The Power of Shame and the Rationality of Trust

Steven Tadelis*
UC Berkeley
Haas School of Business

March 2, 2011

Abstract

A mounting number of studies suggest that individuals are not selfish, which perhaps explains the prevalence of trust among strangers. Models of players who care about their opponents' payoffs have been used to rationalize these facts. An alternative motive is that players care directly about how they are perceived by others. I propose and implement an experimental design that distinguishes perception motives from payoff motives. Participants not only exhibit concerns for perception, but they seem strategically rational by anticipating the change in behavior of their opponents. The approach can explain previously documented behaviors, both in the lab and in the field, and can shed light on some determinants of trust.
JEL classifications C72, C91, D03, D82

*This work was inspired during the Stanford Institute for Theoretical Economics workshop in August 2004. I am grateful to Gary Charness for sharing his experimental design and his experience, and to Pierpaolo Battigalli, Tore Ellingsen, Simon Gervais and Shachar Kariv for detailed comments on earlier drafts. I also thank Yossi Feinberg, Uri Gneezy, Navin Kartik, Mike Katz, Jim Lincoln, John List, Barbara Mellers, John Morgan, Muriel Niederle, Larry Samuelson, Klaus Schmidt, Jonathan Smith and Phil Tetlock for helpful discussions. Victor Bennett and Constança Esteves provided outstanding research assistance. This work has been supported by the National Science Foundation, UC Berkeley's COR grant, and the X-Lab at UC Berkeley's Haas School of Business.

1 Introduction

Trust enables transactions when enforceable contracts are absent, and even when contracts are enforceable, trust can reduce transaction costs between trading partners. It is often argued that trust fosters economic growth and promotes economic development (Fukayama (1995)). Understanding trust, its determinants and its implications for economic outcomes has been a central theme in economic research for at least two decades.[1]

Many studies focused on laboratory experiments that study behavior in the "Trust Game," which is a stylized game of trade without binding contracts that was originally introduced by Berg et al., (1995). A trustor can "trust" a trustee by relinquishing some money, and the trustee, who gains considerably from this act, can either reward the trustor in kind, or can abuse trust and take the entire proceeds. Numerous experiments have confirmed a considerable amount of trust, which is rewarded in kind by unselfish behavior of trustees. By and large these experiments simulated one-shot anonymous interactions where relationship-building (repeated game) effects are absent. This poses a challenge to the standard model of selfishly motivated agents.

This paper develops and implements a stark experimental design that simultaneously answers two related questions. First, is trustworthy behavior motivated by a trustee's concern for how others perceive him? Second, do trustors grant trust as if they expect trustees to be motivated by such concerns for perception? Experimental results confirm a positive answer to both questions. The key experimental design rests on a manipulation of information that players have *after the game is over*, without manipulating the way in which actions affect final payouts.

Altruism has long been recognized as a motive for unselfish behavior, and related impure altruism, or "warm glow" effects, have been advocated as well (Andreoni (1989,1990)). Other motives explored in the literature include concerns for fairness (Fehr and Schmidt (1999), Bolton and Ockenfels (2000)); preferences for reciprocity (Rabin (1993), Dufwenberg and Kirchsteiger (2004)) and identity concerns (Akerlof and Kranton (2000)). Yet none of these pro-social motives can explain why, in one-shot settings, trustees will react to environmental conditions that alter whether others observe their behavior after a game ends.

More recently, however, various forms of a concern for "social esteem" have been explored. The approach typically involves preferences for fairness or altruism that are complemented with a concern for how one is perceived by others (Levine (1998), Benabou and Tirole (2006), Ellingsen and Johannesson (2008), Andreoni and Bernheim (2009)).[2] In these settings "audience effects" emerge and players behave in ways that depend on whether their actions can be observed by others. It is to this body of theoretical work that my paper is related. Its main contribution to this literature is by concentrating solely on concerns for perception, and carefully isolating audience effects from other pro-social motives.

[1] Studies of how trust affects broader issues in economics and finance have attracted considerable attention (e.g., Laporta et al. (1997), Knack and Keefer (1997) and Guiso et al. (2008)).

[2] An older lietrature explores the behavioral implications of concerns for social image, including Bernheim (1994), Ireland (1994) and Glazer and Konrad (1996)).

Also closely related is a recent collection of experimental studies that investigate how obscuring a subject's actions affects their behavior, all of which have focused on variations of the dictator game. For example, when a dictator's behavior is less exposed he behaves more selfishly.[3] In a recent paper, Dana et al. (2006) offer an innovative experimental design in which the dictator can "exit" and conceal the game's existence from a potential recipient, thus obscuring his role. They show that dictators are willing to give up money and choose the exit option rather than having more money to share with a recipient. The paper suggests that "if appearances are the reason the dictator shares, she may wish to keep her endowment private so that the receiver does not expect anything. Then, she does not have to give or feel guilty for not giving."

The results in Dana et al. (2006) directly challenge theories of altruism, fairness and reciprocity. They suggest that people will pay a premium for obscuring their selfish actions, consistent with audience effects.[4] However, an alternative interpretation is that people dislike knowing that others are being disappointed, which is more in line with the explanation offered by Dana et al. This form of "guilt aversion" was advanced by Charness and Dufwenberg (2006), which is discussed in more detail below, and theoretically analyzed by Batigalli and Dufwenberg (2007). The approach I take here contributes to these experimental studies in two significant ways.

First, the trustee cannot manipulate the trustor's beliefs about whether or not a game is being played. Treatments differ only by how much the trustee's actions are exposed, and hence, how easy it is for an audience to infer his actions. As a result, differences in behavior across treatments cannot be explained by a wish not to disappoint.[5] More importantly, I derive robust comparative static effects from a simple model of "shame aversion" and test the specific implications of the model, which are confirmed by the evidence.

Second, by using a trust game to explore audience effects it is possible to simultaneously test both whether trustees behave as if they are "shame averse" (less selfish when their behavior is exposed) and whether trustors behave as if they are strategically rational. Namely, if exposing the trustee's actions causes him to be less selfish, then trustors should be more trusting. The evidence is consistent with strategic rationality.

The next section briefly describes the main idea of manipulating the exposure of a trustee's actions without altering the game's other aspects using the noisy trust game introduced by Charness and Dufwenberg (2006). When a trustee honors trust, there is some probability that

[3] Hoffman et al (1996) showed that in double-blind trials, subjects playing the role of dictators gave smaller amounts as compared to treatments where the experimenter observes giving, suggesting that the experimenter may be a relevant audience. Bohnet and Frey (1999) showed that when dictators and recipients face each other then giving amounts are much higher.

[4] This finding was replicated and refined by Broberg et al. (2007) and by Lazear et al. (2010). Koch and Normann (2008) show that when the recipient is distant from the dictator then giving is unaffected by the recipient's knowledge of the game. Audience effects play prominently in the results of Andreoni and Bernheim (2009), and in Ariely et al. (2009) who interact these effects with material rewards.

[5] Ellingsen et al. (2010) test guilt aversion by eliciting beliefs from participants and sharing them with dictators and trustees. Their evidence suggest that guilt aversion motives are weak. Subtle distinctions with guilt-aversion are discussed in sections 4 and 5.1.

the trustor's payoff is identical to that when trust is abused. Hence, if the trustee's actions are not observed then the trustor cannot distinguish between "bad luck" and "bad behavior."

Section 3 develops an equilibrium model to derive robust comparative static results on exposure when the trustee is shame averse, modeled as a utility loss for the trustee that increases in the trustor's belief that trust was abused. The empirical predictions of the comparative static analysis are used directly to design several experimental treatments that are described in section 4. Results from a series of experiments conducted in 2006 with 171 subjects from the University of California–Berkeley are analyzed in section 5. A discussion follows.

2 Identifying Shame: A Noisy Trust Game

This section briefly describes a trust game originally introduced by Charness and Dufwenberg (2006), and explains how a manipulation of the experimental design can identify preferences that are inconsistent with altruism, fairness or shame aversion, yet consistent with audience effects.[6] Player 1 (the trustor) can trust (T) or not-trust (N) player 2 (the trustee). If trusted, player 2 can cooperate (C) or defect (D). Trust followed by cooperation is Pareto superior to not-trust, but following trust, player 2 must incur a cost to cooperate. Defection, however, imposes a cost on player 1. There is imperfect success to cooperation: with probability $p \in (0, 1)$ cooperation succeeds and player 1 receives a high payoff, while with probability $1 - p$ cooperation fails, and player 1 receives a payoff identical to his payoff from defection. Conditional on cooperating, the pecuniary payoffs to player 2 do not depend on success.

Figure 1 describes two treatments for this game with dollar payoffs that are used in the actual experiments and where $p = \frac{5}{6}$. The left panel shows the "exposure" treatment where player 1 observes player 2's action. The right panel shows the "no exposure" treatment where player 1 cannot distinguish between defection and bad luck, both yielding him 0.

The difference between these two treatments only affects the possible beliefs that player 1 can have about player 2's behavior. A selfish trustee will defect regardless of exposure, and anticipating this a trustor should never trust. Moreover, since the mapping from pairs of actions to (probability distributions over) outcomes is unchanged, theories of altruism, fairness or reciprocity all imply that both players should behave in the exact same way in both treatments.

If subjects exhibit *different* behavior across these two treatments then players must have a concern for ex post perception. If player 2 experiences "shame" when he is perceived as a defector then he should be more likely to cooperate when his behavior is exposed.[7] The resulting

[6] Charness and Dufwenberg (2006, 2011) introduce cheap-talk into the trust game and focus on the role of communication, not exposure. Andreoni and Rao (2011) investigate communication in dictator games.

[7] This is related to a vast literature in psychology that studies "Self-presentation", which is an individual's concern about constructing his or her "public self". Baumeister (1982) notes that "The most common procedure for testing for self-presentational motives is by comparing two situations that are identical in all respects except that some circumstance is public in one situation but private in the other." He continues, "If public awareness makes people change their behavior, it is because they are concerned with what their behavior communicates to

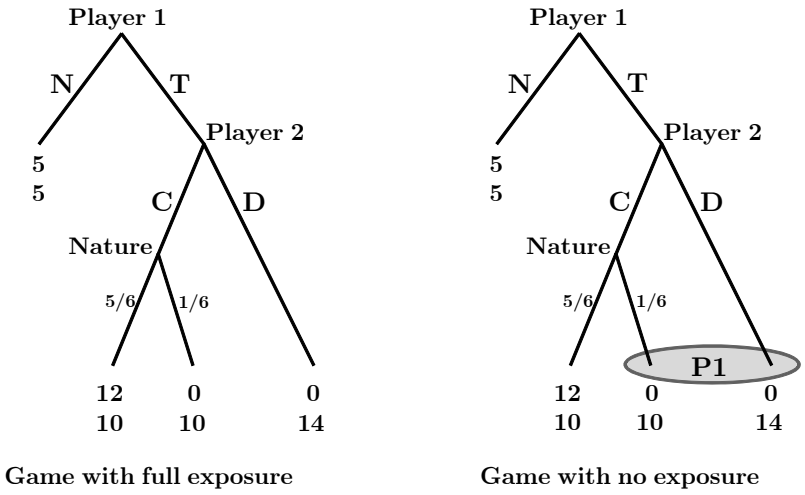

Figure 1: A Noisy Trust Game with and witout Exposure

hypothesis follows:

Hypothesis 1 (The Power of Shame): Player 2 is more likely to cooperate in a game with exposure relative to one with no exposure.

Now, if player 1 anticipates the "power of shame" then trust is more likely to pay off with exposure. This implies the following hypothesis:

Hypothesis 2 (The Rationality of Trust): Player 1 is more likely to trust in a game with exposure relative to one with no exposure.

In the next section I lay out a simple model from which robust comparative statics are derived. These will not only supply solid foundations for the two central hypotheses listed above, but will also lead to a set of well defined treatments that tease out the effects of exposure.

3 The Model

Consider the noisy trust game with the following payoffs: if player 1 chooses N then both players receive $v > 0$. If player 1 chooses T and player 2 chooses C then both get an expected payoff of $c > v$, where player 2 gets c for sure, player 1 gets $\frac{c}{p}$ with probability p and 0 with probability $1 - p$. If player 1 chooses T and player 2 chooses D then player 1 gets 0 and player 2 gets $d > c$.

others." The approach taken here is very much related to this line of research, with the extra step of embedding behavior in a rational strategic setting.

With no pro-social considerations the game has a unique sequential equilibrium: if trusted, player 2 will never cooperate since $d > c$, and in turn player 1 will never trust since $v > 0$.

For simplicity, player 1 is assumed to be risk neutral and selfish so that only monetary payoffs matter to him and an expected amount m_1 yields utility $u_1 = m_1$.[8] Player 2 is also assumed to be risk neutral but has private information about the intensity of his pro-social concerns modeled as preferences over the posterior beliefs that player 1 has about player 2 after the game is over. There are two ways to model such preferences. One used by Ellingsen and Johannesson (2008) would have player 1 form beliefs about the *type* of player 2, and player 2 having preferences over those beliefs. An alternative that I use, which is simpler to incorporate, is that player 1 forms beliefs about the *action* that player 2 chose, and player 2's preferences are over these beliefs.

Formally, player 2 cares about money and about player 1's beliefs about the action that player 2 actually chose. Let $E_1[C] \in [0,1]$ be the posterior belief of player 1 about the probability that player 2 chose to cooperate. Player 2 of type s who receives payment m_2 and who believes his opponent's ex post belief is $E_1[C]$ has utility $u_2 = m_2 - s(1 - E_1[C])$, where s represents player 2's *shame aversion*. Let s be distributed over $[0, \infty)$ with cumulative distribution $F(\cdot)$ and positive density $f(\cdot)$.[9] That is, player 2 suffers when player 1 *thinks* that player 2 defected with positive probability, but player 2 does not care about the *outcome* that player 1 receives. Notice that these preferences have neither altruism, concerns for fairness nor reciprocity considerations. As a consequence, player 2 is not averse to defecting, but is averse to the beliefs that others have about his possible defection.[10]

A "Good" outcome (G) is the one in which player 1's payoff is $\frac{c}{p}$ and a "Bad" outcome (B) yields player 1 a payoff of 0. An exposure technology determines whether player 1 can infer the reason for a bad outcome. The game exhibits *exposure* ($e = 1$) if player 1 observes the action of player 2 and learns why a bad outcome occurred. The game exhibits *no exposure* ($e = 0$) if player 1 learns nothing about the reason for the bad outcome.[11]

[8] Since treatments will only affect exposure of player 2, I focus on social preferences for player 2 only. Including either pro-social concerns for player 1, or risk aversion, will not alter the comparative static results. For a model with social concerns for both players that interact in interesting ways see Ellingsen and Johannesson (2008).

[9] Assuming positive density is convenient but not necessary. Introducing positive measures of certain type-values can introduce mixed strategy equilibria that are less convenient to work with. Also, it suffices to assume that $s \in [\underline{s}, \overline{s}]$ where $\underline{s} < d - c < \overline{s}$ as will be come apparent from the analysis that follows.

[10] More general preferences of the form $u_2(m_2, E_1[C], s) = m_2 - \phi(E_1[C], s)$, where $\phi(\cdot, \cdot)$ increases in both components would suffice for robust comparative statics. Using the alternative formalization as in Ellingsen and Johannesson (2008) of "esteem" would be more in line with beliefs of the form "I think you're a cheater" instead of "I think you have likely cheated me." To capture this alternative let $F^{post}(\cdot)$ be the posterior belief of player 1 about the type of player 2, which is derived from $F(\cdot)$ and from belief about the strategy of each type. Player 2's preferences are over money and $F^{post}(\cdot)$ so that less favorable beliefs about 2's type will hurt more shame-averse types. In this simple environment, these two notions of shame are isomorphic, and hence I resort to the simpler formulation.

[11] It is easy to introduce a more refined exposure technology $\tau \in [0,1]$ as follows: with probability τ there is exposure after the payoffs are determined, and with probability $(1 - \tau)$ there is no exposure. The results derived in the next section generalize to the case where exposure is given by this continuous probability of detection. This τ is similar, but not identical to the parameter p in Andreoni and Bernheim (2009).

As player 1's beliefs about the action of player 2 directly enter player 2's utility, the analysis follows the literature on "Psychological Games" pioneered by Geanakoplos et al. (1989) and developed further for dynamic games by Battigalli and Dufwenberg (2008). I use a straightforward application of sequential equilibrium where in each subgame players play a best response to their beliefs and these beliefs are consistent with Bayes' rule. The best response of player 2 may depend on the beliefs of player 1 (more precisely, the *beliefs of player* 2 *about* the beliefs of player 1). Equilibrium analysis requires that player 1 have correct beliefs about the types and actions of player 2, and that player 2 have correct beliefs about the beliefs of player 1.[12]

Player 1's posterior beliefs will depend on the history of play after the game ends. With exposure this history can be restricted to C and D (the actions of player 2), while with no exposure the history can be restricted to G and B because player 1 does not observe actions. Hence, let $E_1[C|h,e]$ be the posterior belief that player 1 assigns to action C being chosen, conditional on the history $h \in \{C,D,G,B\}$ and exposure $e \in \{0,1\}$. Let $\sigma_s : [0,\infty) \to [0,1]$ denote the strategy of player 2 (the probability that player 2 of types s chooses C).

Clearly, following a good outcome player 1 perfectly infers that player 2 cooperated, hence $E_1[C|G,e] = 1$ for $e \in \{0,1\}$. Also, exposure reveals the action of player 2, and hence $E_1[C|C,1] = 1$ and $E_1[C|D,1] = 0$. Finally, with no exposure bad outcomes may not be inferred perfectly so that $E_1[C|B,0] \in [0,1]$. Furthermore, if $\sigma_s > 0$ for a positive measure of types (implying a positive probability of cooperation) and $\sigma_s < 1$ for a positive measure of types (implying a positive probability of defection) then $E_1[C|B,0]$ is strictly within the interior of the $(0,1)$ interval.

Proposition 1: If $e = 1$ then there is a unique sequential equilibrium characterized by a cutoff type $s^1 = d - c > 0$ such that all types $s \leq s^1$ choose D and all types $s > s^1$ choose C.

With exposure, the unique equilibrium is obvious: since $d > c$, monetary payments provide an incentive to defect. However, since behavior is perfectly observed then the shame-cost of defection is independent of the measure of types who cooperate. Hence, all types $s > d - c$ have a shame-cost of defection that is higher than the monetary benefit.[13]

Proposition 2: If $e = 0$ then there is a unique sequential equilibrium characterized by a cutoff type $s^0 > s^1$ such that all types $s \leq s^0$ choose D and all types $s > s^0$ choose C.

Proposition 2 shows that fewer shame averse types will choose to cooperate with no exposure as compared with exposure. The intuition follows in three steps. First, consider one extreme where player 1's prior beliefs are that *all types cooperate*. His posterior beliefs will never be updated and shame is never imposed. But then, *all types will defect*. Second, consider the other

[12] Battigalli and Dufwenberg (2008) show that adopting sequential equilibrium to such a game is indeed valid and that sequential equilibria exist. Unlike Battigalli and Dufwenberg (2007, 2008), higher order beliefs will not play a role in the game analyzed and are therefore ignored.

[13] A similar analysis follows from preferences over beliefs about types. If all types prefer to be perceived as more shame averse then shame aversion would motivate more shame averse types to act cooperatively. The analysis would be more involved but will carry the same logic.

extreme where player 1's prior beliefs are that no type ever cooperates, then posterior beliefs after a bad outcome must be that *all types defect* and shame is as high as for defection with exposure. But then *all types with* $s > s^1$ *will cooperate*. It follows that in equilibrium some types must defect and some must cooperate, belief updating is therefore in the interior of $(0, 1)$ and shame is less painful than with exposure. Hence, some types at and above s^1 must defect.[14]

Propositions 1 and 2 provide comparative statics that underpin the first hypothesis of section 2, *the power of shame*. To reiterate, player 2 is more likely to cooperate in a game with exposure relative to one with no exposure. In fact, the prediction is more refined and implies monotonic behavior for *every* type. A type who cooperates with no exposure will cooperate with exposure; a type who defects with exposure will defect with no exposure; and some types will switch their behavior from defect to cooperate when their actions become exposed.

Equilibrium behavior of player 1 depends on the distribution of player 2's types. His payoff from trusting with exposure is $(1 - F(s^1))c$ while with no exposure it is $(1 - F(s^0))c$, which is lower. Thus, there are three cases of parameter ranges to consider. First, if the distribution of types is such that $1 - F(s^1) < \frac{v}{c}$ then player 1 should never trust. Second, if $1 - F(s^0) > \frac{v}{c}$ then player 1 should always trust. Last, if $1 - F(s^0) < \frac{v}{c} < 1 - F(s^1)$ then player 1 should not trust with no exposure but should trust with exposure.

Player 1's response to exposure is therefore weakly monotonic. If $1 - F(s^0) < \frac{v}{c} < 1 - F(s^1)$ then his behavior is strictly monotonic in exposure, while otherwise it is unchanged. This underpins the second hypothesis of section 2, *the rationality of trust*. To reiterate, player 1 is more likely to trust in a game with exposure relative to one with no exposure. To get a more refined prediction some heterogeneity for player 1 must be introduced. For instance, players may have heterogeneous subjective beliefs about the distribution of types that comes from past experience (e.g., the prior in this game is the posterior of past experiences). In this richer environment, denote a type of player 1 as β, with beliefs $F(s; \beta)$ about the distribution of shame aversion. One can order player 1 types using first-order stochastic dominance: a type β' is a "higher type" than type β if $F(s; \beta') < F(s; \beta)$ for all s. This extension implies robust comparative statics with heterogeneous individual-level monotonic behavior of *every* type of player 1. Any type who trusts under no exposure will trust under exposure; any type who does not trust under exposure will not trust under no exposure; and some types will switch their behavior from not-trust to trust when player 2's actions are exposed.

The model can be extended further to derive an additional comparative static result that is

[14] The results generalize if noise also follows defection. Imagine that after a choice of D the outcome is good with probability q and bad with probability $1 - q$. As long as $q < p$, the results described above will carry over because Bayes updating will imply that $0 < E[C|B, 0] < E[C|G, 0] < 1$ when both C and D are chosen with positive probability, and the inequality $E[C|B, 0] < E[C|G, 0]$ would drive the result. However, no-trust and defect becomes an equilibrium with no exposure. With this equilibrium the comparative static result holds trivially since trust and cooperation are less than in the case of exposure. Also, as mentioned in footnote 10, a more refined exposure technology $\tau \in [0, 1]$ will imply both type-monotonic behavior and the structure of Bayes updating in this game.

closely related to exposure: the effect of anonymity. The game so far is between two players, where player 1 updates beliefs about player 2. A direct experimental implementation would be, therefore, to have matched-pairs where each player knows who his matched-player is. It is uncommon, however, to expose the identity of players and most studies implement anonymous pairs, imposing a form of no exposure. I concur with Andreoni and Bernheim (2009) who advocate for "the importance and feasibility of studying audience effects with theoretical and empirical precision." (p. 1610) The theoretical framework outlined in this section can help explore the interplay between exposure and anonymity.

To proceed, imagine that there are n players in each role, yet the pairs are anonymous. In this case, loosely speaking, each players 1 "spreads" his updating across all players 2.[15] For example, anonymous matching with exposure means that if some player 1 learns that his anonymous partner chose to defect, he learns neither *who it was*, nor does he learn the outcome of the other pairs. All he knows is that there is a defector in the group of players 2. Hence, anonymity is another form of "imperfect monitoring." However, even with anonymous pairs, exposure lets every player 1 update beliefs about the group of players 2. As a result, extending the game to several pairs of players yields an additional comparative static hypothesis as follows:

Hypothesis 3 (The Weakness of Anonymity): Fixing exposure, a switch from anonymous to matched pairs increases the likelihood of both cooperation and trust.

4 Experimental Design

Hypotheses 1, 2 and 3 imply that anonymity and exposure can be interacted in four treatments as described in Table 1.

	No exposure		Exposure
Anonymous	AN	$\Rightarrow$	AE
	$\Downarrow$		$\Downarrow$
Matched Pairs	MN	$\Rightarrow$	ME

Table 1: Treatments Derived from Hypotheses 1, 2 and 3

Consider anonymous pairs with no exposure (treatment AN). Incentives to cooperate (and trust) will increase if either behavior is exposed (treatment AE) or if pairs become matched (treatment MN). Similarly, switching from matched pairs with exposure (treatment ME) to either treatments MN or AE will reduce incentives to cooperate (and trust). In the table, arrows

[15] Precisely, every player 1 has beliefs over the distribution of n actions and outcomes. When observing the history of his game, only one of the n components of his belief is revealed, and Bayes updating is less extreme about the symmetric group of platers 1. Each player 2 understands that his actions impose a negative externality on the beliefs of players 1, and as a consequence, the incentives to cooperate are weaker.

mark the direction in which audience effects are stronger. Note that it is not possible to rank treatments AE and MN since one form of exposure increases while the other decreases.[16]

Two extra treatments not directly derived from the theory can shed light on more subtle aspects of audience effects. These treatments both exhibit *public exposure* where the actions of players 2 are *announced to all the participants* in the room. They differ in that one treatment has anonymous pairs with public exposure (treatment AP) whereas the other treatment has matched pairs with public exposure (treatment MP).

These additional treatments are useful to explore the concept of "guilt-from-blame" that is introduced by Batigalli and Dufwenberg (2007), where player 2 dislikes being "blamed" for bad outcomes.[17] If shame is the primary motivator then it is the *announcement of behavior* and not the *identity-matching* that should matter. In both treatments AP and MP exposure is identical, so shame motives should have the same effect. If, however, guilt-from-blame is present, then anonymous pairs implies that the trustor cannot assign blame directly, an externality is created, and the level of cooperation in treatment AP should be less than that of MP.

The experimental game was directly taken from Charness and Dufwenberg (2006) and followed their implementation closely. Sessions were conducted at UC Berkeley's X-Lab in a large classroom divided into two sides by a center aisle. Participants were randomly seated at private tables with dividers between them. Twelve sessions were conducted. Two pilot sessions included only one treatment each, AN and AE, to ensure that the experiment was manageable and easy to implement. Each of the next four sessions (totalling 34 pairs) included the four treatments in Table 1. The AP and MP treatments were added to the other four treatments in each of the last six sessions, so those included six treatments each (totalling 53 pairs).

There were 10-30 participants per session with a total of 86 participants in the role of player 1 and 85 in the role of player 2.[18] Smaller sessions (5-6 pairs) lasted for 30-40 minutes, and large sessions (10-15 pairs) lasted for about one hour. The average payout was about \$14, which included a \$7 show-up fee. In each session participants were referred to as "A" or "B" (for players 1 and 2 respectively). A coin was tossed to determine which side of the room were A-players. Personal identification numbers were assigned to participants (A1, A2,..., B1, B2,...) who were informed that they will be used to choose pairings, track decisions and determine payoffs. Before each treatment began, a new random draw of pairings was used to prevent a repeated game

[16] The ranking of AE and MN will depend on the number of pairs and on beliefs about the distribution $F(\cdot)$, the latter being impossible to control for.

[17] Batigalli and Dufwenberg (2007) also define "simple guilt" where player 2 dislikes having player 1 disappointed by getting less than he expects to. The predictions derived in hypotheses 1 and 2 above cannot be derived from simple guilt. They can be derived from guilt-from-blame since either removing anonymity or removing obscurity can intensify the blame that player 1 can bestow upon player 2. These subtleties are related to the psychology literature that compares guilt with shame. As noted by Tangney (1995), "there is a long-standing notion that shame is a more "public" emotion than guilt, arising from public exposure and disapproval, whereas guilt represents a more "private" experience arising from self-generated pangs of conscience."

[18] There were 5 sessions for which there were an odd number of players, and in these cases one player was matched with two opponents, one of them randomly assigned to determine the payoff of the player who is matched twice.

between any pair of individual players.

The payoffs in Figure 1 were used in all treatments with numbers representing dollar amounts. A-players received a sheet with two options, "*In*" (for "trust") and "*Out*" (for "no-trust.") B-players received a sheet with two options, "*Roll*" (for "cooperate") and "*Don't Roll*" (for "defect") a six-sided die. First, A-players record their choice after which their sheets were collected. Next, B-players record their choice without knowing the actual choice of their paired A-player. The instructions (see Appendix B) explained that a B-player's choice would be relevant only if his paired A-player chose *In*. This conventional "strategy method" guarantees an observation for every B-player. After the B-players recorded their decision, the resolution of the 6-sided die roll was recorded for every B-player on the back of their decision sheet *regardless* of their choice. This was explained to the participants in advance to allow for anonymity of B-players who chose *Don't Roll* in the anonymous treatments. The actual resolution of the die was relevant only if (*In, Roll*) had been chosen by the pair of players. The outcome corresponding to a success occurred only if the die came up 2, 3, 4, 5, or 6 (hence, $p = \frac{5}{6}$).

Note that the mere presence of an experimenter, or the belief of a subject that someone is observing their behavior, will act as a motivator for shame averse players. The experimental design, however, keeps the presence of the experimenter constant across treatments and only manipulates exposure to other subjects.

5 Experimental Results

5.1 The Power of Shame

Table 2 presents the means and exact confidence intervals for the trustee's behavior in each of the 6 treatments based on the binomial distribution. On the left are the results of the complete panel from all the sessions and all the treatments.[19] These results are also shown in Figure 2. As predicted, there is more cooperation in treatment ME relative to MN and in AE relative to AN, while AE and MN cannot be ranked. The magnitudes of changing exposure and anonymity are impressive. (In what follows, 95% confidence intervals are bracketed.) The mean of cooperative behavior practically doubles from 0.2 $[0.12, 0.30]$ in the AN treatment to 0.38 $[0.27, 0.48]$ in the AE treatment and 0.4 $[0.3, 0.51]$ in the MN treatment. These are almost doubled again to 0.75 $[0.65, 0.84]$ in the ME treatment.

Interestingly, the behavior of B-players in the public treatments AP (0.73 $[0.58, 0.84]$) and MP (0.78 $[0.65, 0.89]$) cannot be distinguished, and is also indistinguishable from behavior in the ME treatment. This suggests that identity-matching per se does not affect behavior, and that exposure to one player seems to have the same shame effects as exposure to the whole group. This

[19] These do not include the two pilot sessions, one for the AN treatment and the other for the AE treatment. When these data are included then the AN treatment has 102 observations, a mean of .21 and a standard error of .04. The AE treatment has 100 observations, a mean of .36 and a standard error of .048. Adding the pilot results maintains the conclusions and strengthens the confidence intervals for these treatments.

is consistent with shame aversion but not with guilt aversion.[20] This complements the results in Ellingsen et al. (2010) who, using a clever treatment of exposing expectations, find little to no evidence of guilt aversion.

The right side of Table 2 shows results using only the first two treatments from each session. The reported results are less precisely estimated since there are fewer observations. Still, the general pattern appears consistent with the theoretical results implying that they are not driven by the later treatments.[21]

Table 3 shows the results from a linear regression of the behavior of B-players who played multiple treatments (i.e., excluding the pilot sessions). The dependent variable y_{it} is the choice of player i in treatment t. With mutually exclusive treatments (making the linear model valid) the predicted values of y_i will be the expected value of Y conditional on the treatments, which is just the proportions of B-players that played *Roll*. Robust Huber-White standard errors are clustered at the individual level to account for the heteroskedastic variance of the y_i values.[22]

Column (1) contains the analysis for the first four treatments for all multi-treatment sessions. Column (2) contains only the AP and MP treatments for the last 6 sessions. Both columns (1) and (2) yield almost identical results to the binomial results in Table 2. Column (3) pools all treatments for the subjects that underwent the last six sessions, each with six treatments. The F-tests, assessing whether dummies for the treatments are jointly zero appear at the bottom of the table. One can reject the hypothesis, given the p-value of almost zero, that the dummies are jointly equal to zero in all columns.

More meaningful F-tests relate to the theoretical predictions and test whether the coefficients on some dummy variables are significantly different from those of other dummy variables. Indeed, as Figure 2 suggests, the hypothesis that the behavior is the same in the MN and ME treatments is rejected with a p-value close to 0. Similarly, the data rejects that the behavior is the same in the AN and AE treatments. However, one cannot reject the hypothesis that behavior is similar in the ME, AP and MP treatments at conventional levels.

[20] Clearly, exposure to all players cannot be weaker than exposure to one, and perhaps more cooperative behavior should be expected for the public treatments. However, there are two reasons that strictly more cooperative behavior would fail to emerge. First, since cooperation rates are at around 75-80% with exposure to one player, there's not much room for improvement. It may be reasonable to believe that 20% of people are just selfish. Second, the theoretical implications of exposure to more people depends on how players believe that other players communicate. If one believes that exposure to one person is enough to get the "rumor mill" churning, then it can be almost as effective as exposure to a larger set of players.

[21] This alleviates the concern that subjects "figure out" what the experimenter expects as they play through treatments in the same session. Using only the first treatments includes too little data to offer meaningful results.

[22] More precisely, this models the discrete choice problem faced by B-players between choosing *Roll* or *Don't Roll*, where $Y_i \sim$ Bernoulli $(p = \delta_j T_j, j = 1, ...6)$. Hence, $E[(Y_i \mid T_j, j = 1...6) = p = \delta_j T_j, j = 1, ...6)$, where $VAR\,(Y_i \mid T_j, j = 1...6) = p(1-p)$, and the errors are therefore heteroskedastic.

5.2 The Rationality of Trust

Turning to the trustor's behavior, the experimental results again corroborate the predictions of the theory. Table 4 presents the means and exact confidence intervals for all the treatments on the left side, and for the first two treatments in each session on the right side. As before, these are the exact means and confidence intervals based on the binomial distribution.[23] The results of Table 4 are also shown in Figure 3. Table 5 displays linear regression results for the behavior of the A-players similarly to Table 2 for the B-players.

Again, the magnitudes of changing exposure and anonymity are impressive. The mean of trusting behavior almost doubles from 0.3 $[0.21, 0.41]$ in the AN treatment to 0.56 $[0.45, 0.67]$ in the AE treatment and to 0.55 $[0.44, 0.65]$ in the MN treatment. These increase by almost 50% to 0.74 $[0.64, 0.83]$ in the ME treatment. Behavior in the two public treatments AP (0.66 $[0.52, 0.79]$) and MP (0.74 $[0.6, 0.85]$) cannot be statistically distinguished.

The relevant F-tests are whether the coefficients on some dummy variables are significantly different from those of other dummy variables. The stark results suggested by Figure 3 are verified by formal statistical tests. The hypothesis that the behavior in the MN and ME treatments is the same is rejected with a p-value of 0.0002. The data also rejects that the behavior in the AN and AE is the same. However, one cannot reject the hypothesis that behavior is similar in the ME, AP and MP treatments at conventional levels.

5.3 Monotonic Individual Behavior

The model of Section 3 suggests a more stringent test of the theory beyond aggregate behavior in that *every* individual's behavior should be *monotonic*: when exposure or matched pairs are introduced then there is a higher likelihood that any given B-player will choose cooperate, and that any given A-player will choose trust.

To investigate whether monotonicity is violated, it is necessary to define what a violation of monotonicity is for each of the two players. For B-players, cooperating in AN and not in AE, MN or ME would constitute a violation. The reason is that there is more exposure in either AE, MN or ME as compared with AN. If the power of shame is strong enough to induce cooperation in the AN treatment then it must be more than enough in the other treatments. Similarly, cooperating in AE or MN and not in ME would constitute a violation.

Like the B-players, A-players should not violate monotonicity when moving from the AN treatment to the AE treatment, or when moving from the MN treatment to the ME treatment. However, things are more subtle for some of the other transitions because A-players may learn something in the middle of the experiment following the AE or ME treatments. For instance, if an A-player trusted in the AE session and learned that his trust was abused then his posterior of

[23] As in Table 1, these results do not include the two pilot sessions, one which ran the AN treatment and the other running the AE treatment. When these data are added then the AN treatment has 103 observations, a mean of .33 and a standard error of .046. The AE treatment has 101 observations, a mean of .51 and a standard error of .050. Thus, the pilot results do not alter the conclusions.

the distribution of types is worse, and he may then *rationally* choose not to trust in a treatment with more information. This in turn implies that a violation of monotonicity by an A-player can be rationalized by Bayes updating following a bad experience in an AE treatment. However, if a player A trusted in the MN (or AN) treatment, he must trust in the ME (or AE) treatment since he does not learn about behavior.

B-players exhibit a high degree of monotonicity: only 12 out of 85 individuals violate monotonicity (of these, three players exhibited two violations and one player exhibited three violations). Players A have significantly more violations of monotonicity: 24 players out of 86. However, a positive and significant correlation exists between A players whose trust was abused in a treatment with exposure, and the same players exhibiting a non-monotonicity after learning about the abuse. Of the 24 individuals, 20 exhibit reversals that are consistent with rational non-monotonicity, i.e., they occur after trust in the AE treatment was violated. Only four of the 86 A-players exhibit an irrational non-monotonicity.

An alternative explanation is that a taste for reciprocity accounts for this non-monotonic behavior (e.g., Dufwenberg and Kirchsteiger (2004)). After being abused, the A-player "punishes" the B-player. It is possible, but arguably less convincing because of the random pairings across treatments. Reciprocity implies that players should believe that the probability of re-matching is high enough, which is questionable with groups of six to twelve pairs.

6 Discussion

6.1 Beyond Trust Games and the Lab

The presence of pro-social behavior has been documented in the lab and the field. I will briefly discuss results from some other studies and discuss how they are consistent with shame aversion.

6.1.1 Dictator Games

Dana et al. (2006) find that a third of participants were willing to exit a \$10 dictator game and take \$9 to avoid the receiver ever knowing that a dictator game existed. As discussed earlier, altruism or concerns for fairness cannot explain choosing the (\$9, \$0) exit outcome over the \$10 dictator game that includes the (\$9, \$1) outcomes. They conclude that "giving often reflects a desire not to violate others' expectations rather than a concern for others' welfare per se." This statement is very much in line with guilt aversion as described in Batigalli and Dufwenberg (2007). Shame aversion can also explain these results, yet guilt averse preferences are not consistent with all the experimental findings of Section 5 above.[24]

Andreoni and Bernheim (2009) also explore dictator games and suggest an explanation of the well documented 50-50 division norm. They employ preferences where people like to be

[24] Shame aversion also explains another finding in Dana et al. (2006) in a game where the receiver had no knowledge of what determins how much money he received. Almost no dictators exited from this game.

perceived as fair, which shares a similar flavor to the shame averse preferences introduced here. They also manipulate the recipients' ex post information in a subtle way, resulting in audience effects that are consistent with a perception concerns. Notice that the role of B-players above is similar to that of a dictator with two actions (but not "constant-sum"), supplying further evidence of audience effects. The trust game, however, goes a step further and demonstrates the rational response of A-players who seem, through their behavior, to acknowledge that their fellow B-players are motivated by shame aversion.

6.1.2 Voting and Public Good Provision

There have been many attempts to rationalize the fact that many people turn out to vote in spite of their vote having little to no effect on outcomes, yet voting itself is costly. Common explanations invoke pro-social behavior where people are "public minded," and hence they show up at the ballots. This implies that if access to voting becomes less costly then turnout should increase. Using data from Swiss elections before and after mail-in voting was introduced, Funk (2010) presents results that fly in the face of this conventional wisdom.

Mail-in voting clearly reduces voting costs substantially, yet overall turnout did not increase on average. In fact, voter turnout *decreased* in smaller communities while slightly increasing in larger ones. Turnout also decreased in communities where poll station had shorter operating windows. Funk (2010) suggests that voters wish to be *perceived* as public minded, very much consistent with shame-averse preferences. In smaller communities and in those with shorter poll operating hours, it is more likely that someone you know will be at the station at any given time, akin to a higher level of exposure. Mail-in voting, however, gives individual an excuse that effectively resembles no exposure at all.

Another related application is the use of public fund-raising in religious gathering places such as churches. Soetevent (2005) conducted a field experiment in thirty Dutch churches where offerings were gathered randomly using either a "closed" collection bag or an open collection baskets. When using baskets, attendants' contributions can be identified by those sitting next to them. Initially, contributions increased by 10% when baskets were used, though this positive effect of using baskets petered out over the experimental period (29 weeks). Also, the coins collected show that churchgoers switch to giving larger coins when exposed baskets were used. Open baskets are akin to an increase in exposure, resulting in behavior that is consistent with shame aversion.

6.2 Shame and Reputational Concerns

The repeated games literature teaches us that players who only care about material payoffs will effectively have indirect preferences over the beliefs of others. This is particularly pronounced for repeated games of incomplete information with imperfect monitoring (see, e.g., Mailath and Samuelson, 2006), where a player's digression cannot easily be detected by his counterparts. Per-

haps players in the experiments described earlier treat the game *as if* it were a repeated game with the other people in the room. Exposure eliminates imperfect monitoring, enhancing reputational concerns and causing more cooperation. This can rationalize many of the experimental results without resorting to pro-social preferences. (Note, though, that repeated games do suffer from the well known problem of multiple equilibria.)

Two reasons render this alternative explanation less convincing than shame aversion. First, the pool of applicants is drawn from several thousands of students and staff members at UC Berkeley.[25] For repeated interaction to convincingly explain behavior, either the probabilities of future interaction or the marginal losses from potential future interactions need to be sufficiently high. Both seem unlikely given the pool of applicants. Second, and more striking, repeated interaction should predict that the level of cooperation in the ME treatment is significantly lower than the AP and MP treatments, which is not the case. If "what goes around comes around," then exposure to more people should act as a significantly strong motivator.[26]

The relation between shame aversion and reputation is meaningful though because shame aversion may act as reduced form preferences that facilitate trust, instead of relaying on complex history-dependent strategies. If most human contact involves repeated interactions then preferences may have evolved to best fit these circumstances. This idea dates back at least to Frank (1987) who argues that human emotions are shaped by natural selection.[27]

6.3 Concluding Remarks: Policy and Strategy

This study was motivated by an important question: what determines trust, and how can trustworthy behavior be encouraged? If shame aversion provides an answer, then the next question is to what extent can individuals, organizations and policy-makers design institutions that leverage shame aversion?

Legal scholars have discussed the role of emotions in the law (see, e.g., Posner (2001)), and shame has been applied to crime deterrence for some time. For example, in the Spring of 2005 Oakland's City Council President Ignacio De La Fuente was quoted saying that "We're going to shame the out-of-towners and locals who drive to our neighborhood to look for prostitutes." The strategy was to post pictures of the offenders on large billboards throughout the city. The experiment was never truly executed because of legal battles, but it presents an example of how shame can affect public policy.[28]

[25] Indeed, casually observing the students as they waited to be seated suggests that very few of them knew each other before participating in the experiment.

[26] Repeated interaction can still explain this (and many other behaviors) by assuming that it is enough for one person to learn your action, after which word-of-mouth immediately spreads to all future potential partners. For a study of pro-social preferences and reputation effects in the field see List (2007).

[27] There has been a flourishing formal literature that explores the conditions for evolutionary stability of non-selfish preferences. See, e.g., Heifetz et al. (2007) and the references therein.

[28] The initiative was legally challenged and as a result the pictures put up on the billboards were blurred enough to not be recognizable. The program was then shelved by the summer of 2005. For a recent discussion of shame

An interesting business case dates back to the early nineteenth century. The Utopian idealist, Robert Owen, adopted a novel incentive mechanism to raise the standard of goods produced in mills in New Lanark, Scotland. Above each machinist's workplace, a cube with four colored faces was installed (black, blue, yellow and white, in ascending order of quality). Depending on the quality of the work, a different color was displayed for all others to see, but no formal rewards or punishments were used. "Owen merely walked through the factory each day looking at the worker and then the monitor, and never said a word. Complaints by workers of unfair ratings could be made directly to Owen. Initially, there were many black marks, but over time, the colors changed from predominantly blue, to yellow, to white. By this device, Owen claimed to have prevented misconduct." (Bloom, 2003). This mechanism seems to rely on shame aversion to promote better performance.

Peer pressure and social exposure may be a fruitful method for business organizations to provide incentives. Mas and Moretti (2009) study the productivity of cashiers in a national supermarket chain. They define individual productivity as the number of items scanned per second, and find that when high productivity cashiers are added to the current pool of cashiers, the average productivity of the *other* cashiers increases. They find, however, that productive workers induce a productivity increase only in workers that are *in their line-of-vision*. Once again, exposure seems to play a central role in motivating behavior.

The potential impact of shame aversion on contractual relationships and gains from trade can be significant. Economic relationships may suffer from moral hazard and opportunism when contracts cannot be fully specified or enforced. This "self interest seeking with guile," as Williamson put it (1975, p. 26), is at the heart of studying contractual relations and institutional design. Interestingly, despite these hazards, problems of moral hazard are not rampant and the use of external litigation is often the exception rather than the rule.

It is possible, though not easily proven, that shame averse preferences are responsible for trustworthy behavior when contracts are incomplete and external enforcement is fragile. Consider the case where two parties can engage in some transaction for which a contract cannot be completely specified in advance. Agency theory suggests that information should be gathered to support exchange *only if* it helps enforcement. However, if parties can reveal information that causes exposure of poor behavior, then shame aversion calls for investing resources in revealing such information even when it cannot be formally used in an externally enforced transaction. Clearly, the efficacy of shame will most likely be related to the actors with whom one transacts, and the social structure in which one does business.[29] Once again, the analogy between shame

as a crime deterent see "Shame, Stigma, and Crime: Evaluating the Efficacy of Shaming Sanctions in Criminal Law," Harvard Law Review, Vol. 116, No. 7 (May, 2003), pp. 2186-2207.

[29] If the engagement is one of a repeated kind then information that cannot be taken to court can still play an important role in supporting repeated-game like strategies that enforce adequate behavior. Granovetter (1985) addresses the issue of trust and malfeasance in transactions and argues that the selfish-rational model of economics, and the over-socialized views of "generalized morality" (agents completely internalize social norms), are both inadequate. Instead, his "embeddedness" theory argues that "the on-going networks of social relations between

aversion and reputational concerns is apparent.

References

Akerlof, George A. and Rachel E. Kranton (2000) "Economics and Identity," *Quarterly Journal of Economics*, **115(3)**:715-753.

Andreoni, James (1989) "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, **97(6)**:1447-1458.

Andreoni, James (1990) "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving," *Economic Journal,* **100(401)**:464-477.

Andreoni, James and B. Douglas Bernheim (2009) "Social Image and the 50-50 Norm," *Econometrica*, **77(5**):1607–1636.

Andreoni, James and Justin Rao (2011) "The Power of Asking: How Communication Affects Selfishness, Empathy, and Altruism," forthcoming, *Journal of Public Economics.*

Ariely, Dan, Bracha, Anat, and Meier, Stephan (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, **99(1)**:544–555.

Battigalli, Pierpaolo and Martin Dufwenberg (2007), "Guilt in Games," *American Economic Review Papers & Proceedings*, **97(2)**:170-76.

Battigalli, Pierpaolo and Martin Dufwenberg (2008), "Dynamic Psychological Games," *Journal of Economic Theory*, **144(1)**:1-35.

Baumeister, Robert F. (1982) "A Self-Presentational View of Social Phenomena," *Psychological Bulletin*, **91(1)**:3-26.

Benabou, Roland and Jean Tirole (2006) "Incentives and Prosocial Behavior," *American Economic Review*, **96(5)**:1652-1678.

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

Bernheim, B. Douglas (1994) "A Theory of Conformity," *Journal of Political Economy*, **102(4)**:841–877.

Bloom, Martin (2003) "Editorial—Primary Prevention and Education: An Historical Note on Robert Owen," *Journal of Primary Prevention*, **23(3)**:275-281.

Bohnet, Iris, and Bruno S. Frey (1999): "The Sound of Silence in Prisoner's Dilemma and Dictator Games," *Journal of Economic Behavior & Organization*, **38(1)**:43–57.

Bolton, Gary E. and Axel Ockenfels (2000) "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, **90(1)**:166-193.

Broberg, Tomas, Tore Ellingsen and Magnus Johannesson (2007) "Is Generosity Involuntary?" *Economics Letters*, **94(1)**: 32–37.

people discourage malfeasance." However, embeddedness theory acknowledges that social networks alone will not deter malfeasance. It may be that shame aversion can offer a mechanism through which some of the ideas of embeddedness operate.

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

Charness, Gary and Martin Dufwenberg (2011) "Participation," forthcoming, *American Economic Review.*

Dana, Jason, Daylian M. Cain and Robyn Dawes. (2006) "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games," *Organizational Behavior and Human Decision Processes,* **100(2)**:193-201.

Dufwenberg, Martin and Georg Kirchsteiger. (2004) "A Theory of Sequential Reciprocity," *Games and Economic Behavior,* **47(1)**:268-98.

Ellingsen, Tore and Magnus Johannesson. (2008) "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, **98(3)**, 990-1008, 2008.

Ellingsen, Tore, Magnus Johannesson, Gaute Torsvik and Sigve Tjøtta (2010) "Testing Guilt Aversion, " *Games and Economic Behavior* **68(1)**:95-107.

Fehr, Ernst and Klaus M. Schmidt (1999) "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics,* **114(3)**:817-68.

Frank, R.H. (1987) "If Homo Economicus Could Choose His Own Utility Function, Would He Choose One With a Conscience?" *American Economic Review*, **77(4)**:593-604.

Fukuyama, Francis (1995) *Trust: The Social Virtues and the Creation of Prosperity,* New York: The Free Press.

Funk, Patricia (2010) "Social Incentives and Voter Turnout: Theory and Evidence," *Journal of the European Economic Association,* **8(5)**:1077-1103.

Geanakoplos, John, David Pearce and Ennio Stacchetti (1989), "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, **1(1)**:60–79.

Glazer, Amihai and Kei A. Konrad (1996) "A Signaling Explanation for Charity," *American Economic Review*, **86(5)**:1019–1028.

Granovetter, Mark (1985) "Economic Action and Social Structure: the Problem of Embeddedness," *American Journal of Sociology*, **91(3)**:481-93.

Guiso, Luigi, Paola Sapienza and Luigi Zingales (2008) "Trusting the Stock Market," *The Journal of Finance*, **63(6)**:2557-2600.

Heifetz, Aviad, Chris Shannon and Yossi Spiegel (2007) "What to Maximize if You Must," *Journal of Economic Theory*, **133(1)**:31-57.

Hofman, Elizabeth, Kevin McCabe, and Vernon Smith (1996) "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review*, **86(3)**:653–660.

Ireland, Nornam, J. (1994) "On Limiting the Market for Status Signals," *Journal of Public Economics*, **53(1)**:91–110.

Knack, Stephen and Philip Keefer "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation," *Quarterly Journal of Economics*, **112(4)**:1251-1288.

Koch, Alexander K., and Hans T. Normann (2008) "Giving in Dictator Games: Regard for Others or Regard by Others?" *Southern Economic Journal*, **75(1)**:223-231.

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

Lazear, Edward, Ulrike Malmendier and Roberto Weber (2010) "Sorting, Prices, and Social Preferences," NBER Working Paper No. 12041.

Levine, David K. (1998) "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, **1(3)**:593–622 .

List, John A. (2007) "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, **114(1)**:1-37.

Mailath, George J. and Larry Samuelson (2006) *Repeated Games and Reputations: Long-Run Relationships*, Oxford University Press, Oxford and New York.

Mas, Alexandre and Enrico Moretti (2009) "Peers at Work," *American Economic Review*, **99(1)**:112–45

Posner, Richard A. (2001) *Frontiers of Legal Theory,* Harvard University Press, Cambridge, MA.

Rabin, Matthew (1993) "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, **83(5)**:1281-1302

Soetevent, Adriaan R. (2005) "Anonymity in Giving in a Natural Context – A field Experiment in 30 Churches," *Journal of Public Economics* **89(11-12)**:2301– 2323.

Tangney, June (1995) "Recent Advances in the Empirical-Study of Shame and Guilt," *American Behavioral Science*, **38(8)**:1132-45.

Williamson, Oliver E. (1985) *The Economic Institutions of Capitalism*, New York: Free Press.

Appendix A: Proofs

Proof of Proposition 1: The proof is straightforward and follows from Bayes updating. With exposure, $E_1[C|C,1] = 1$ and $E_1[C|D,1] = 0$. This implies that a type s receives c from choosing C, while he receives $d - s$ from choosing D, implying that D is a strict best response for all $s < s^1 = d - c$ and C is a strict best response for all $s > s^1$. □

Proof of Proposition 2: With no exposure player 1's beliefs satisfy $E_1[C|G,0] = 1 \geq E_1[C|B,0]$. A pooling equilibrium where all types (measure 1) choose $\sigma_s = 0$ cannot exist because if all types choose D then $E_1[C|B,0] = 0$, but then all types $s > d - c$ will prefer to choose C, a contradiction. A pooling equilibrium where all types choose $\sigma_s = 1$ cannot exist either because if all types choose C then $E_1[C|B,0] = 1$, but then all types will prefer to choose D, a contradiction. It follows that in any equilibrium both C and D are chosen by a positive measure of types and $1 > E_1[C|B,0] > 0$. Given any such beliefs of player 1, player 2's expected utility from d is monotonically decreasing in type s implying that if some type s chooses to cooperate then all types $s' > s$ will cooperate, and if some type s chooses to defect then all types $s' < s$ will defect as well. It follows that any equilibrium must have a cutoff type $s^0 > 0$ so that $\sigma_s = 0$ for all $s < s^0$ and $\sigma_s = 1$ for all $s > s^0$. The cutoff type must be indifferent between choosing C and D implying the equilibrium equation

$$c - (1-p)(1 - E_1[C|B,0])s^0 = d - (1 - E_1[C|B,0])s^0,$$

giving the cutoff type

$$s^0 = \frac{d-c}{1 - E_1[C|B,0]} > s^1.$$

Last, to prove that this equilibrium is unique, notice that in any equilibrium with a cutoff type s^0, $E_1[C|B,0]$ depends on s^0. When $s^0 = 0$ then correct beliefs imply that $E_1[C|B,0] = 1$, and $E_1[C|B,0]$ is monotonically decreasing in s^0 with $\lim_{s^0 \to \infty} E_1[C|B,0] = 0$. Hence, there is a unique s^0 that satisfies the equilibrium equation above. □

Appendix B: Experiment Instructions

Thank you for participating in this session. The purpose of this experiment is to study how people make decisions in a particular situation. There will be time for questions after the explanation. Please do not speak to other participants during the experiment.

You will receive $7 for participating in this session. You may also receive additional money, depending on the decisions made (as described below). Upon completion of the session, this additional amount will be added to the $7 fee and the total will be paid to you individually and privately.

During the session you will have several decisions to make. For each decision you will be paired with another person randomly, and the random pairing will be reshuffled for each of the decisions. For some decisions you will not know who you are paired with, while for others you will.

Decision tasks :

In each pair, one person will have the role of A, and the other will have the role of B. The amount of money you earn depends on the decisions made in your pair.

First persons A will make their choices. On the designated decision sheet, each person A will indicate whether he or she wishes to choose IN or OUT. If A chooses OUT, A and B each receives $5. We will collect these sheets after the choices have been indicated.

Second, persons B will indicate whether he or she wishes to choose ROLL or DON'T ROLL (a die). Note that B will not know whether his paired A has chosen IN or OUT; however, since B's decision will only make a difference when A has chosen IN, we ask B's to presume (for the purpose of making this decision) that A has chosen IN. B's will then turn over their decision sheets.

Third, I will pass by each B and roll a six-sided die, recording the number 1 through 6 on the reverse side of the decision sheet, without observing the decision. Then, these sheets will be collected, and matched to the collected sheets from the A persons.

If A has chosen IN and B chooses DON'T ROLL, then B receives $14 and A receives $0. If A chose IN and B chooses ROLL, B receives $10 and the roll of the die determines A's payoff. If the die comes up 1, A receives $0; if the die comes up 2-6, A receives $12. (All of these amounts are in addition to the $7 show-up fee.) The payoff information from the pair of tasks is summarized in the chart below:

Decisions	A receives	B receives
A chooses OUT	$5	$5
A chooses IN, B chooses DON'T ROLL	$0	$14
A chooses IN, B chooses ROLL, die = 1	$0	$10
A chooses IN, B chooses ROLL, die = 2,3,4,5, or 6	$12	$10

Sometimes A's who receive $0 for a given pair of decisions will be told whether their paired person chose DON'T ROLL or whether they chose ROLL and the die roll was 1. Your final payoff will be determined by randomly choosing one of the outcomes that you participated in, and adding that to your $7 show-up fee.

Table 2: Percent of B-players who chose Roll (Cooperate)

Treatment	actual sessions, all treatments				first two treatments only			
	Obs	Mean	Std. err.	95% conf.	Obs	Mean	Std. err.	95% conf.
AN	85	.2	.043	[.121,.301]	15	.33	.122	[.118,.616]
AE	85	.376	.053	[.274,.488]	49	.35	.068	[.217,.496]
MN	85	.4	.053	[.295,.512]	17	.29	.111	[.103,.560]
ME	85	.753	.047	[.647,.840]	55	.75	.059	[.610,.853]
AP	51	.725	.062	[.583,.841]	8	.75	.153	[.349,.968]
MP	51	.784	.058	[.647,.887]	9	.89	.105	[.518,.997]

Table 3: Behavior of B players

Linear Probability Model, Dependent variable: Player B chose Roll (Cooperated)			
	(1)	(2)	(3)
AN	0.200	-	0.196
	(0.044)		(0.057)
AE	0.377	-	0.411
	(0.053)		(0.070)
MN	0.400	-	0.431
	(0.054)		(0.071)
ME	0.753	-	0.804
	(0.047)		(0.057)
AP	-	0.725	0.725
		(0.063)	(0.064)
MP	-	0.784	0.784
		(0.058)	(0.059)
Test	F (4,84)	F(2,50)	F(6,50)
F-stat value	66.65	128.72	66.62
Significance	0.0000	0.000	0.000
N	340	102	306
No. of Clusters	85	51	51

Robust standard errors are in parentheses
(clustered at the individual level)

Table 4: Percent of A-players who chose In (Trust)

Treatment	actual sessions, all treatments				first two treatments only			
	Obs	Mean	Std. err.	95% conf.	Obs	Mean	Std. err.	95% conf.
AN	86	.302	.050	[.208,.411]	16	.375	.121	[.152,.646]
AE	86	.558	.054	[.447,.665]	51	.588	.069	[.441,.724]
MN	86	.547	.054	[.435,.654]	16	.563	.124	[.229,.802]
ME	86	.744	.047	[.639,.832]	54	.815	.053	[.686,.907]
AP	53	.660	.065	[.517,.785]	9	.556	.166	[.212,.863]
MP	53	.736	.061	[.597,.847]	9	.889	.105	[.518,.997]

Table 5: Behavior of A-players

Linear Probability Model, Dependent variable: Player A chose In (Trust)			
	(1)	(2)	(3)
AN	0.302	-	0.340
	(0.05)		(0.066)
AE	0.558	-	0.585
	(0.054)		(0.069)
MN	0.547	-	0.604
	(0.054)		(0.068)
ME	0.744	-	0.660
	(0.048)		(0.066)
AP	-	0.660	0.660
		(0.066)	(0.066)
MP	-	0.736	0.736
		(0.061)	(0.062)
Test	F (4,85)	F(2,52)	F(6,52)
F-stat value	80.09	75.84	38.18
Significance	0.0000	0.000	0.000
N	344	106	318
No. of Clusters	86	53	53

Robust standard errors are in parentheses
(clustered at the individual level)

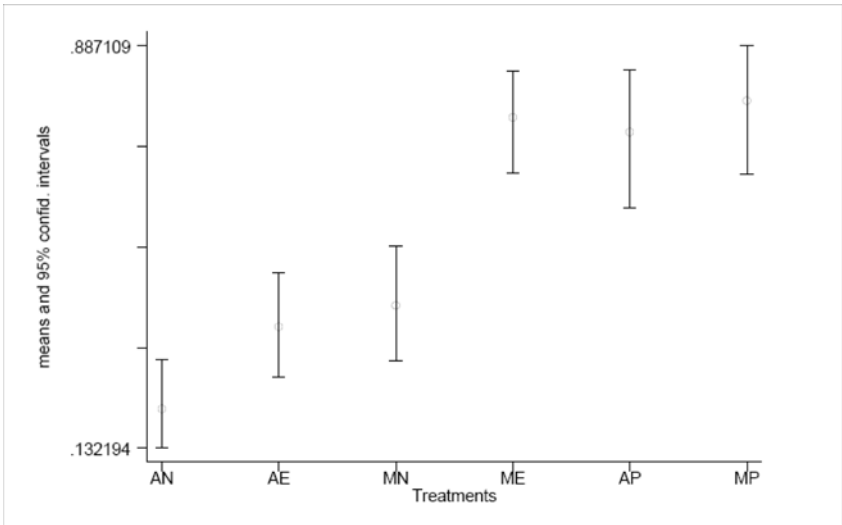

Figure 2: Percentage of B-Players Who Choose Cooperate in the Six Treatments

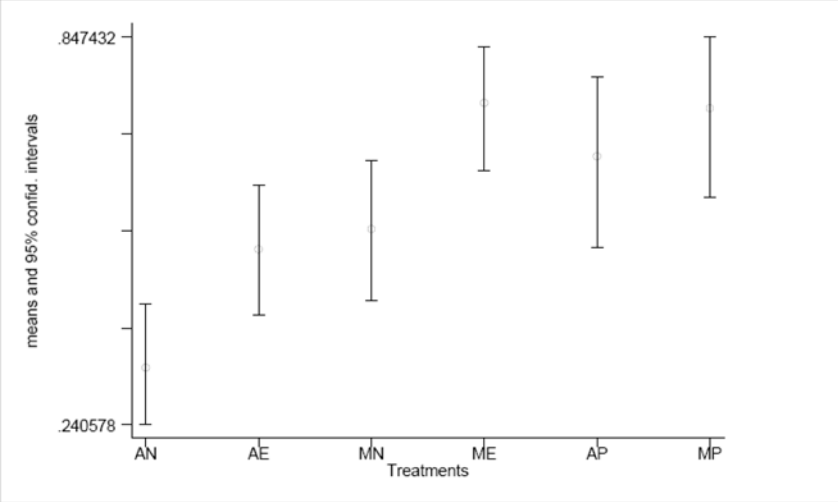

Figure 3: The Precentage of A Players Who Chose Trust in the Six Treatments