2.** Guilt aversion implies that  $s_2$ ,  $\beta_2(s_2)$ , and  $s_1$  are positively correlated; but  $s_2$  and  $\bar{\gamma}_2(s_2)$  are not correlated, nor are  $s_1$  and  $\bar{\gamma}_1(s_1)$ .<sup>23</sup>

*Proof.* See Appendix A.

Note that guilt aversion proper (as modeled by Battigalli and Dufwenberg [2007]) is agnostic as to any relationship between  $s_i$  and  $\gamma_i(s_i)$ . Generally a guilt averse player is thought to be caring only about the extent to which she lets her own opponent down: therefore, if a Trustee

---

<sup>22</sup>  $E_{\beta_2}[m_1|h^0]$  is the *expected value* of  $m_1$ , calculated with respect to  $\beta_2$ , at the initial history.

<sup>23</sup> A similar result applies to models allowing for efficiency concerns combined with maximin preferences, as in Charness and Rabin [2002] and López-Pérez [2008]. Kranz [2010] also allows for efficiency concerns while using a flexible framework: there, if one defines a moral norm so that it prescribes Trusters to cooperate and Trustees to reciprocate, then the model will imply perfect correlation of strategies and second-order beliefs.

is motivated only by guilt aversion, then she will be indifferent as to what other players in the same role believe or do; this implies that  $s_i$  and  $\bar{\gamma}_i(s_i)$  are not correlated.<sup>24</sup>

Another candidate explanation is represented by consensus effects (Ross *et al.* [1977]). As mentioned above, if only consensus – and not conformism – is present, then there will be a *causal relationship from own behavior to own beliefs* about others in the same role, and thus there should not be an effect of providing exogenous information.<sup>25</sup>

$$s_i \Rightarrow \gamma_i(s_i). \quad (6)$$

**Proposition 3.** The consensus effect implies that  $\gamma_2(s_2)$ ,  $s_2$ ,  $\alpha_1(s_2)$ , and  $s_1$  are positively correlated; but  $s_2$  and  $\bar{\gamma}_2(s_2)$  are not correlated, nor are  $s_1$  and  $\bar{\gamma}_1(s_1)$ .<sup>26</sup>

*Proof.* See Appendix A.

**Hypothesis 1 (Either consensus effects or guilt aversion or inequity aversion or reciprocity).** Manipulating  $\bar{\gamma}_2(s_2)$  has no impact on  $s_2$ .

Before proceeding it should be stressed that, for the purposes of formulating the hypotheses, we will presume that only one motivation (as modeled by the relevant theory) is at play, and not a combination of motivations.

---

<sup>24</sup> While Proposition 2 refers to the theoretical notion of guilt aversion of Battigalli and Dufwenberg [2007], we cannot entirely exclude the possibility that the desire to avoid feeling guilty of having let someone down plays a role in determining behavior choices in our design. In other words it is possible that – although our Trustees are only informed of other Trustees' expectations (about other Trustees' behavior) – they might infer the *respective Trustor's expectations* from such exogenous information; and hence they might choose to behave in a way so as to avoid disappointing the respective Trustor. However, since Ellingsen *et al.* [2010] have not been able to find any guilt aversion in a trust game with *exogenous transmission of the respective Trustor's expectations*, we can reasonably assume that guilt aversion is not present in our setting either.

<sup>25</sup> Note that the consensus effect is not an other-regarding preference, hence the psychological term – for players affected by consensus – is simply null, *i.e.*  $f(\beta_i(s_i), s_i) = 0$ .

<sup>26</sup> The following reasoning implies a further step: if Trustees believe that other Trustees behave likewise and that Trustors think likewise, then it follows that  $\beta_2(s_2) \sim \gamma_2(s_2)$ . This is indeed Ellingsen *et al.*'s argument. (Note that, still, that implies no correlation between  $\bar{\gamma}_2(s_2)$  and  $\beta_2(s_2)$ .)

The next candidate motivation is “pure conformity”, which is a tendency for subjects to behave as “it is thought” that most players in the same role would behave, *regardless of any material consequences*. Such a tendency is supposed to arise from a desire to fit in with the purported majority, no matter what. In the case of the Trustee, this implies that a subject will suffer a psychological disutility whenever she deviates from some diffused belief about the Trustees’ modal behavior. More specifically, there are a number of possible ways to re-define the psychological term so as to capture the tendency for the Trustee to think that she is expected to behave like most other Trustees are believed to be behaving: naturally, any such psychological term must be defined to be an increasing function of the act of matching the exogenous belief about the Trustees’ modal behavior. In Appendix A we propose one specific psychological term, as an illustration, but for the purposes of the current analysis it will suffice to say that if only pure conformity – and not consensus – is present, then there will be a *causal relationship from second-order beliefs about the peers’ behavior to own behavior*, and thus there will be an effect of providing exogenous information:

$$\bar{\gamma}_2(s_2) \Rightarrow s_2. \tag{7}$$

**Proposition 4.** A purely conformist motivation implies that  $\bar{\gamma}_2(s_2)$  and  $s_2$  are positively correlated.

*Proof.* See Appendix A.

Note that, in the case of receiving an exogenous belief  $\bar{\gamma}_2(c) > 50\%$ , the interpretation is that a purely conformist individual will take such a belief to signal that it is considered normal to cooperate and, therefore, she will believe that she is expected to cooperate and she will in fact cooperate so as to fit in.<sup>27</sup> This leads to our next empirical hypothesis.

---

<sup>27</sup> Whether an individual  $i$  decides to cooperate – when playing as a Trustor – may depend on *both* the exogenous belief about the Trustors’ modal behavior *and* the exogenous belief about the Trustees’ modal behavior. In fact it should be recalled that, in our setting, subjects play both roles: first *as a Trustee* and then *as a Trustor*, each time with a different counterpart. Hence, expression (1) above  $u_1(s_1 = a) = 3 \cdot \alpha_1(c) + 0 \cdot (1 - \alpha_1(c)) + f(\cdot)$  can be re-defined as  $u_1(s_1 = a) = 3 \cdot \bar{\gamma}_2(c) + f(\bar{\gamma}_1(a))$ ; as for the Trustor’s psychological term, for the purposes of this investigation we will simply assume it to be an increasing function of the (act of

**Hypothesis 2 (*Pure conformity*).** Manipulating  $\bar{\gamma}_2(s_2)$  has always an impact on  $s_2$ .

We conclude this section by discussing another possible pattern, which is implied by the motivation we refer to as “self-serving conformism”, that is, a tendency for subjects to *conditionally* behave as “it is thought” that most players in the same role would behave. As mentioned above, if the actual behavior of one’s peers is unobservable, then it is likely that one’s first- and second-order beliefs about the peers’ behavior will coincide. However, while purely conformist subjects will necessarily update their second-order beliefs upon receiving the exogenous information, self-servingly conformist subjects will update their beliefs as long as it is convenient to them. In brief – in the event of any difference between *first-* and *induced second-order* beliefs (about the peers’ behavior) – self-servingly conformist subjects will choose to act in accordance with the belief that best serves their interests.

Formally, consider strategies  $s_i$  and  $s'_i$ , and prior and transmitted beliefs  $\gamma_i(s_i)$  and  $\bar{\gamma}_i(s'_i)$ , respectively: self-servingly conformist individuals will “adopt” the transmitted belief  $\bar{\gamma}_i(s'_i)$ , if and only if they are better off by doing so. In Appendix A we propose an illustrative psychological term that captures this motivation, but for the purposes of the current analysis it will suffice to say that – in the case of Trustees – self-serving conformism is characterized by the following relationships:

---

matching the) exogenous belief about the Trustors’ modal behavior. Note that it is reasonable to conjecture that such a psychological term – via  $\bar{\gamma}_1(a)$  – may not be decisive in driving Trustors’ behavior because, whenever  $\bar{\gamma}_2(c) > 33\%$ , then the expected value of the Trustor’s material payoff from choosing  $a$  is greater than that from  $b$ , for any value of  $\bar{\gamma}_1(a)$ . Nevertheless, the fact that a Trustor makes use of the (previously received) exogenous belief  $\bar{\gamma}_2(c)$  so as to calculate her expected material payoff is itself consistent with conformist preferences.

$$\begin{aligned} \bar{\gamma}_2(s'_2) \Rightarrow s'_2 \text{ if } \gamma_2(c) > 50\%, \\ \text{or} \\ \gamma_2(s_2) \Rightarrow s_2 \text{ if } \gamma_2(c) < 50\%. \end{aligned} \tag{8}$$

In sum, a self-servingly conformist individual will follow the transmitted belief  $\bar{\gamma}_i(s'_i)$  if and only if it points out to a behavior that yields a higher payoff (than the one the individual herself had anticipated the peers would get).

**Proposition 5.** Self-serving conformism implies that  $\bar{\gamma}_2(s'_2)$  and  $s'_2$  are positively correlated iff  $\gamma_2(c) > 50\%$ .

We will further expand on the notion of self-serving conformism in the following sections,<sup>28</sup> for the moment we limit ourselves to deriving a clear-cut hypothesis that is consistent with such a notion.

**Hypothesis 3 (Self-serving conformism).** Manipulating  $\bar{\gamma}_2(s_2)$  has an impact on  $s_2$  conditional on holding prior belief  $\gamma_2(c) > 50\%$ .

## V. Results

### 1. Summary statistics and models

Table 1 summarizes the data for the decision to cooperate and the stated belief.

---

<sup>28</sup> It should be noted that – in the case of Trustors – the effect of the exogenous belief (about the Trustors' modal behavior) on behavior is symmetrical; that is, self-serving conformism implies that  $s'_1$  and  $\bar{\gamma}_1(s'_1)$  are positively correlated if  $\gamma_1(a) < 50\%$ . In fact, it is reasonable to assume that Trustors who think it normal not to cooperate (holding low priors about cooperation, i.e.  $\gamma_1(a) < 50\%$ ) would not cooperate themselves, *unless* they received an exogenous belief signaling that it is instead considered normal to cooperate (i.e.  $\bar{\gamma}_1(a) > 50\%$ ): in that case they would find it convenient to adopt the exogenous belief, and would choose to cooperate instead. On the other hand, Trustors who think it normal to cooperate (holding high priors about cooperation, i.e.  $\gamma_1(a) > 50\%$ ) would more likely cooperate, in the hope of getting a material payoff of 3 instead of 1: so, if they received an exogenous belief signaling that it is not considered normal to cooperate, they would not want to change their beliefs and would cooperate anyway. Later on we will provide a more detailed argument supporting this point.

Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	110	.5818	.4955	0	1
$\gamma_i(\textit{share})$	110	44.43	24.91	0	99
<i>in</i>	110	.6818	.4679	0	1
$\gamma_i(\textit{in})$	110	59.3	27.35	0	100

**Table 1** - Summary statistics: m and M indicate the minimum and maximum values, respectively.  $\gamma_i(\textit{share})$  and  $\gamma_i(\textit{in})$  denote a subject's stated guess about the percentage of other participants that will choose *share* and *in*, respectively.

Note that the decision to cooperate is dichotomous, taking on value 1 when a subject chooses *c* (i.e. the Trustee *shares*, in Part I) or *a* (i.e. the Trustor *opts in*, in Part II), and taking on value 0 otherwise. Also note that the stated beliefs  $\gamma_i(\cdot)$  are expressed as percentages. It should be highlighted that – while the majority of both Trustees and Trustors cooperated (58% and 68%, respectively) – beliefs about the other subjects go, on average, in opposite directions: specifically, Trustees expected on average that most other Trustees would *not* cooperate, while Trustors expected on average that most other Trustors *would* cooperate (44% and 59%, respectively). To put it into perspective, this means that 68% of Trustors cooperated and subjects expected, on average, 44% of Trustees to cooperate back.

Denote by  $d^H \bar{\gamma}_i(\cdot)$  a *dummy for the transmitted belief of low/high cooperation* taking on value 1 if, for example,  $\bar{\gamma}_i(\textit{share}) > 0.5$  (that is, if  $\bar{\gamma}_i(\textit{share}) \sim 0.75$ ). Now, in order to start exploring the relationship between beliefs and behavior, column 1 of Table 2 presents the following probit regressions:

- the model in the top panel of column 1 consists of the Trustee's decision as the dependent variable, and of the Trustee's *stated belief*  $\gamma_i(\textit{share})$  as predictor for Part I;
- the model in the bottom panel of column 1 consists of the Trustor's decision as the dependent variable, and of the Trustor's *stated belief*  $\gamma_i(\textit{in})$  as predictor for Part II.

	(1)	(2)	(3)	(4)	(5)
<i>share</i>					
$\gamma_i(\text{share})$ : prior about Trustees	.023*** (.005)		<b>.015**</b> (.006)		
$d^H \bar{\gamma}_i(\text{share})$ : belief transmitted Pt I		.077 (.242)	<b>-.830</b> (.557)		
$\gamma_i(\text{share}) \cdot d^H \bar{\gamma}_i(\text{share})$ : interact. I			<b>.029**</b> (.013)		
<i>constant</i>	-.777*** (.265)	.166 (.173)	<b>-.586</b> (.361)		
Pseudo R2	0.128	0.000	<b>0.167</b>		
AIC	134.398	153.429	<b>132.523</b>		
Obs.	110	110	<b>110</b>		
<i>in</i>					
$\gamma_i(\text{in})$ : prior about Trustors	.029*** (.005)		.039*** (.008)		<b>.040***</b> (.009)
$d^H \bar{\gamma}_i(\text{in})$ : belief transmitted Pt II		.190 (.251)	1.257* (.672)		<b>1.303*</b> (.756)
$\gamma_i(\text{in}) \cdot d^H \bar{\gamma}_i(\text{in})$ : interact. II			-.019* (.011)		<b>-.021*</b> (.011)
$\gamma_i(\text{share})$ : prior about Trustees				.014*** (.005)	<b>.0182***</b> (.006)
$d^H \bar{\gamma}_i(\text{share})$ : belief transmitted Pt I					<b>.638**</b> (.321)
<i>constant</i>	-1.142*** (.314)	.382** (.171)	-1.821*** (.514)	-.138 (.254)	<b>-2.903***</b> (.751)
Pseudo R2	0.242	0.004	0.268	0.054	<b>0.346</b>
AIC	108.180	141.023	108.660	134.077	<b>101.886</b>
Obs.	110	110	110	110	<b>110</b>

**Table 2** - Probit regression coefficients: in brackets are robust standard errors (\*, \*\*, and \*\*\* indicate  $p < 0.10$ ,  $p < 0.05$  and  $p < 0.01$ , respectively, for the relevant Z Statistic). The top panel refers to the Trustee's decision (*i.e.* Part I), whereas the bottom panel refers to the Trustor's decision (*i.e.* Part II).

Column 1 seems to indicate a strongly significant positive effect of prior beliefs on behavior. Yet, given that we cannot establish the direction of the causal relationship by using the model in column 1, we cannot accept or reject any hypothesis.



Hence, we move on to the analysis of the effect of introducing an exogenous variation in beliefs. The model in the top panel of column 2 consists of the Trustee's decision as the dependent variable, and of the Trustee's *transmitted belief* dummy  $d^H \bar{\gamma}_i(\text{share})$  as predictor. Similarly the model in the bottom panel of column 2 consists of the Trustor's decision as the dependent variable, and of the Trustor's *transmitted belief* dummy  $d^H \bar{\gamma}_i(\text{in})$  as predictor. Column 2 seems to indicate a non-significant positive effect of transmitting beliefs on behavior, which – taken together with the figures in column 1 – seems to suggest that there is evidence of consensus *and not* of pure conformity. Yet, given that there could potentially be interaction effects between *stated* beliefs and low/high *transmitted* beliefs, controlling for any such interaction might elucidate the analysis.

Thus, the model in the top panel of column 3 consists of the Trustee's decision as the dependent variable, and of the following predictors: (i) the Trustee's *stated belief*, (ii) the *transmitted belief* dummy, and (iii) their *interaction* (i.e.  $\gamma_i(\text{share}) \cdot d^H \bar{\gamma}_i(\text{share})$ ).<sup>29</sup> It turns out that, while the transmitted belief dummy is non-significant, both the stated belief and the interaction variable are significant at the 5%-level: this suggests that *transmitting information has an effect on Trustees' behavior, conditional on their priors*. Most importantly, comparing the values of the Akaike information criterion – AIC – for models 1-3 of Table 2 (*top panel* only), we can conclude that the best specification is that of model 3, having the minimum value of 132.523.<sup>30</sup> This provides evidence in support of our self-serving conformism hypothesis.

Likewise, we move on to analyze the effect of an exogenous variation in beliefs on Trustors' behavior, taking care to control for interaction effects between stated beliefs and low/high transmitted beliefs. The model in the bottom panel of column 3 consists of the Trustor's decision as the dependent variable, and of the Trustor's *stated belief*, *transmitted*

---

<sup>29</sup> Note that the inclusion of the *stated belief* variable is justified by the non-significant negative correlation between the stated beliefs and the transmitted belief dummy (correlation coefficient of  $-0.145$ ,  $p = 0.130$ ). In fact, even a very slight co-variation between stated and transmitted beliefs can confound the estimate of the effect of exogenously transmitting beliefs, unless one controls for any interaction effects.

<sup>30</sup> See Burnham and Anderson [2002] for a discussion of model selection techniques.

*belief* dummy and their *interaction* as predictors. Here both the transmitted belief and the interaction variable are significant, but this time only at the 10%-level. In this respect it seems that influencing the behavior of Trustees with high priors is easier than influencing the behavior of Trustors with high priors. Later on we will provide a detailed argument supporting this point.

We proceed to discuss a couple of alternative specifications for models predicting Trustors' behavior. The model in column 4 captures the reasoning of subjects with standard, rational self-centered preferences: in fact, if Trustors are selfish utility-maximizers, then their behavior should be well predicted by their beliefs about the opponents' behavior. It should be recalled that, in Part I of the experiment, we elicited Trustees' beliefs about the behavior of other Trustees; it follows that – from the viewpoint of a subject taking a decision in Part II – the beliefs we elicited in Part I now represent Trustors' expectations about the opponents' behavior. So, column 4 seems to indicate a strongly significant positive effect of *prior beliefs about the opponents' behavior* on own behavior. Yet, if one compares the values of the Akaike information criterion for models 1-4 of Table 2 (*bottom panel* only), then model 4 should be discarded. To sum up, the relatively high value of AIC for model 4 points out in the direction of some conformity motive.

Lastly, the model in column 5 regresses a Trustor's decision on the following variables: (i) the stated belief about other Trustors' behavior; (ii) the transmitted belief dummy *about other Trustors' behavior*; (iii) the interaction between i and ii (*i.e.*  $\gamma_i(in) \cdot d^H \bar{\gamma}_i(in)$ ); (iv) the stated belief about Trustees' behavior; (v) the transmitted belief dummy *about Trustees' behavior*. In brief, the transmitted belief dummy about Trustees' behavior turns out to be significant at the 5%-level,<sup>31</sup> while the rest of the coefficients remain very close to those of models 3 and 4. Finally, comparing the values of the Akaike information

---

<sup>31</sup> It should be stressed that the fact that the transmitted belief about Trustees' behavior is significant is itself consistent with conformist preferences, since there is no reason (other than some form of conformity) for a subject to rely on other participants' beliefs about the opponents' behavior. In fact, the structure of the game was such that no participant had access to private signals, as instead is the case in informational cascade models. See also footnote 27.

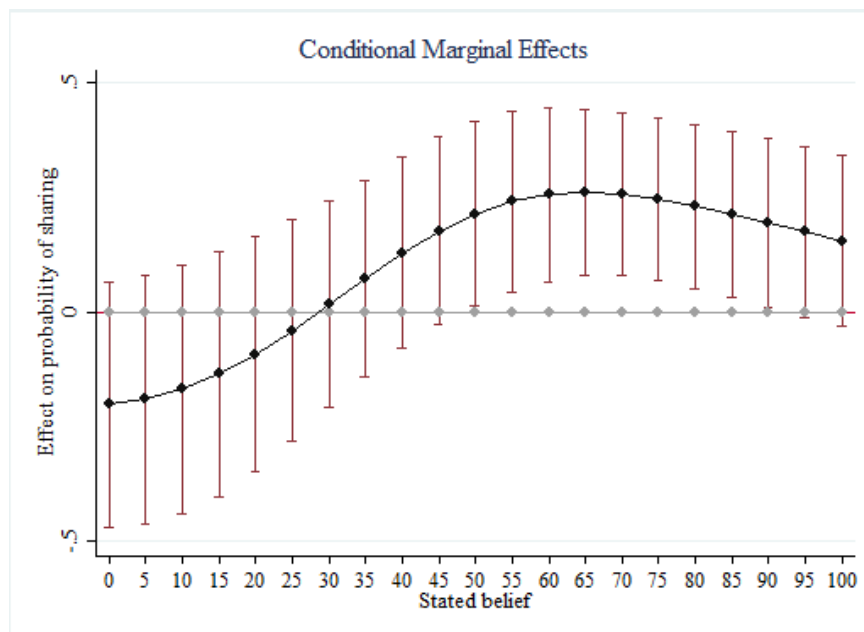
criterion across all (bottom) models of Table 2, we can conclude that the best specification is that of model 5, having the minimum value of 101.886. Again, we take it as evidence in support of our self-serving conformism hypothesis.

## 2. Size of impact and interpretation

In what follows we shall provide an interpretation of the above results while focusing on the models singled out by the Akaike information criterion. After we will have analyzed the effect of the treatment on behavior, we will move on to examine the effect that transmitting beliefs in Part I has on the way subjects form beliefs in Part II.

### 2.a Treatment effect on behavior

**Analysis of Trustees' behavior.** First, because the impact of exogenously transmitting beliefs varies with a Trustee's prior belief, it is useful to compute the discrete change in predicted *sharing* (for the transmitted belief  $d^H \bar{\gamma}_i(\text{share})$ ) at each value of the stated belief  $\gamma_i(\text{share})$ . Figure 2 below graphs such differences, along with 95%-confidence intervals.



**Figure 2** - Analysis of the Trustee's decision. The horizontal axis measures values of the *stated belief*  $\gamma_i(\text{share})$  in 5% increments. The vertical axis measures the discrete change from the base level (*i.e.* from  $d^H \bar{\gamma}_i(\text{share}) = 0$ ), with 95%-confidence intervals.

We can conclude that, for values of the prior belief greater than 50% (*i.e.* for those Trustees who believed that most other Trustees would cooperate), the marginal effect of transmitting some peers' beliefs was positive and significant at the 5%-level.<sup>32</sup> On the other hand, for values of the prior belief less than 50%, the marginal effect of transmitting some peers' beliefs was non-significant.<sup>33</sup> Once again this clearly supports the notion of self-serving conformism, which suggests that – when conformist individuals are uncertain as to which behavior is currently considered normal – they will choose to take the action that best serves their interests. In summary, self-servingly conformist individuals will follow the peers' expectations, only if that is convenient to them. For Trustees with high priors it is indeed convenient to conform to the low transmitted belief – if they receive it – because their expected payoff in the case of conformity to some regular pattern of cooperation is 3, whereas their expected payoff in the case of a pattern of non-cooperation is 6.<sup>34</sup> So, if they get a low transmitted belief, they will consistently adjust their behavior. Conversely, for Trustees with low priors it is not convenient to adopt an eventual high transmitted belief.

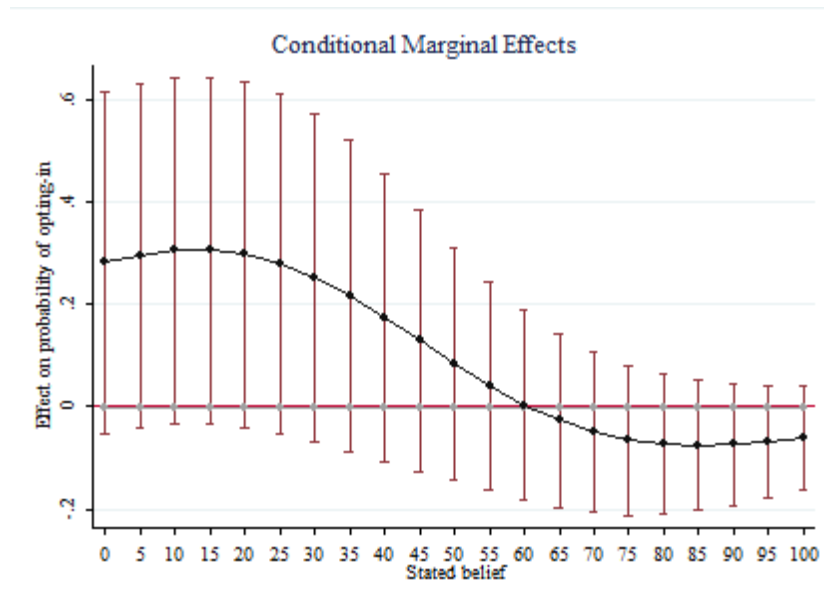
**Analysis of Trustors' behavior.** We move on to analyze the impact of transmitting beliefs on the Trustor's decision, using model 5 of Table 2 above. Again, it is useful to compute the discrete change in predicted *opting-in* (for the transmitted belief  $d^H \bar{\gamma}_i(in)$ ) at each value of the stated belief  $\gamma_i(in)$ , while keeping the other predictors at their mean values. Figure 3 below graphs such differences, along with 95%-confidence intervals.

---

<sup>32</sup> One exception is that the effect was positive and significant at the 10%-level for individuals whose prior beliefs approached unity (*i.e.*  $\gamma_i(share) \geq 95\%$ ).

<sup>33</sup> More precisely, there was a negative non-significant effect for individuals whose stated beliefs were less than 30%, and a positive non-significant effect for individuals with stated beliefs within the range 30%-40%.

<sup>34</sup> According to the norms literature, in order to determine that a *social* norm is at work one would need to elicit both empirical and normative expectations, and verify that they are consistent with the observed behavior. We chose to focus on the empirical beliefs in order for us to establish a direct parallel between our investigation and previous studies whose results may have been driven by consensus effects. We believe that our results are sufficient to demonstrate that conformity is present (whereas we do not claim that this is necessarily a case of conformity to a “social norm”; see Bicchieri [2006], Sontuoso [2013], and Bicchieri and Sontuoso [2015]).



**Figure 3** - Analysis of the Trustor's decision. The horizontal axis measures values of the *stated belief*  $\gamma_i(in)$  in 5% increments. The vertical axis measures the discrete change from the base level, along with 95%-confidence intervals, while the other predictors are being held at their mean values (*i.e.* the change from  $d^H \bar{\gamma}_i(in) = 0$ , when  $\gamma_i(share) = 44.43\%$  and  $d^H \bar{\gamma}_i(share) = 0.5$ ).

Figure 3 seems to suggest that Trustors who held a low prior belief were, on average, positively affected by the transmitted belief: it should be noticed that Figure 3 looks somewhat symmetrical to Figure 2 above. The data show that, for those Trustors who believed that few other Trustors would cooperate (*i.e.* for values of the prior belief less than 30%), the marginal effect of transmitting some peers' beliefs was positive and significant at the 10%-level. On the other hand, for relatively high values of the prior belief, the marginal effect of transmitting some peers' beliefs was non-significant.<sup>35</sup> Once again this is due to the presence of self-serving conformist preferences, which in this case work in the opposite direction to that of Trustees' preferences. Indeed, for Trustors with high priors it is not convenient to switch to the low transmitted belief – if they receive it – because their expected payoff in the case of conformity to some regular pattern of (mutual) cooperation is 3, whereas

---

<sup>35</sup> More precisely, there was a positive non-significant effect for individuals whose stated beliefs were within the 30%-60% range, and a negative non-significant effect for individuals with stated beliefs greater than 60%.

their expected payoff in the case of a pattern of non-cooperation is 1. So, if they get a low transmitted belief, they will choose to disregard it. Conversely, for Trustors with very low priors it might be convenient to switch to an eventual high transmitted belief, since they would interpret it as a signal that it is considered normal to cooperate; and, therefore, if they get a high transmitted belief they will consistently adjust their behavior.

Before proceeding, it is important to stress that our analysis of Trustors' behavior has focused on the marginal effects of manipulating the beliefs about other Trustors' behavior. We believe that such effects are weaker than are the respective effects for Trustees (as shown in Figure 2), because of the small difference in payoff between *opting out* (*i.e.* \$1) and the worst-case scenario that can occur after a Trustor has *opted in* (*i.e.* \$0); more explicitly, we believe that such a small payoff difference may have encouraged Trustors to take the risky option no-matter-what, thereby weakening the effect of transmitting beliefs *about other Trustors' behavior*.<sup>36</sup> That said, we must recall that our best model to predict a Trustor's choice (*i.e.* model 5, Table 2) presents also a strongly significant effect of transmitting beliefs *about Trustees' behavior*, which is itself consistent with some conformity hypothesis.<sup>37</sup> In this connection, we shall now move on to the analysis of the effect that transmitting beliefs in Part I has on the way subjects form beliefs in Part II.

## **2.b A further test of the consensus effects hypothesis**

In this subsection we aim at providing additional evidence in favor of the assumption that consensus is *not the only* force driving a belief-behavior correlation in our trust game. We will consider the null hypothesis that Trustors are affected solely by consensus, and adapt

---

<sup>36</sup> Another possible explanation for the lower impact of feedback on Trustors' behavior might be due to the fact that – as Bicchieri *et al.* [2011] argue – “Trustworthiness is a social norm, but trusting is not”, in that the authors find that experimental subjects do not view trust as being normative, whereas they view reciprocation as normative in trust games.

<sup>37</sup> Note that Ellingsen *et al.*'s [2010] focus was on Trustees' rather than on Trustors' behavior, so we cannot compare results about Trustors' behavior. On the other hand, the strongly significant effect of exogenously varying Trustees' beliefs, conditional on priors, clearly suggests that self-serving conformism is at least part of what has driven the correlation between beliefs and behavior in previous studies of Trustees' behavior.

Ellingsen *et al.*'s [2010] argument as follows.<sup>38</sup> In Part II of the experiment each Trustor believes that Trustees generally behave as she herself did in Part I (when she acted as a Trustee). So, *if a Trustor is affected solely by consensus*, then the Trustor's behavior must be correlated with the very individual's behavior in Part I; further, she will believe that other Trustors must behave like her (see Proposition 3 above).

If that is the case, the following must be true: consider an individual in Part II of the experiment (*i.e.* a Trustor); the individual's belief about other Trustors' behavior must be correlated with the action the very individual took in Part I. So, *divide subjects according to the action taken* in Part I (as a Trustee), and consider each group of subjects in turn: here, within each group, a Trustor's belief about other Trustors' behavior must *not* vary with the belief we exogenously transmitted in Part I (if subjects are affected solely by consensus). In this respect, we note that there is by definition no reason other than some form of conformity for some Trustees' beliefs to influence a Trustor's beliefs about some other Trustees' behavior.<sup>39</sup>

Hence, for each of the two groups, the null hypothesis is that the mean of a Trustor's prior belief (about other Trustors' behavior) is the same whether the subject – when playing as a Trustee – got a high transmitted belief or not. Formally, the null hypotheses for the two groups are:

- a) for Trustees who did *not* cooperate (*i.e.*  $share = 0$ ),  
 $\gamma_i(in)(given\ d^H\bar{\gamma}_i(share) = 0) = \gamma_i(in)(given\ d^H\bar{\gamma}_i(share) = 1)$ ;
- b) for Trustees who *did* cooperate (*i.e.*  $share = 1$ ),  
 $\gamma_i(in)(given\ d^H\bar{\gamma}_i(share) = 0) = \gamma_i(in)(given\ d^H\bar{\gamma}_i(share) = 1)$ .

---

<sup>38</sup> Ellingsen *et al.*'s argument «utilizes the consensus effect twice: Trustees believe that other trustees *behave* likewise and that trustors *think* likewise» (Ellingsen *et al.* [2010], footnote 4, p. 96, italics in original). See also our footnote 26 above.

<sup>39</sup> On the other hand, a Trustor would have reason to take into account the Trustees' beliefs about other Trustees' behavior, if the Trustor thought that every other subject *i*'s stated belief was *directly* correlated to *i*'s decision to share. We reckon that such a scenario is fairly improbable as it would require a high level of strategic sophistication.

- a) For the group of subjects who did not cooperate in Part I, the Wilcoxon-Mann-Whitney test suggests that there is *not* a statistically significant difference between the underlying distributions of Trustors' beliefs, conditional on receiving or not a high transmitted belief (as a Trustee), that is,  $z = -0.803$ , with  $p = 0.422$ .
- b) For the group of subjects who cooperated in Part I, the Wilcoxon-Mann-Whitney test suggests that there is *indeed* a statistically significant difference between the underlying distributions of Trustors' beliefs, conditional on receiving or not a high transmitted belief (as a Trustee), that is,  $z = -2.273$ , with  $p = 0.023$ .

This means that subjects who cooperated as Trustees are most likely to *take into account the information we provided* in Part I, *when forming a belief in the role of Trustors* (in Part II): we believe that this pattern is consistent with a conformist attitude, and therefore we reject the hypothesis that Trustors are affected solely by consensus.

Why is it the case that only subjects who cooperated as Trustees are likely to take into account the information we provided in Part I? Recall that, as shown in Figure 2 above, Trustees with low priors would just disregard any transmitted beliefs *by not cooperating*. This implies that if one is looking out for potentially conformist individuals in Part II, then one will more likely find them among those subjects who cooperated in Part I.

### ***2.c Disentangling explanations***

Having provided conclusive evidence in support of our self-serving conformism hypothesis, we shall now turn to compute a crude measure of the strength of the *consensus effect* relative to that of *pure conformity*.

To that end it is useful to review the implications of the consensus effect; that is, if consensus is present, then there will be a *positive correlation* between own behavior and own beliefs about the strategy taken by subjects in the same role. Although we are not directly able to infer the direction of causality underlying this particular correlation, if one is to detect the individuals who are affected by consensus effects, Proposition 3 above implies this tentative identification strategy: consider an individual  $i$  who chooses a certain action with a higher probability than the average player ( $g$ ) does; then, individual  $i$  is affected by consensus effects if  $i$  gives a higher *estimate of some player  $j$  choosing that action* than the estimate given by the average player ( $g$ ).



For the purposes of this exercise it is important to note that the average stated belief was 52.87% for the Trustees who chose to share while it was 32.69% for the Trustees who chose not to share. Further, the average stated belief was 69.2% for the Trustors who chose to opt in, whereas it was 38.08% for the Trustors who chose not to opt in. Now, in order to identify all Trustees affected by consensus effects, we will have to look at subjects who cooperated and held a prior  $\gamma_i(\text{share}) > 0.32$ ; also we will have to look at subjects who did not cooperate and held a prior  $\gamma_i(\text{share}) < 0.52$ . This gives us a total of 83 Trustees (75.45%) affected by consensus effects. Similarly, in order to identify all Trustors affected by consensus effects, we will have to look at subjects who cooperated and held a prior  $\gamma_i(\text{in}) > 0.38$ ; also we will have to look at subjects who did not cooperate and held a prior  $\gamma_i(\text{in}) < 0.69$ . This gives us a total of 95 Trustors (86.36%) affected by consensus effects.

It should be noted that the above figures do not give us an idea of the real magnitude of the consensus effect relative to that of pure conformity, since those figures also account for subjects who are affected by both consensus and *some form* of conformist preferences. Therefore – in order to tell consensus effects from pure conformity – we shall now focus on Trustees and, in particular, on *Trustees engaging in cooperative behavior*. Given that, we propose to identify four categories of subjects:

- i) those who are affected by consensus effects (*and not* by any form of conformity);
- ii) those who are affected by pure conformity (*and not* by self-serving conformism *nor* consensus effects);
- iii) those who are affected by *either* consensus *or* pure conformity *or* both (*and not* by self-serving conformism);
- iv) those who are affected by *neither* consensus *nor* any form of conformity.

Table 3 below summarizes the data for the four categories of interest, with respect to Trustees engaging in cooperative behavior.

	Explanation	Proportion of the subjects who shared
i)	<i>Consensus only</i>	34.37%
ii)	<i>Pure conformity only</i>	15.62%
iii)	<i>Consensus or pure conformity</i>	10.94%
iv)	<i>Neither consensus nor (any) conformity</i>	12.50%

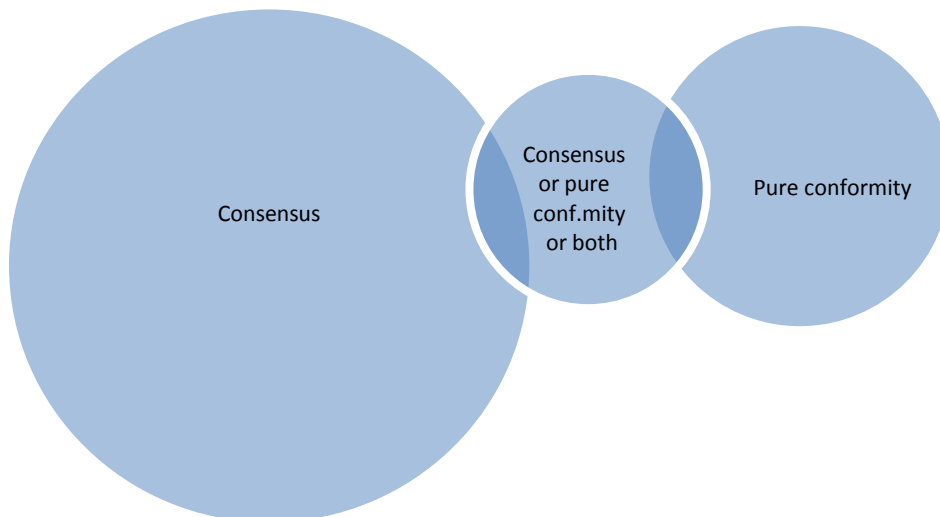
**Table 3** - Relative strengths of consensus and pure conformity, with reference to Trustees' cooperative behavior. (Note that the remaining 26.57% of subjects who shared in Part I includes individuals affected by consensus or some form of conformity or both.<sup>40</sup>)

Note that the Trustees belonging to class *i*) can be identified with those participants who held a prior  $\gamma_i(\text{share}) > 0.32$ , and were transmitted a low belief (*i.e.*  $d^H \bar{\gamma}_i(\text{share}) = 0$ ). The Trustees belonging to class *ii*) can be identified with those participants who held a prior  $\gamma_i(\text{share}) \leq 0.32$ , and were transmitted a high belief. The Trustees belonging to class *iii*) can be identified with those participants who held a prior  $0.32 < \gamma_i(\text{share}) < 0.50$ , and were transmitted a high belief.<sup>41</sup> Lastly, the Trustees belonging to class *iv*) can be identified with those participants who held a prior  $\gamma_i(\text{share}) \leq 0.32$ , and were transmitted a low belief. The frequencies of the Trustees belonging to classes *i*)-*iii*) are summarized in the below bubble chart.

---

<sup>40</sup> Specifically, 26.57% accounts for individuals who held a prior  $\gamma_i(\text{share}) > 0.50$  and were transmitted a high belief.

<sup>41</sup> It is clear that these subjects' behavior is consistent with consensus effects or pure conformity or both (but not self-serving conformism). Think of a subject using her experience about the "normal behavior" she observed in some past interactions in order to guide her current behavior *and* to inform her estimation of the current peers' behavior; the transmitted information will then reinforce her convictions.



**Figure 4** - Consensus vs. pure conformity with respect to Trustees' cooperative behavior.

Before proceeding to some robustness checks, we note that Table 3 above suggests that a large fraction of the data on *Trustees' cooperative behavior* is consistent with the notions of consensus and pure conformity, with consensus being the stronger effect.<sup>42</sup> Finally we note that the fraction of behavior consistent with neither consensus nor any form of conformity is quite small (*i.e.* 12.50%).

---

<sup>42</sup> Given that, in the case of Trustees, self-serving conformism is most observable when subjects do not cooperate, the analysis of the relative strengths of consensus and pure conformity – with reference to *non-cooperators* – shows that almost all behaviors are consistent with self-serving conformism. For example, note that most of those Trustees who did not cooperate, held a prior  $\gamma_i(\text{share}) < 0.52$  and were transmitted a high belief (*i.e.*  $d^H \bar{\gamma}_i(\text{share}) = 1$ ), might have been affected by both consensus and self-serving conformism. Similarly, all those Trustees who did not cooperate, held a prior  $\gamma_i(\text{share}) \geq 0.52$  and were transmitted a low belief (*i.e.*  $d^H \bar{\gamma}_i(\text{share}) = 0$ ), might have been affected by both pure conformity and self-serving conformism. An analogous argument applies to class *iii*). In short, 93.5% of the non-cooperative behavior of Trustees is consistent (also) with self-serving conformism. On a different note, as regards Part II, the analysis of such relative effects on Trustors presents additional complexities in that one would need to account also for the beliefs transmitted in Part I.

## VI. Robustness checks

In order to make sure that neither the rate of cooperative behavior nor the beliefs about the others' behavior vary *if* beliefs are incentivized, we ran a few sessions that implemented slightly modified experimental designs. That is, a *control treatment* ("T0") was designed to be identical to our main treatment above, except that there was no belief transmission at all. Then, a *belief-incentivizing treatment* ("T1") was designed to be identical to T0, except for an incentivizing scheme for the elicitation of beliefs.

Specifically, given that a crucial element of conformist preferences involves expectations – in order to induce participants to state their true beliefs – a few sessions presented the following incentivizing scheme. As usual, before entering their decisions, each subject was asked to guess how many of the other participants (in the same role) would choose each action: yet in T1 subjects were also told that they would receive an additional payment of £2 if their guess differed by no more than 5 percentage points from the realized value.<sup>43</sup> Table 4 summarizes the data for the decision to cooperate and the stated belief with reference to each treatment, T0 and T1.

---

<sup>43</sup> If a subject's guess differed by more than 5 percentage points from the realized value, that subject would receive no additional belief payment; note that Charness and Dufwenberg [2006] implemented the same incentivizing scheme (to elicit subjects' beliefs about the opponents). Each subject was not informed about the correctness of her guesses until the end of Part II.

Treatment T0					
Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	53	.509	.504	0	1
$\gamma_i(\textit{share})$	53	38.622	25.979	0	99
<i>in</i>	53	.641	.484	0	1
$\gamma_i(\textit{in})$	53	50.415	30.917	0	100
Treatment T1					
Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	46	.586	.497	0	1
$\gamma_i(\textit{share})$	46	45.695	25.488	0	100
<i>in</i>	46	.608	.493	0	1
$\gamma_i(\textit{in})$	46	62.543	25.844	5	100

**Table 4** - T0 and T1 summary statistics: m and M indicate the minimum and maximum values, respectively.  $\gamma_i(\textit{share})$  and  $\gamma_i(\textit{in})$  denote a subject's stated guess about the percentage of other participants that will choose *share* and *in*, respectively.

We can now proceed to check whether the rate of cooperative behavior or the beliefs about the others' behavior vary when beliefs are incentivized. We consider *beliefs* first: for the Trustee's decision (*i.e.* Part I), the null hypothesis is that the mean for  $\gamma_i(\textit{share})$  is the same for T0 and T1; for the Trustor's decision (*i.e.* Part II), the null hypothesis is that the mean for  $\gamma_i(\textit{in})$  is the same for T0 and T1. Thus, as for Part I, the Wilcoxon-Mann-Whitney test suggests that there is not a statistically significant difference between the underlying distributions of  $\gamma_i(\textit{share})$  for T0 and T1 ( $z = -1.296$ ,  $p = 0.195$ ). As regards Part II, the Wilcoxon-Mann-Whitney test suggests that there is mild evidence against the null hypothesis ( $z = -1.827$ ,  $p = 0.067$ ). In light of this result, the analysis of the *decision to cooperate* (*i.e.* a test of whether offering to pay for beliefs affected behavior or not) becomes critical.

Hence, for the Trustee's decision, the null hypothesis is that the mean for *share* is the same for T0 and T1; for the Trustor's decision, the null hypothesis is that the mean for *in* is the same for T0 and T1. As for Part I, the Wilcoxon-Mann-Whitney test suggests that there is not a statistically significant difference between the underlying distributions of *share* for T0 and T1 ( $z = -0.769$ ,  $p = 0.442$ ). More interestingly, as regards Part II, the Wilcoxon-

Mann-Whitney test suggests that there is no evidence against the null hypothesis ( $z = 0.335$ ,  $p = 0.737$ ). We can conclude that offering or not-offering to pay for beliefs is very unlikely to affect behavior or beliefs (incentivizing beliefs may, at most, induce Trustors to overestimate the fraction of the other Trustors that will choose to opt in).

Finally, if neither offering (in T1) nor not-offering (in T0) to pay for beliefs changes behavior – when subjects are *not* shown an aggregate measure of the others' beliefs – then one can reasonably assume that offering or not-offering to pay for beliefs will not change behavior even when subjects *are* shown an aggregate measure of the others' beliefs. This corroborates the findings of our main treatment.

## **VII. Concluding remarks**

This essay has defined a class of conformist preferences, and presented a test to determine if there are causal effects of second-order beliefs on behavior. While it is true that a substantial part of the data on Trustees' cooperative behavior can be consistent with the consensus effects hypothesis, as we look at the whole sample of subjects' decisions we find conclusive evidence in support of our *self-serving conformism hypothesis*. Such a hypothesis posits that exogenous beliefs about the peers' behavior conditionally influence one's behavior, with the strength and direction of the impact depending on one's prior beliefs. Indeed, a significant number of our subjects chose to adjust their strategy to the transmitted belief when it was in their interest to do so.

In short, some individuals have a tendency to follow the behavior, attitudes or judgments of others, with the others' observed or purported behavior being considered appropriate or normal within a certain social group. Naturally, situations like mixed-motive games render successful coordination on shared rules of behavior more difficult because of individual incentives to deviate. Nevertheless, especially in small social groups – where the social distance is little and the individual actions more visible – the desire to fit in with the group or to just maintain a nice social image often acts as a counter-incentive to deviating from collective rules of behavior (*e.g.* peer norms). And yet, in cases where there is some uncertainty as to which norm is currently at work, some individuals have been observed to exhibit a tendency to choose to follow the pattern that is more advantageous to them.

In this paper we show that such self-servingly conformist preferences substantially affect decisions in trust games. Once again, the interpretation we propose is the following: when conformist individuals are uncertain as to which behavior is currently considered normal, they will choose to take the action that best serves their interests. In the case of our trust game, “high-prior” Trustees, who got a low transmitted belief, acted as if they convinced themselves that it is considered normal to behave selfishly as Trustees. Instead “low-prior” Trustors, who got a high transmitted belief, acted as if they convinced themselves that it is considered normal to mutually cooperate in a trust game.

It should be highlighted that such a pattern is consistent with a desire to reduce *cognitive dissonance* (Festinger [1957]) or, more specifically, a desire to minimize any difference between one’s behavior and some beliefs about what is right to do. As Rabin noted in a pioneering paper, «[b]ecause it is unpleasant, people prefer to reduce cognitive dissonance. There are two ways to do so. As economists generally assume, people can change their behavior. Or – much less familiar to an economist – people can change their beliefs» (Rabin [1994], p. 178). Rabin went on to model an individual’s difficulty of holding false beliefs with a cost function such that a utility-maximizing individual would trade off his preference for feeling “moral” with the cost of holding false beliefs. Here, instead, one may well assume that a self-servingly conformist individual faces no cost in changing her beliefs whenever she is provided with “more convenient” exogenous beliefs.

To conclude, our analysis makes us believe that part of what has driven the correlation between beliefs and behavior in previous trust game experiments might have been motivated by a *causal relationship running from beliefs to behavior*, thereby implying that consensus is not the only driving force of such correlation. We further believe that our design can determine that the observed causal effects are due to conformist preferences, and not to other belief-dependent motivations like guilt aversion. In fact, our investigation may suggest that the traditional notion of guilt aversion could be seen as a special case of the more general notion of conformist preferences: while players affected by guilt aversion proper adapt their behavior to the expectations of the specific subject they are matched with, conformist preferences imply a tendency for individuals to adapt their behavior so as to be in line with some collective belief about the behavior of others. This tendency is thought to arise from a desire to fit in with the purported majority. Hence – in the case of conformist preferences – a

player is not simply motivated by a desire not to disappoint her matched, *individual* coplayer. Nevertheless, a conformist player would not want to hurt her coplayer *if* the coplayer's expectations were based on a diffused belief about the *majority's* behavior.

## Appendix A

### *Proofs*

**Proof of Proposition 2.** If a Trustee is guilt averse and believes that the opponent expects her to cooperate (*i.e.*  $\beta_2(c) = 1$ ), then she will be inclined to cooperate (provided that her own guilt sensitivity is large enough). That is,  $u_2(c) \equiv 3 > 6 - 3k_2 \equiv u_2(d)$  if  $k_2 > 1$ . Hence (for guilt averse players with sensitivity  $k_2 > 1$ ),  $s_2$  and  $\beta_2(s_2)$  will be positively correlated. Finally, if a subject is motivated only by guilt aversion, then she will be indifferent as to what other players in the same role believe or do; this implies that  $s_i$  and  $\bar{\gamma}_i(s_i)$  are not correlated.

**Proof of Proposition 3.** The causal relationship in expression (6) obviously implies that  $s_2$  and  $\gamma_2(s_2)$  are correlated. Now, assume that an individual – when playing as a Trustee – chooses some strategy, say,  $s_2 = c$  (for whatever reason): because of the consensus effect, upon playing as a Trustor the very individual will be more likely to believe that most Trustees behave as she did, which implies that  $\alpha_1(c) \equiv \gamma_2(c)$ . Here, this in turn implies that that individual will be more likely to choose  $s_1 = a$  (as dictated by material self-interest). Finally note that the consensus effect only entails a causal relationship from own behavior to own beliefs (about others' behavior), and thus there is no correlation between own behavior and *someone else's* beliefs, *e.g.* between  $s_2$  and  $\bar{\gamma}_2(s_2)$ .



**The psychological term for subjects affected by *pure conformity*.**

In order to suggest an example of a specific functional form – and to establish a direct parallel with empirical studies of guilt averse preferences – we will adapt the Trustee’s psychological term from expression (5) above, as follows. That is, we assume the Trustee’s psychological term to be given by the additive inverse of any (positive) difference between the expected value of the Trustor’s payoff, *this time derived from*  $\bar{\gamma}_2$ , and the payoff  $m_1(s_2)$  the Trustor actually ends up with:

$$f(\beta_2(s_2), s_2) = -x_2[\max\{0, E_{\beta_2=\bar{\gamma}_2}[m_1|h^0] - m_1(s_2)\}], \quad (1A)$$

where  $x_2$  denotes a (game-specific) “conformity constant” that is the same for every Trustee of the given game. As it will be clear, for expression (1A) to capture the majority-conformity bias in our trust game *it is assumed that*  $x_2 = 2$  (and that  $\beta_2 = \bar{\gamma}_2$ ). Note that the interpretation of  $x_2$  is quite different from that of the player’s guilt sensitivity (*i.e.*  $k_2$ ) of expression (5): in fact,  $k_2$  captures the extent to which an individual player is guilt averse (hence, the larger is the player’s sensitivity, the more will she be inclined to meet the opponent’s expectations). On the other hand, here  $x_2$  is a constant; that is, it is the game-specific value that makes it possible for an exogenous belief *different from 50%* to actually swing the behavior of all the conformist players who receive it (in our case  $x_2 = 2$ ). Also notice that the fact that the expected value is now calculated with respect to the exogenous belief  $\bar{\gamma}_2(s_2)$  means that the Trustee thinks that she is expected to behave like most other Trustees are thought to be behaving. In brief, in the event of transmission of exogenous beliefs, the purely conformist subject would unconditionally adopt the others’ beliefs (*i.e.* the belief that some other players in the same role hold about the strategy taken by most other players).

**Proof of Proposition 4.** Plugging (1A) into expressions (3) and (4), it follows that  $\bar{\gamma}_2(s_2)$  and  $s_2$  are positively correlated. In fact, if a Trustee is told that  $\bar{\gamma}_2(c) > 50\%$  – which means that some other Trustees believe that most other Trustees will *cooperate* – then she will choose strategy  $s_2 = c$ , since  $u_2(c) \equiv 3 > 6 - 3x_2 \cdot \bar{\gamma}_2(c) \equiv u_2(d)$ , given  $x_2 = 2$ . If, instead, the

Trustee is told that  $\bar{\gamma}_2(c) < 50\%$  – which means that some other Trustees believe that most other Trustees will *defect* – then she will choose strategy  $s_2 = d$ , since now  $u_2(c) \equiv 3 < 6 - 3x_2 \cdot \bar{\gamma}_2(c) \equiv u_2(d)$ , given  $x_2 = 2$ .

**The psychological term for subjects affected by *self-serving conformism*.**

In this case the Trustee’s psychological term can be defined as follows:

$$f(\beta_2(s_2), s_2) = \begin{cases} -x_2[\max\{0, E_{\beta_2=\bar{\gamma}_2}[m_1|h^0] - m_1(s_2)\}] & \text{if } \gamma_2(c) > 50\% \\ 0 & \text{if } \gamma_2(c) < 50\% \end{cases}, \quad (2A)$$

with  $x_2 = 2$ . In plain words, the Trustee’s psychological term is defined to be null whenever one believes that most other Trustees will not cooperate (*i.e.*  $\gamma_2(c) < 50\%$ ); otherwise it is defined as it was defined in (1A). This means that subjects would like to fit in with the group, but they feature a self-serving bias in the belief formation process.

**Appendix B**

***Procedure, experimental instructions and screenshots***

The experiment was run with *zTree* (Fischbacher [2007]) in the ExpReSS Lab at Royal Holloway, University of London, between February and May 2012; subjects were recruited via emails forwarded across all faculties at Royal Holloway. A total of 209 subjects participated in the experiment; each session consisted of one of the three treatments (no subject could participate in more than one session). Each session took around 45 minutes and average earnings were £8 (including a £3 show-up fee), with minimum and maximum payments being £4 and £14, respectively. Paper instructions and transcripts of *zTree* screenshots are shown below (the *main treatment* is labeled as “T2”).

## General instructions for participants

Thank you for participating in this study.

Please note that it is prohibited to communicate with other participants during the experiment. If you have a question once the experiment has begun, please raise your hand and an assistant will come to your desk to answer it. Violation of this rule leads to immediate exclusion from the study and from all payments.

You will never learn the identity of the other participants, neither before nor after the study; and not one of the other participants will learn anything about your identity. Also, no other participant will learn what you earn during the experiment: upon completion of the session, the amount of money you will have earned will be paid out individually and privately. Hence, no other participant will know your choices and how much money you earn in this experiment.

You will receive £3 for participating in this session; additionally you also receive money depending on the decisions made (as described in the next paragraphs).

The experiment consists of two parts ("Part I" and "Part II"), each involving one simple decision task; your payment at the end of the session will be calculated as follows.

Your payment

= £3 (show-up fee) + any amount earned in Part I + any amount earned in Part II

In what follows we describe the procedure for Part I.

## Part I

There are two types of participants, participants "A" and participants "B".

You will be assigned a type and paired with **one other participant** who was assigned another type than you.

This part consists of two steps, which you will perform with the particular participant you are paired with.

Step 1: Participant A must choose between the following two options. The first option ("OUT") gives a payout of £1 to both participants. The second option ("IN") is to instead transfer both pounds (i.e. £2 in total) to participant B and leave further decisions to him/her. If participant A transfers the 2 pounds to participant B, they will be tripled and participant B will receive  $3 \times 2 = 6$  pounds.

Step 2: Only if participant A chooses the second option ("IN"), participant B will then decide if he/she transfers £3 back to participant A and keeps £3 for himself/herself OR if participant B keeps all the £6 for himself/herself.

## Procedure for the two steps

### Step one: Decision of participant A

It is up to participant A to choose one of the 2 options (OUT or IN): EITHER both participants receive £1 each OR the money and further decisions are transferred to participant B.

**If participant A chooses the option OUT**, both of you will receive £1. In this case participant B cannot change the payout allocation and the first part ends.

#### **As a result**

At the end of step one, there are two possible situations.

- If participant A has transferred the £2 to participant B (option IN), participant B has £6 and participant A has nothing.
- If participant A has chosen the option OUT, both of you have £1.

### Step two: Decision of participant B

**If participant A has transferred the money to participant B (option IN)**, then B receives £6 and it is now up to participant B to decide about the distribution of the £6 between the two participants. Participant B can EITHER:

- transfer £3 back to participant A and keep £3 for himself/herself

OR

- keep all the £6 for himself/herself and leave nothing to participant A.

After participant B's decision this part is completed and the earnings for both participants will be determined according to B's decision.

The above information is summarised in the following table:

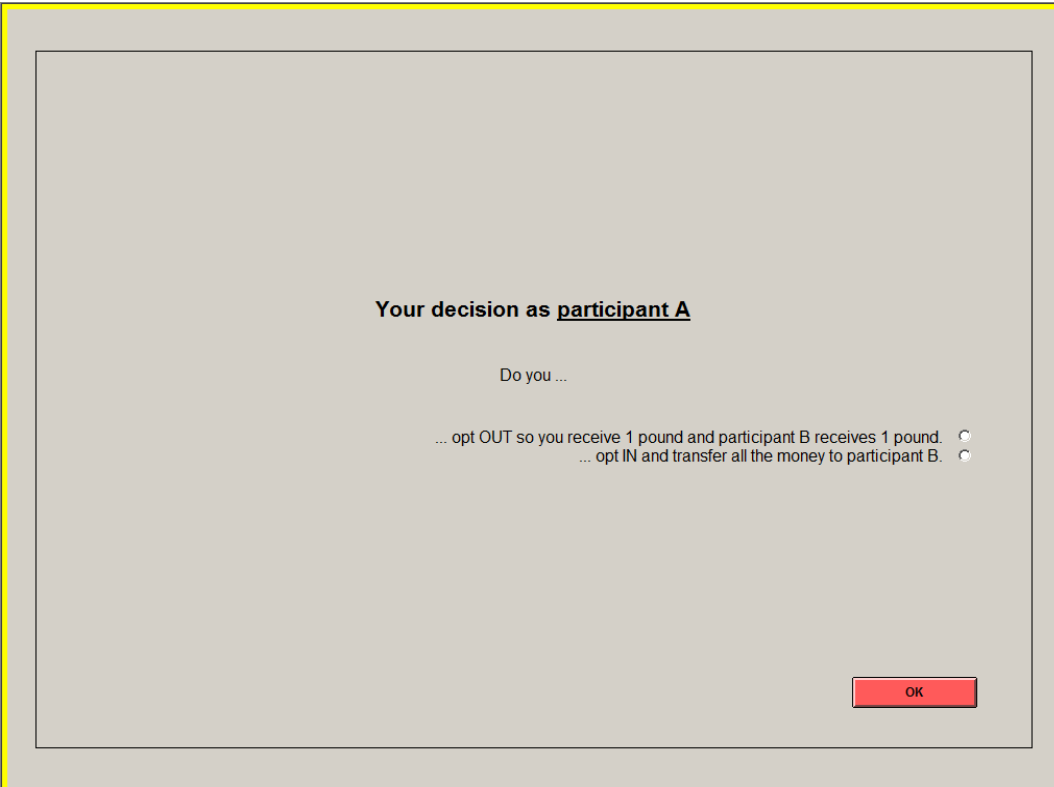
		A's income	B's income
A chose OUT		£1	£1
A chose IN	B keeps all	£0	£6
	B transfers half	£3	£3

## Specific procedure and on-screen instructions for Part I

### You are assigned the role of participant B

Note that you will complete the above-described two steps only once.

Step 1: Participant A decides by entering his/her choice on the screen shown below.



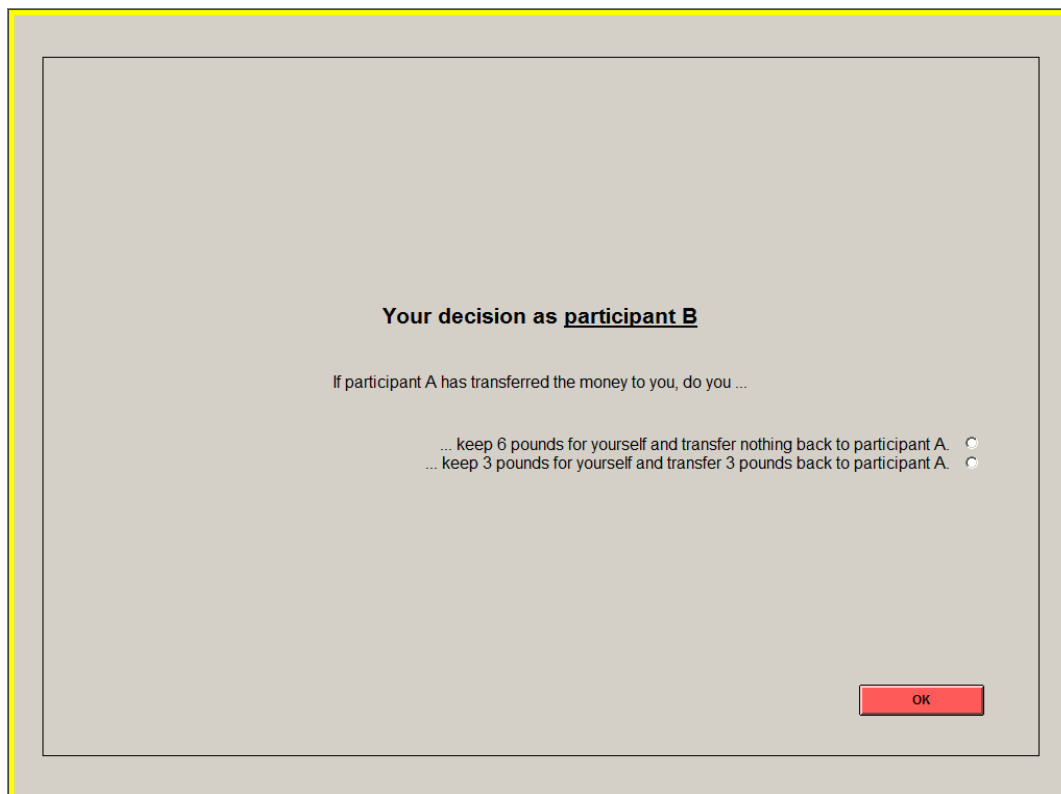
The screenshot shows a grey rectangular window with a yellow border. Inside the window, the text reads: "Your decision as participant A". Below this, it says "Do you...". There are two radio button options: "... opt OUT so you receive 1 pound and participant B receives 1 pound." and "... opt IN and transfer all the money to participant B.". At the bottom right of the window is a red button labeled "OK".

Step 2: We will ask you (participant B) how you would like to divide the £6 between participant A and yourself. Note that your answer will have an effect only if participant A does choose to transfer the money to you (option IN).

Participant A will not know your decision when he/she submits his/her own decision.

As explained above, **you decide on whether to transfer half the money to participant A or keep all the £6 for yourself.**

You will enter your choice on the following screen:



The screenshot shows a grey dialog box with a yellow border. The text inside reads: "Your decision as participant B" followed by "If participant A has transferred the money to you, do you ...". There are two radio button options: "... keep 6 pounds for yourself and transfer nothing back to participant A." and "... keep 3 pounds for yourself and transfer 3 pounds back to participant A.". A red "OK" button is located in the bottom right corner.

## Control questions

Please answer the following control questions. Please contact the study organizer if you have any questions.

1. Participant A has chosen IN. You then choose to transfer half the money back to participant A.

What is the income of participant A? .....

What is the income of participant B (yourself)?.....

2. Participant A has chosen IN. You then choose to keep all the money for yourself.

What is the income of participant A? .....

What is the income of participant B (yourself)?.....

3. Participant A has chosen OUT.

What is the income of participant A? .....

What is the income of participant B (yourself)?.....

Please feel free to ask questions at any point if you feel you need some clarification. Please do so by raising your hand.

We will start with Part I once the instructions are clear to everyone. Are there any questions?



## Part II

We are now ready to undergo the last part of the study. This part has exactly the **same two-step procedure as in Part I.**

The payouts are the same as before and are summarised in the following table:

		A's income	B's income
A chose OUT		£1	£1
A chose IN	B keeps all	£0	£6
	B transfers half	£3	£3

The only difference is that you are assigned a different type in this part than in the previous part.

**You are now assigned the role of participant A.**

Again, you will be paired with one other participant. **This other participant will be a different person than the one you were paired with in Part I.**

Please refer to your paper handout or ask an assistant if you need reminding of the procedure.

## **[Transcript of on-screen messages]**

### **Treatments T0-T1**

Screen 1 (Part I)

#### **You are assigned the role of participant B**

Prior to entering your decision as participant B, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants B).

In other words, we ask you to guess how many of today's participants B (excluding yourself) will choose to transfer half the money back, and how many of today's participants B will keep all the money for themselves.

*Please enter your guess by positioning the below slider to the desired percentage.*

*[The below line is only for treatment T1.]*

Note: You can earn some additional income if your guess is correct. If your guess differs by no more than 5 percentage points from the realized value, at the end of the study you will receive an additional payment of £2. Otherwise, you do not receive an additional income.

Screen 2 (Part I)

*Enter 2<sup>nd</sup> mover decision.*

Screen 3 (Part II)

*Insert instructions for Part II here.*

Screen 4 (Part II)

**You are assigned the role of participant A**

Prior to entering your decision as participant A, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants A).

In other words, we ask you to guess how many of today's participants A (excluding yourself) will choose IN, and how many of today's participants A will choose OUT.

*Please enter your guess by positioning the below slider to the desired percentage.*

[The below line is only for treatment T1.]

Note: You can earn some additional income if your guess is correct. If your guess differs by no more than 5 percentage points from the realized value, at the end of the study you will receive an additional payment of £2. Otherwise, you do not receive an additional income.

Screen 5 (Part II)

*Enter 1<sup>st</sup> mover decision.*

Screen 6

*Outcome.*

## **Treatment T2**

Screen 1 (Part I)

### **You are assigned the role of participant B**

Prior to entering your decision as participant B, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants B).

In other words, we ask you to guess how many of today's participants B (excluding yourself) will choose to transfer half the money back, and how many of today's participants B will keep all the money for themselves.

\*\*\*\*\*first lower part of screen 1\*\*\*\*\*

*Please enter your guess by positioning the below slider to the desired percentage.*

\*\*\*\*\*second lower part of screen 1 [to appear after subjects have entered their guesses]\*\*\*\*\*

*A sample of other participants B in this session expects on average that  $<x>\%$  will transfer half the money, whereas  $<100-x>\%$  will keep all the money.*

TRANSFER HALF:  $x\%$

KEEP:  $(100-x)\%$

Screen 2 (Part I)

*Enter 2<sup>nd</sup> mover decision.*

Screen 3 (Part II)

*Insert instructions for part II here.*

Screen 4 (Part II)

**You are assigned the role of participant A**

Prior to entering your decision as participant A, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants A).

In other words, we ask you to guess how many of today's participants A (excluding yourself) will choose IN, and how many of today's participants A will choose OUT.

\*\*\*\*\*first lower part of screen 4\*\*\*\*\*

*Please enter your guess by positioning the below slider to the desired percentage.*

\*\*\*\*\*second lower part of screen 4[to appear after subjects have entered their guesses]\*\*\*\*\*

*A sample of other participants A in this session expects on average that  $\langle x \rangle\%$  will OPT IN, whereas  $\langle 100-x \rangle\%$  will OPT OUT.*

IN:  $x\%$

OUT:  $(100-x)\%$

Screen 5 (Part II)

*Enter 1<sup>st</sup> mover decision.*

Screen 6

*Outcome.*

## Acknowledgments

We are grateful to Dirk Engelmann, David Rojo-Arjona, Avichai Snir, Robert Sugden, and Gari Walkowitz for their helpful comments. Also, we thank Bjoern Hartig for managing the experimental lab and for programming the zTree code used in the experiment.

## VIII. References

- Akerlof, George A.** 1980. "A Theory of Social Custom, of Which Unemployment May Be One Consequence" *Quarterly Journal of Economics*, 94(4): 749-775.
- Allport, Floyd H.** 1933. *Institutional Behavior*. Chapel Hill: University of North Carolina Press.
- Anderson, Lisa R. and Charles A. Holt.** 1997. "Information Cascades in the Laboratory" *American Economic Review*, 87(5): 847-862.
- Andreoni, James and Douglas Bernheim.** 2009. "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects" *Econometrica*, 77(5): 1607-1636.
- Asch, Solomon E.** 1956. "Studies of Independence and Conformity: A Minority of One Against A Unanimous Majority" *Psychological Monographs*, 70(9): 1-70.
- Bacharach, Michael, Gerardo Guerra and Daniel J. Zizzo.** 2007. "The Self-Fulfilling Property of Trust: An Experimental Study" *Theory and Decision*, 63(4): 349-388.
- Banerjee, Abhijit V.** 1992. "A Simple Model of Herd Behavior" *Quarterly Journal of Economics*, 107(3): 797-817.
- Battigalli, Pierpaolo and Martin Dufwenberg.** 2007. "Guilt in Games" *American Economic Review P&P*, 97(2): 170-176.
- Berkowitz, Alan D. and H. Wesley Perkins.** 1986. "Problem Drinking Among College Students: A Review of Recent Research" *Journal of American College Health*, 35: 21-28.
- Bernheim, B. Douglas.** 1994. "A Theory of Conformity" *The Journal of Political Economy*, 102(5): 841-877.
- Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian and Katherine L. Milkman.** 2015. "The Effect of Providing Peer Information on Retirement Savings Decisions" *Journal of Finance*, 70(3): 1161-1201.
- Bicchieri, Cristina.** 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
- Bicchieri, Cristina and Alessandro Sontuoso.** 2015. "I Cannot Cheat on You after We Talk: Communication and Norms in Mixed-Motive Games" in *The Prisoner's Dilemma*, ed. Martin Peterson. Cambridge: Cambridge University Press.
- Bicchieri, Cristina and Erte Xiao.** 2009. "Do the Right Thing: But Only if Others Do So" *Journal of Behavioral Decision Making*, 22(2): 191-208.
- Bicchieri, Cristina, Erte Xiao and Ryan Muldoon.** 2011. "Trustworthiness Is A Social Norm, but Trusting Is Not" *Politics, Philosophy and Economics*, 10(2): 170-187.
- Bikhchandani, Sushil, David Hirshleifer and Ivo Welch.** 1992. "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades" *Journal of Political Economy*, 100(5): 992-1026.
- Blanco, Mariana, Dirk Engelmann, Alexander Koch and Hans-Theo Normann.** 2014. "Preferences and Beliefs in A Sequential Social Dilemma: A Within-Subjects Analysis" *Games and Economic Behavior*, 87(C): 122-135.
- Bolton, Gary E. and Axel Ockenfels.** 2000. "ERC: A Theory of Equity, Reciprocity, and Competition" *American Economic Review*, 90(1): 166-193.
- Boyd, Robert and Peter J. Richerson.** 1985. *Culture and the Evolutionary Process*. Chicago: Chicago University Press.
- Burnham, Kenneth P. and David R. Anderson.** 2002. *Model Selection and Multimodel Inference*. New York: Springer-Verlag.
- Charness, Gary and Martin Dufwenberg.** 2006. "Promises and Partnership" *Econometrica*, 74(6): 1579-1601.
- Charness, Gary and Matthew Rabin.** 2002. "Understanding Social Preferences with Simple Tests" *Quarterly Journal of Economics*, 117(3): 817-869.
- Cialdini, Robert B. and Noah J. Goldstein.** 2004. "Social Influence: Compliance and Conformity" *Annual Review of Psychology*, 55(1): 591-621.
- Costa-Gomes, Miguel A., Steffen Huck and Georg Weizsäcker.** 2014. "Beliefs and Actions in The Trust Game: Creating Instrumental Variables to Estimate the Causal Effect" *Games and Economic Behavior*, 88: 298-309.
- Costa, Dora L. and Matthew E. Kahn.** 2013. "Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from A Randomized Residential Electricity Field Experiment" *Journal of the European Economic Association*, 11(3): 680-702.
- Dana, Jason, Roberto A. Weber and Jason Xi Kuang.** 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness" *Economic Theory*, 33(1): 67-80.

- Dawes, Robyn M.** 1989. "Statistical Criteria for Establishing A Truly False Consensus Effect" *Journal of Experimental Social Psychology*, 25(1): 1-17.
- Dhaene, Geert and Jan Bouckaert.** 2010. "Sequential Reciprocity in Two-Player, Two-Stage Games: An Experimental Analysis" *Games and Economic Behavior*, 70(2): 289-303.
- Dufwenberg, Martin and Uri Gneezy.** 2000. "Measuring Beliefs in an Experimental Lost Wallet Game" *Games and Economic Behavior*, 30(2): 163-182.
- Dufwenberg, Martin and Georg Kirchsteiger.** 2004. "A Theory of Sequential Reciprocity" *Games and Economic Behavior*, 47(2): 268-298.
- Ellingsen, Tore, Magnus Johannesson, Sigve Tjøtta and Gaute Torsvik.** 2010. "Testing Guilt Aversion" *Games and Economic Behavior*, 68(1): 95-107.
- Engelmann, Dirk and Martin Strobel.** 2000. "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given" *Experimental Economics*, 3(3): 241-260.
- Falk, Armin, Ernst Fehr and Urs Fischbacher.** 2008. "Testing Theories of Fairness—Intentions Matter" *Games and Economic Behavior*, 62(1): 287-303.
- Fehr, Ernst and Simon Gächter.** 2000. "Fairness and Retaliation: The Economics of Reciprocity" *The Journal of Economic Perspectives*, 14(3): 159-181.
- Fehr, Ernst and Klaus M. Schmidt.** 1999. "A Theory of Fairness, Competition, and Cooperation" *Quarterly Journal of Economics*, 114(3): 817-868.
- Festinger, Leon.** 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.
- Fischbacher, Urs.** 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments" *Experimental Economics*, 10(2): 171-178.
- Gächter, Simon, Daniele Nosenzo and Martin Sefton.** 2013. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association*, 11(3): 548-573.
- Gaviria, Alejandro and Steven Raphael.** 2001. "School-Based Peer Effects and Juvenile Behavior" *Review of Economics and Statistics*, 83(2): 257-268.
- Guerra, Gerardo and Daniel J. Zizzo.** 2004. "Trust Responsiveness and Beliefs" *Journal of Economic Behavior & Organization*, 55(1): 25-30.
- Huck, Steffen, Dorothea Kübler and Jörgen Weibull.** 2012. "Social Norms and Economic Incentives in Firms" *Journal of Economic Behavior & Organization*, 83(2): 173-185.
- Kandel, Eugene and Edward P. Lazear.** 1992. "Peer Pressure and Partnerships" *Journal of political Economy*, 100(4): 801-817.
- Kapur, Sandeep.** 1995. "Technological Diffusion with Social Learning" *Journal of Industrial Economics*, 43(2): 173-195.
- Kranz, Sebastian.** 2010. "Moral Norms in A Partly Compliant Society" *Games and Economic Behavior*, 68(1): 255-274.
- Krupka, Erin L. and Roberto A. Weber.** 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3): 495-524.
- López-Pérez, Raúl.** 2008. "Aversion to Norm-Breaking: A Model" *Games and Economic Behavior*, 64(1): 237-267.
- Manski, Charles F.** 1993. "Identification of Endogenous Social Effects: The Reflection Problem" *Review of Economic Studies*, 60(3): 531-542.
- Rabin, Matthew.** 1994. "Cognitive Dissonance and Social Change" *Journal of Economic Behavior and Organization*, 23(2): 177-194.
- Ross, Lee, David Greene and Pamela House.** 1977. "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes" *Journal of Experimental Social Psychology*, 13(3): 279-301.
- Schram, Arthur and Gary Charness.** 2015. "Inducing Social Norms in Laboratory Allocation Choices" *Management Science*, 61(7): 1531-1546.
- Shue, Kelly.** 2013. "Executive Networks and Firm Policies: Evidence from The Random Assignment of MBA Peers" *Review of Financial Studies*, 26(6): 1401-1442.
- Sontuoso, Alessandro.** 2013. "A Dynamic Model of Belief-Dependent Conformity to Social Norms" *MPRA Paper 53234*, University Library of Munich, Germany.
- Sugden, Robert.** 2000. "The Motivating Power of Expectations" in *Rationality, Rules and Structure*, ed. Julian Nida-Rümelin and Wolfgang Spohn. Amsterdam: Kluwer.
- Thöni, Christian and Simon Gächter.** 2014. "Peer Effects and Social Preferences in Voluntary Cooperation" *CESifo Working Paper Series 4741*, CESifo Group Munich.
- Vanberg, Christoph.** 2008. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations" *Econometrica*, 76(6): 1467-1480.
- Xiao, Erte and Cristina Bicchieri.** 2010. "When Equality Trumps Reciprocity" *Journal of Economic Psychology*, 31(3): 456-470.