

# The Power of Shame and the Rationality of Trust

Steven Tadelis\*  
UC Berkeley  
Haas School of Business

April 2, 2008

## Abstract

Experimental evidence and a host of recent theoretical ideas take aim at the common economic assumption that individuals are selfish. The arguments made suggest that intrinsic “social preferences” of one kind or another are at the heart of unselfish, pro-social behavior that is often observed. I suggest an alternative motive based on “shame” that is imposed by the extrinsic beliefs of others, which is distinct from the more common approaches to social preferences such as altruism, a taste for fairness, reciprocity, or self-identity perception. The motives from shame are consistent with observed behavior in previously studied experiments, but more importantly, they imply new testable predictions. A new set of experiments confirm both that shame is a motivator, and that trusting players are strategically rational in that they anticipate the power of shame. Some implications for policy and strategy are discussed. *JEL* classifications C72, C91, 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 for detailed comments on an earlier draft. I also thank Yossi Feinberg, Uri Gneezy, Shachar Kariv, Navin Kartik, Mike Katz, Jim Lincoln, John List, Barbara Mellers, John Morgan, Muriel Niederle, Larry Samuelson, 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 and by UC Berkeley’s X-Lab at the Haas School of Business.

# 1 Introduction

Despite its simple structure and extreme assumptions, the selfish rational choice model that has been the work-horse of economic analysis works surprisingly well at the market level. However, when put to tests of individual decision making, human behavior departs from theoretical predictions in one notable and systematic way: individuals often exhibit “pro-social” behavior, in which they sacrifice some of their own monetary payoff to increase (and sometimes decrease) the payoff of others. (See Camerer 2003 for an excellent summary of these results.)

Beyond the simple experimental settings, pro-social behavior is manifested in daily life by a variety of so called acts of kindness and contributions to the public good. This suggest that individuals have “other-regarding” preferences, for which many mechanisms have been proposed. Prominent among these are altruism (direct and Indirect, e.g., Andreoni, 1990); inequity Aversion (Fehr and Schmidt, 1999); preferences for fairness and reciprocity (Rabin, 1993, Dufwenberg and Kirchsteiger, 2004); regards for self-identity (Akerlof and Kranton, 2000), or combinations of intrinsic and reputational concerns (Benabou and Tirole, 2006; Levitt and List, 2007).

Many of these proposed explanations are primarily driven by intrinsic considerations—they are an expression of one’s preferences over outcomes that include the payoffs of others. Such preferences can be viewed as a minor conceptual departure from the standard selfish model in that the payoffs of others are components of the relevant outcomes over which preferences are defined. Loosely speaking, such intrinsic mechanisms might fit under the broad title of “guilt”: being selfish causes a player to experience some intrinsic loss of utility. This definition of guilt, though broad, is useful to categorize a set of preferences where *only the outcomes* matter through some *intrinsic preference ordering*.<sup>1</sup>

In a similarly loose manner, to be made precise soon, I refer to “shame” mechanisms as a *distinct* motivator from guilt mechanisms. 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.” This distinction, which has been used extensively by psychologists<sup>2</sup>, seems to imply that from a decision theoretic perspective preferences that incorporate shame must present a concern over the perception, or beliefs of others, above and beyond any intrinsic preferences over

---

<sup>1</sup>This is different from a recent description of preferences incorporating guilt as modelled by Battigalli and Dufwenberg (2007, 2008). They define guilt as a loss of utility for player  $i$  who believes that he either disappointed player  $j$  (simple guilt), or how much  $j$  believes that  $i$  believes that he lets  $j$  down (guilt from blame).

<sup>2</sup>Tangney (1995) offers a competing view that she advocates regarding one’s perception of oneself (shame) versus that of one’s actions (guilt), and a long list of references dating back to the 50’s that discuss these two approaches. A self-help teen website put this approach in more accessible language by stating that “Guilt and Shame are closely connected emotions, we tend to feel guilty when we have violated rules or not lived up to expectations and standards that we set for ourselves. If we believe that we “should” have behaved differently or we “ought” to have done better, we likely feel guilty. Shame involves the sense that we have done something wrong that means we are “flawed,” “no good,” “inadequate,” or “bad” and is usually connected to the reactions of others. Anytime you catch yourself thinking ‘if they knew \_\_\_\_\_ then they would not like me or would think less of me,’ you are feeling shameful.” [http://www.teenhealthcentre.com/articles/publish/article\\_93.shtml](http://www.teenhealthcentre.com/articles/publish/article_93.shtml)

physical outcomes and payoffs.<sup>3</sup> This paper takes a step towards incorporating such preferences into an otherwise standard model of rational choice, and offers answers to three related questions.

First, if both shame-based (extrinsic) and guilt-based (intrinsic) preferences can motivate pro-social behavior, is it possible to empirically distinguish these two channels of motivation? I demonstrate that the answer is yes. The idea is that one can manipulate the ability of others to form their perception of an individual's behavior *without* changing the way in which the individual's actions affect the distribution of physical outcomes. As I will argue, this offers a clean experimental design to test whether shame has motivating power.

Second, if shame is confirmed as a motivator, do players act *as if* they perceive the power of shame in the behavior of others so as to direct their own choices? This is a subtle and important question because the premise of equilibrium analysis in rational choice models is that players not only understand the consequence of their own actions, but that they have a rationally informed theory about the way in which other players make their decisions. Using a standard "game of trust," I show that players rationally anticipate the behavior of others, and as such, adapt their own behavior in a way that is consistent with rational choice equilibrium analysis, validating the equilibrium approach that game theory subscribes to in such settings.

Finally, as shame is established as a motivator, and as people seem to understand its consequences, a natural question is what are the public policy and strategy implications for the design of public and private institutions? I offer some directions through which this channel of motivation can inform the way firms and individuals interact, as well as some public policy questions that can be influenced when taking such preferences into account.

The idea behind distinguishing extrinsic preferences (shame) from intrinsic preferences (guilt) is simple, and can be described by the following caricature. Imagine that there is an outdoor restaurant on the oceanfront at which tips are only accepted cash. It is known that the wind is sometimes strong enough to blow away part or all of the cash into the ocean where it is lost forever. Understanding this, you decide how much to leave on the table after your dining experience. Hence, when the waiters collect their loot, they sometimes experience generous tips and other times have no tips at all, without the ability to know for sure if it was a stingy customer or a nasty gust of wind. Now imagine that the restaurant installs cameras that are focused on the center of each table and accurately record the amount of tips left on the table and display it to the waiters. The physical environment, however, is unchanged. Hence, if you are purely motivated by intrinsic pressures then your behavior should be unchanged after cameras are installed because they have no effect on how your actions result in the final outcomes. However, if you care about what the waiter thinks about you, then with the cameras he will know for sure how much you left behind and you will not be able to hide behind the excuse that "the wind blew the money away." In this lies the testable hypothesis: if your behavior changes, it must be that you care

---

<sup>3</sup>Preferences for fairness and reciprocity (Rabin, 1993, Dufwenberg and Kirchsteiger, 2004) are often modeled by players caring not only about the material payoffs of others, but also about their intentions and beliefs about how they themselves behave. As will become apparent, the way in which beliefs of others play a role in "shame" is rather distinct from these concerns of fairness and reciprocity.

about the waiter’s beliefs about you (or of the beliefs of others who have access to the camera footage). The next step, which is to test whether people are strategic rational, is by considering the behavior of the waiters. If cameras are installed, and they infer that their better service is more likely to be rewarded, then they will improve their service after the cameras were installed even when these devices only record the behavior of their patrons.

In Section 2 I introduce a simple game of trust (in the spirit of Berg, Dickhaut and McCabe, 1995) to formally capture preferences for shame together with standard equilibrium analysis. To formally model how the perception of others matters to an individual, I resort to the well established game theoretic structure using a prior distribution of “types” for players, from which ex post beliefs are formed. I endow some players with preferences over the beliefs of others, implying a “psychological game” as introduced by Geanakoplos et. al. (1989). However, the tools introduced by Geanakoplos et. al. to treat the dynamic nature of the game do not adequately capture sequential rationality. For this reason I adapt a rather common notion of sequential equilibrium in the spirit of the recent work of Battigalli and Dufwenberg (2007).<sup>4</sup>

In the game, player 1 can either trust player 2, or can exit with a safe outside option. If trusted, player 2 can return in kind by cooperating, but at a sacrifice of additional funds that he receives from defecting. A selfish player 2 will therefore never cooperate, and anticipating this player 1 should never trust. Noise is added to introduce an element of luck, as in Charness and Dufwenberg (2006): following trust by player 1 and cooperation by player 2, there is a small chance that player 1 receives the same low payoff that would result from player 2 defecting. A second layer of noise can then be manipulated, which is whether player 1 *observes* the actual actions of player 2, or whether he just observes his own final payoffs without observing player 2’s choice. In the former case, player 1 knows exactly what player 2 did, while in the latter, he cannot be certain whether a low payoff is the result of player 2 defecting, or the result of bad luck. In this lies the method for identifying if the belief of others affects one’s behavior.

In particular, if player 2 is motivated by extrinsic preferences (over the beliefs of player 1) then the theory suggests a testable implication on the behavior of player 2. Namely, if player 1 can establish whether bad outcomes correspond to selfish behavior of player 2 then player 2 should be more likely to act cooperatively. This should not happen if the sole motivation for pro-social behavior is intrinsic. This offers an answer to the first question posed in this study: extrinsic preferences driven by some form of “shame” can be distinguished from intrinsic preferences driven by some form of “guilt.”

Answering the second question posed above, on the ability of players to correctly predict the behavior of others, is a consequence of equilibrium analysis. If player 2 is indeed motivated by extrinsic preferences then a *rational* player 1 will anticipate the change in player 2’s incentives following a change in the informational environment. If the informational environment implies that player 2 is more likely to cooperate, then player 1 should be more inclined to trust player

---

<sup>4</sup>The framework developed by Battigalli and Dufwenberg (2008) does not include incomplete information, yet as they discuss in Section 6.2 of their paper, it naturally extends to it.

2. Thus, higher levels of trust can be explained as a rational response to the incentives provided by extrinsic preferences, and finding this behavioral response would imply that rational players act *as if* they perceive the power of shame.

This simple theoretical exercise is then applied to a series of laboratory experiments in which information is manipulated along the lines described above. The experimental results strongly support the hypotheses that are derived from the theoretical analysis. Thus, without ruling out that intrinsic motivations explored in previous studies are drivers of pro-social behavior, the motivation implied by extrinsic preferences is very strong and robust, as is the equilibrium response of rational players. This is in line with two complementary papers by Andreoni and Bernheim (2007) and Dana et. al. (2006) who also confirm the effect of extrinsic preferences in an alternative setting of a dictator game, but do not address the so-called best response of rational players to others with these preferences.

Following the experimental analysis, a discussion focuses on three main issues. First, shame-based preferences can be considered as a reduced form of more complex reputational preferences that players would exhibit in repeated games with incomplete information (e.g., Kreps et. al., 1982). Despite the fact that players are playing a one-shot game, they may perceive the situation as part of a repeated game, either rationally or as a rule of thumb, which is consistent with some of the experimental results. To what extent the players can rationally consider the experimental interaction as part of a repeated game is a judgement call, but several results shed doubt on this interpretation. It may be, however, that such preferences may be a result of evolutionary pressures that began with a form of reputational concerns.

A second issue raised in the discussion is the ability of extrinsic preferences to explain the behavior observed in other experimental settings, and what kinds of new experiments would the theory suggest as good testing grounds. Finally, I offer some thoughts on the implication of extrinsic preferences to both private sector strategy (contractual relationships) and public policy, which is an attempt to give some partial answers to the final question that this study raises.

The paper contributes to the growing literature in economics that is mentioned above, which tries to identify motives for pro-social behavior. It also contributes to two established literatures in social psychology. The first is a vast literature that studies “Self-presentation,” which is concerned with an individual’s concern about constructing his or her “public self”, either to please the audience they interact with or to become one’s ideal self through the eyes of others. 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 others.” This paper is very much related to this line of research, with the extra step of embedding behavior in an rational strategic setting.

The second literature tries to identify the difference between shame and guilt (see, e.g., Tangney, 1995 and Niedenthal et. al. 1994). Using narrative situations described to subjects, the

research tries to identify the difference between guilt and shame by the emotions that people claim the narrative gives rise to personally. This of course assumes that when one person identifies as situation as imposing “shame” then he or she means what the researcher has in mind. (See Smith et. al. (2002) for a discussion of this issue, and for related results.) This issue of subjectivity is circumvented in my study since it focuses *only on the behavior* of subjects in two environments that differ in the amount of behavioral exposure. This is consistent with the psychological view that shame is a publicly driven emotion, or one that is generated by a feeling of inadequacy in the eyes of others. Note, though, that I refrain from carefully defining guilt, nor do I offer treatments that vary any amount of personal intrinsic pressures that may be associated with guilt, however broadly defined.

## 2 Modelling Shame: A Noisy Trust Game

In his section I offer a precise definition of extrinsic preferences that are motivated by psychological descriptions of “shame” as such an extrinsic preference. As described in the introduction, shame is defined as an emotion that affects behavior through the anticipation of what others will think about the player in question. Hence, it is natural to assume that the beliefs of others include some undesirable trait, or undesirable behavior, that they can attribute to the player in question. The incentive mechanism follows from the player’s aversion to being thought of as having that trait or acting in an inadequate way. This is the approach that will be spelled out in what follows, where players will have some trait of how “shameful” they are, and this will drive their behavior.

In contrast to the approach taken here, Battigalli and Dufwenberg (2007) introduce preferences that depend on the beliefs of others in a different way. Instead of having “shameful” types and incomplete information, they endow players with expectations over monetary outcomes, and these beliefs form the basis for disappointment (i.e., receiving less than you expect.) Thus, the preferences of players in Battigalli and Dufwenberg (2007) are over money and over the fulfilled or disappointed expectations of others about outcomes (or in a more complex framework, beliefs over beliefs about the disappointment of others).

To offer a general model that introduces such “shameful” types, however, may be context dependent. This is particularly true if one views preferences over the beliefs of others as a reduced-form rule of thumb that replaces more complex behavior in repeated games where reputational concerns are present. I defer this discussion to section 6.1, and turn to a simple trust game to illustrate the idea of extrinsic preferences for shame, and the behavioral consequences and predictions if such preferences are present. Section 6.2 expands on how to consider further extensions in a variety of different games.

## 2.1 Perfect Information and Monetary Payoffs

Imagine a simple trust game in which 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 of zero, identical to his payoff from defection. Conditional on cooperating, the pecuniary payoffs to player 2 do not depend on success. This trust-game has the structure used in the experiments of Charness and Dufwenberg (2006), and with perfect information can be described by the game in Figure 1.

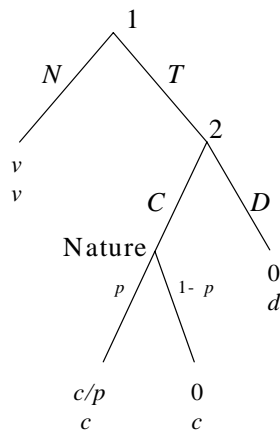


Figure 1: A Simple Trust Game

In this game, if player 1 chooses  $N$  then both players receive a payoff of  $v > 0$ , while if player 1 chooses  $T$  and player 2 chooses  $C$  then both get an expected payoff of  $c > v$ . Following trust, the cost of cooperating for player 2 is  $d - c > 0$ . These payoffs imply that the game has a unique equilibrium: player 2 will never cooperate if he is trusted since  $d > c$ , and in turn player 1 will never trust since  $v > 0$ .

Unlike most trust games (see, e.g., Camerer 2003), there is added noise by having Nature move after player 2's cooperative action. The effect is that even when player 2 acts cooperatively, there is some chance that player 1 will receive the same low payoff he gets from player 2's choice of defection. Hence, if player 1 were only to observe his own payoffs, a payoff of 0 can be attributed to either selfish behavior of player 2, or just bad luck.<sup>5</sup>

<sup>5</sup>This “technology” is present in Charness and Dufwenberg (2006), but they do not make explicit use of the noise in their theoretical or experimental analysis. Instead, they use it as a form of “hidden action” in that it relates the experiment to the standard principal-agent model.

## 2.2 Incomplete Information and Preferences with Shame

To capture pro-social behavior, I add a simple layer of incomplete information. For simplicity, player 1 is assumed to be risk neutral and selfish in the sense that only pecuniary payoffs matter to him,<sup>6</sup> so that if player 1 receives an expected amount  $m_1$  then his utility is  $u_1 = m_1$ .

Player 2 is also assumed to be risk neutral, but in contrast, has pro-social preferences. As argued in the introduction, one can think of either intrinsic preferences (guilt) or extrinsic preferences (shame) as being distinct emotions that motivate pro-social behavior. I formalize the difference between guilt and shame as follows: *guilt* is associated with the intrinsic cost of “cheating” player 1 in that player 1 is harmed, or potentially harmed by player 2’s action (or, as in Battigalli and Dufwenberg’s “simple guilt,” is disappointed from 2’s actions.) This, for example, can be captured with altruism, inequity aversion or other mechanisms that make player 2 care about the *outcomes* of player 1. *Shame*, in contrast, is associated with the extrinsic cost of having player 1 *believe* that player 2’s behavior was inadequate, or that player 2 is himself inadequate in some way, regardless of the actual outcomes that player 1 is dealt.

Guilt is a hard dimension to manipulate externally without changing the set of outcomes, and hence may be hard to identify on its own. Furthermore, manipulating a person’s intrinsic guilt propensity would be challenging.<sup>7</sup> As I will soon show, it is possible to manipulate a player that is extrinsically motivated by shame. As such, I introduce shame aversion as the sole source of pro-social preferences and ignore guilt related motivation all together.

Formally, player 2 cares about money and about player 1’s beliefs about the strategy that player 2 actually chose. Let  $E[\sigma] \in [0, 1]$  be the posterior belief of player 1 about the probability that player 2 chose to cooperate. Player 2 of type  $s$  that receives payment  $m_2$  and whose opponent’s ex post belief is  $E[\sigma]$  has utility

$$u_2 = m_2 - (1 - E[\sigma])s,$$

where  $s$  is player 2’s *shame aversion*. 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. As a consequence, player 2 is not averse to defecting, but is averse to the beliefs that others have about his possible defection. This is how extrinsically motivated shame is differentiated in the model from intrinsically motivated guilt. For convenience, I assume that  $s$

---

<sup>6</sup>Of course, one can argue that by not trusting player 2, player 1 may be acting in an offensive way, which should potentially result in feelings of guilt or shame. I ignore this possibility for simplicity, though at a later stage this can be incorporated into a richer set of experiments. In particular, noise can be added after player 1 chooses “trust” in a way that may cause termination of the game, so that a player 2 who is not called upon to move will not necessarily know whether player 1 was not-trusting, or whether trust was followed by a noisy exogenous termination.

<sup>7</sup>Psychologists have used the method of “priming” to try and modify the intrinsic feelings of individuals, so as to increase their feeling of guilt. (See, e.g., Zemack-Rugar, Bettman and Fitzsimons (2007) and the references therein.) It is unclear whether priming can differentially change the marginal “guilt” cost of selfish behavior, which would be necessary to evaluate guilt as an incentive for pro-social behavior.

is distributed over  $[0, \infty)$  with cumulative distribution  $F(\cdot)$  and positive density  $f(\cdot)$ .<sup>8</sup>

Two issues regarding my model of shame are worth noting. First, the specific form of multiplicative preferences that are linear in  $s$  are very convenient, but the results demonstrated below carry over to more general settings. A more general form of preferences with shame can be given by  $u_2(m_2, E[\sigma], s) = m_2 - \phi(E[\sigma], s)$ , where  $\phi(E[\sigma], s)$  increases in both components, so that either “harsher” beliefs or more shame aversion creates more disutility.

Second, notice that the model captures shame preferences as player 2’s aversion to player 1’s belief over 2’s action, and not 2’s type. Given the language used in the social psychology literature described earlier, it seems more adequate that shame is associated with player 1’s belief that player 2’s type is not desirable, as in “I think you’re a cheater” instead of “I think you have likely cheated me.” To capture this alternative, and possibly more realistic notion of shame, consider  $F^{post}(\cdot)$  to be the posterior belief of player 1 about the type  $s$  of player 2, which is derived from  $F(\cdot)$  and from player 1’s belief about the strategy of each type,  $\sigma_s$ ,  $s \in [0, \infty)$ . Now consider the payoffs of player 2 to be  $u_2(m_2, F^{post}(\cdot), s) = m_2 - \phi(F^{post}(\cdot), s)$  where  $\phi(F^{post}(\cdot), s)$  increases in both components as before, but now it increases in  $F^{post}(\cdot)$  in the sense of (inverse) first-order-stochastic-dominance. That is, less favorable beliefs about 2’s type (more likely to have a lower  $s$ ) will imply a higher value of  $\phi(F^{post}(\cdot), s)$ , as will a higher value of  $s$ . In this simple environment, these two notions of shame are isomorphic, and hence I resort to the formulation proposed above that is much simpler to use.

On a final point, notice that there is a departure from the standard definition of a game because payoffs are not functions of actions and types alone, but also of the beliefs that some players have about others. In particular, player 1’s belief about the action of player 2 directly effects the payoffs of player 2, which is the way in which the model captures extrinsic motivation. This is in the spirit of a small literature on “Psychological Games,” pioneered by Geanakoplos et. al. (1989), and developed further by Battigalli and Dufwenberg (2008).<sup>9</sup>

### 2.3 Exposure and Inference

To complete the structure of the game I introduce an exogenous “exposure technology.” This basically refers to player 1’s ability to decipher the noisy outcome of a payoff of 0, and attribute it either to bad luck, or to an uncooperative action by player 2. For convenience, I refer to a “Good” outcome ( $G$ ) as the outcome in which player 1 gets a payoff of  $c/p$ . (This is a consequence

---

<sup>8</sup>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.

<sup>9</sup>In Charness and Dufwenberg (2006) the beliefs of player 1 also enter into the preferences of player 2, but their notion of “guilt aversion” is based on 2’s incentives not to *disappoint* player 1 (the notion of “simple guilt” in Battigalli and Dufwenberg, 2008), which is different from the notion of “shame” I introduce here. It is in fact possible to explain the behavior demonstrated in their experiments without resorting to the beliefs of others, and using some form of preferences over breaking promises. This, for example, could be captured by preferences over outcomes alone if ex post outcomes include the messages sent by players earlier in the game, and some players do not like to go back on their proposed behavior (the latter now being part of the outcome space.)

of player 1 choosing to trust, player 2 choosing to cooperate, and nature being favorable.) I refer to a “Bad” outcome ( $B$ ) as the outcome of player 1 receiving a payoff of 0, which is either a consequence (following trust of player 1) of player 2 defecting, or of player 2 cooperating and nature being unfavorable.

I say that the game exhibits *full exposure* ( $e = 1$ ) if player 1 observes the action of player 2. With full exposure, if player 1 chooses to trust player 2, and the outcome is a failure, then player 1 will learn whether the payoff of 0 is due to player 2 choosing  $D$ , or whether it was because of nature choosing the payoffs  $(u_1, u_2) = (0, c)$  following player 2’s choice of  $C$ . The game exhibits *no exposure* ( $e = 0$ ) if player 1 learns nothing about the reason for failure. (Recall that after a success player 1 must correctly infer that player 2 cooperated.)

The set-up can be generalized to intermediate levels of exposure. Namely, one can introduce an exposure technology  $\tau \in [0, 1]$  as follows: with probability  $\tau$  there is full 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.

## 2.4 Equilibrium Analysis

I consider a natural adaptation of sequential equilibrium to the setting of this game where in each subgame, player’s are playing a best response to their beliefs, and these beliefs are consistent with Bayes’ rule.<sup>10</sup> The best response of a “good” player 2 depends on his payoffs, which in turn depend on his actions and on the beliefs of player 1 (more precisely, the beliefs of player 2 *about* the beliefs of player 1). Equilibrium analysis will require the beliefs of player 1 about the type and actions of player 2 to be correct, and that player 2’s beliefs about the beliefs of player 1 are correct as well.<sup>11</sup>

Let  $\sigma_s \in [0, 1]$  be the probability that a player 2 of types  $s$  chooses  $C$  (the strategy of type  $s$ ), and let  $E[\sigma|h, e]$  be the posterior probability that player 1 assigns to player 2 choosing  $C$  conditional on the history  $h \in \{C, D, G, B\}$  and the game’s exposure  $e \in \{0, 1\}$ . The history set encompasses the fact that with full exposure, history can be restricted to  $C$  and  $D$ , the actions of player 2, while with no exposure, history is restricted to  $G$  and  $B$ , since player 1 only observes monetary payoffs but not actions.<sup>12</sup>

It is convenient to first describe some facts about player 1’s conditional posterior belief. Clearly, following a good outcome player 1 perfectly infers that player 2 cooperated, hence

---

<sup>10</sup>The analysis in Battigalli and Dufwenberg (2007, 2008) suggests that adopting the notions of sequential equilibrium to these kinds of games is indeed valid. They also show that sequential equilibria exist in these games, and address the issue of hierarchies of beliefs as well.

<sup>11</sup>Unlike Battigalli and Dufwenberg (2008), higher order beliefs will not play a role in the game analyzed, and as such are ignored.

<sup>12</sup>A complete history would always include success or failure, and will include actions only when  $e = 1$ . However, if actions are known then outcomes have no effect on the relevant beliefs, and therefore the restriction is without loss.

$E[\sigma|G, e] = 1$  for  $e \in \{0, 1\}$ . Also, full exposure reveals the action of player 2, and hence  $E[\sigma|C, 1] = 1$  and  $E[\sigma|D, 1] = 0$ . Finally,  $E[\sigma|B, 0] \in [0, 1]$ , and furthermore, if  $\sigma_s > 0$  for a positive measure of types  $s$  and  $\sigma_s < 1$  for a positive measure of types then  $E[\sigma|B, 0]$  is strictly within the interior of the  $(0, 1)$  interval.

Since exposure is an environmental characteristic of the game, I will solve for the equilibrium behavior of the players for both cases of full and no exposure, and use the difference to describe the comparative statics of changing the level of exposure. I begin with the case of full exposure:

**Proposition 1** *If  $e = 1$  then there exists a unique sequential equilibrium characterized by a cut-off type  $\bar{s}^1 = d - c > 0$  such that all types  $s < \bar{s}^1$  choose  $D$  ( $\sigma_s = 0$ ) and all types  $s > \bar{s}^1$  choose  $C$  ( $\sigma_s = 1$ ).*

**Proof.** First, observe that behavior will exhibit monotonicity in types: if some type  $s$  chooses to cooperate, then all types  $s' > s$  will choose to cooperate as well, and if some type  $s$  chooses to defect, then all types  $s' < s$  will as well. Second, observe that we cannot have a pooling equilibrium where all types choose  $\sigma_s = 0$ . To see this, if all types choose  $\sigma_s = 0$  then it must be that

$$d - (1 - E[\sigma|D, 1])s \geq c - (1 - E[\sigma|C, 1])s \text{ for all types } s,$$

but since  $E[\sigma|C, 1] = 1$  and  $E[\sigma|D, 1] = 0$ , this inequality is violated for all types  $s < d - c$ , a contradiction. Similarly, all types cannot pool and choose  $\sigma_s = 1$  because the reverse violation will occur. This, together with the monotonicity of behavior, implies that there is a cut-off type. The argument that rules out pooling equilibria immediately identifies the cutoff type as  $\bar{s}^1 = d - c$ . ■

With full exposure, the unique equilibrium is obvious: since  $d > c$ , monetary payments provide an incentive to defect. However, since behavior is perfectly observed then shame aversion creates a cost to defection. Perfect observability implies that the cost of defection are independent of the measure of types who cooperate, and as such, all types  $s > d - c$  have a cost of defection that is higher than the monetary benefit.<sup>13</sup>

Similar forces are at work when there is no observability, and in particular, it is easy to see that once again behavior is monotonic in types. However, when actions are not observed by player 1 then the cost of defection depends on the beliefs of player 1 about the measure of types that cooperate. Still, the structure of the unique equilibrium is maintained as the following proposition shows.

**Proposition 2** *If  $e = 0$  then there exists a unique sequential equilibrium characterized by a cut-off type  $\bar{s}^0 > d - c$  such that all types  $s < \bar{s}^0$  choose  $D$  ( $\sigma_s = 0$ ) and all types  $s > \bar{s}^0$  choose  $C$  ( $\sigma_s = 1$ ).*

---

<sup>13</sup>A similar analysis would result from preferences over beliefs about types (what kind of person I am), rather than beliefs about actions (what I did). Namely, if all types like people to think that they are very shame averse, then shame aversion would make those who care more act cooperatively. The analysis would be somewhat more involved, but essentially carry the same flavor.

**Proof.** First, the same logic applied in the proof of proposition (1) implies that there cannot be a pooling equilibrium where all types choose  $\sigma_s = 0$  or  $\sigma_s = 1$ , and the monotonicity of behavior implies that the only candidate for a sequential equilibrium is one with a cut-off type,  $\bar{s}^0$ . Since player 1 only observes the outcome, there are only two relevant posteriors:  $0 < E[\sigma|B, 0] < E[\sigma|G, 0] = 1$ , where both inequalities are a consequence of the fact that some, but not all types are choosing  $C$ . To verify that a cutoff equilibrium exists there must be some type  $\bar{s}^0$  such that<sup>14</sup>

$$d - (1 - E[\sigma|B, 0])\bar{s}^0 = c - (1 - p)(1 - E[\sigma|B, 0])\bar{s}^0$$

or,

$$\bar{s}^0 = \frac{d - c}{p(1 - E[\sigma|B, 0])} > \bar{s}^1.$$

This value of  $\bar{s}^0$  is unique because when  $\bar{s}^0 = 0$  then  $E[\sigma|B, 0] = 1$ , and  $E[\sigma|B, 0]$  is an decreasing function of the cutoff type  $\bar{s}^0$  with  $\lim_{\bar{s}^0 \rightarrow \infty} E[\sigma|B, 0] = 0$ . ■

Proposition (2) shows that with no exposure there is a smaller measure of shame averse types that will choose to cooperate compared with full exposure. It is still the most shame averse types that will cooperate, but there is a positive measure of shame averse types in the interval  $[\bar{s}^1, \bar{s}^0]$  that would have cooperated with full exposure, but will not cooperate with no exposure.

Two issues are worth noting. First, if some noise is present after defection as well as after cooperation then the results would generalize. Namely, imagine that after a choice of  $D$  the outcome can be good ( $c/p$  for player 1) with some probability  $q$ , and bad (0 for player 1) with probability  $1 - q$ . As long as  $q < p$ , the results described above will carry over because Bayes updating will imply that  $0 < E[\sigma|B, 0] < E[\sigma|G, 0] < 1$ , and the inequality  $E[\sigma|B, 0] < E[\sigma|G, 0]$  would drive the result.

Second, as mentioned at the end of Section 2, one can introduce an exposure technology  $\tau \in [0, 1]$  that is continuous, and does not have the extreme comparison of full versus no exposure. (With probability  $\tau$  there is full exposure after the payoffs are determined, and with probability  $(1 - \tau)$  there is no exposure.) The results derived above generalize to the continuous case because of the type-monotonic behavior, and the structure of Bayes updating in this game.

### 3 Empirical Implications

I now spell out the empirical implications that follow from the immediate application of the analysis in the previous section.

#### 3.1 The Power of Shame

Recall from Proposition (2) above that  $\bar{s}^0 > \bar{s}^1$ , implying that more exposure increases the measure of types who cooperate, and hence, it immediately follows that,

<sup>14</sup>The equality is actually  $d - (1 - E[\sigma|B, 0])\bar{s}^0 = c - p(1 - E[\sigma|G, 0])\bar{s}^0 - (1 - p)(1 - E[\sigma|B, 0])\bar{s}^0$ , but  $E[\sigma|G, 0] = 1$ .

**Hypothesis 1: The Power of Shame.** When moving from a game with no exposure to one with full exposure, the likelihood of cooperation by players 2 increases.

I coin this hypothesis the “power of shame” since it identifies a intervention that manipulates the *direct incentives* from the extrinsic motivator I call shame that has been demonstrated in the model. Notice that if people are motivated by other social preferences that are not dependent on the beliefs of others, e.g., altruism, fairness, reciprocity and self identity, which are intrinsic-based “guilt” preferences, then changes in exposure *should not* have an effect on the outcomes observed. This is the first prediction that differentiates the extrinsic nature of shame from other social preferences.

### 3.2 The Rationality of Trust

The “power of shame” hypothesis has implications about the behavior of player 2 in equilibrium: an exposure to shame will increase the pool of players who will choose to cooperate. As a result, equilibrium analysis implies that player 1 should anticipate the “power of shame” with a response of, loosely speaking, more trusting behavior. This offers an empirical test of equilibrium behavior, or rational “best response” behavior in the setting of this simple trust game.

As mentioned in the introduction, this is an important and subtle point. If the power of shame hypothesis is confirmed for actors in the role of player 1, and if players 2 do not act *as if* they perceive the power of shame, then violating equilibrium behavior in this simple setting would shed some doubt on the use of equilibrium concepts even in the most simple games. If, however, players 1 adapt their own behavior in a way that is consistent with rational choice equilibrium analysis, validating the equilibrium approach that game theory subscribes to in such a simple setting offers some credibility to the use of equilibrium analysis in strategic settings.

To make this hypothesis precise, one has to introduce some variation in the behavior of players 1 that, for simplicity, is not in the formal structure of the game described above.<sup>15</sup> For example, player 1 may also be motivated by what player 2 believes about him. In this context, notice that player 2 *always* infers the behavior of player 1, and as a result, may think “poorly” of a player 1 that chooses not to trust. This will cause some types of player 1 (with high “shame aversion”) to trust, and others not to, with a similar cut-off structure to the behavior of the types of player 2.

Then, if one manipulates the shame incentives that player 2 faces using exposure as described above, the equilibrium effect on the types of player 1 should be that more types of player 1 should have an incentive to trust player 2 when there is exposure. The reasoning is that due to the increased likelihood that trust will be reciprocated, some “marginal” types of player 1 that chose not to trust with no exposure will now choose to trust when there is full exposure. This implies the following testable hypothesis:

---

<sup>15</sup>To be loyal to the model above, the fact that player 1 only cares about money implies that either he should trust or not, depending on his belief about the probability that player 2 cooperates. Hence, if empirically implemented, there is no variation across different players 1, since they should share the same beliefs.

**Hypothesis 2: The Rationality of Trust.** When moving from a game with no exposure to one with full exposure, the likelihood of trust by players 1 should increase.

I coin this hypothesis the “rationality of trust” since it identifies an intervention that affects the *indirect equilibrium incentives* to trust by player 1 through the manipulation of player 2’s shame incentives. Thus, if shame is a motivator for player 2 then changes in exposure should not only have an effect on the choice of player 2, but they should also, through rational expectations, have an effect on the choice of player 1. This is the second prediction that differentiates the theory of shame from other social preferences.

### 3.3 The Weakness of Anonymity

Aside from turning exposure “on” and “off,” the theory implies that the strength of exposure will be a function of whether or not the players know the identity of their partners. Consider, for example, a game in which there are two players 1 and two players 2 with two matched pairs. If every player knows with whom he is matched then after the game is played, every player 1 is able to infer something about the behavior—and type—of the player he is matched with.

If, however, players are anonymous and do not know who they are matched with, then players 1 must “spread” their updating across both players 2 and cannot infer as strongly about either because Bayes updating is “diluted” across the two players 2. For example, with full exposure, if some player 1 learns that his paired player 2 chose to defect, but does not learn about the outcome of the other pair, then he knows that there is a defector but he does not know who it is. Hence, anonymity is another form of imperfect monitoring.<sup>16</sup>

**Hypothesis 3: The Weakness of Anonymity.** When moving from a game with anonymity to one without anonymity in pairs, the likelihood of cooperation by players 2 and of trust by players 1 should increase.

Combining hypothesis 3 with hypotheses 1 and 2 offers an interesting “monotonicity” of behavior: No exposure with anonymity offers the weakest incentives to cooperate (and trust), while full exposure without anonymity offers the strongest incentives. By hypotheses 1 and 2, anonymity with full exposure offers stronger incentives than anonymity with no exposure, and similarly, no-anonymity with full exposure offers stronger incentives than no-anonymity with no exposure. Hence, anonymity with full exposure, and no-anonymity with no exposure, both have incentives in between the weakest and strongest incentives, but they themselves cannot be ranked.<sup>17</sup>

---

<sup>16</sup>This can easily be incorporated into the model by having  $N$  players in each role. Without anonymity the analysis is the same as described earlier. With anonymity, every player 2 knows that his actions will impose a weaker effect on the beliefs of players 1 since his actions have a  $\frac{1}{N}$  effect on the updating of beliefs. Hence, the incentives to cooperate are weaker both with full and with no exposure.

<sup>17</sup>If incorporated into the model, the ranking of anonymity with full exposure and no-anonymity with no exposure will depend on the number of pairs playing and on the shape of the distribution  $F(\cdot)$ . It would be a bit heroic and ad hoc to claim that one can predict this ranking.

## 4 Experimental Design

The experimental design follows the analysis described above, and is based very closely on the experiment used in Charness and Dufwenberg (2006). 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 with one treatment each, four with four treatments and six with six treatments. There were 10-30 participants per session.

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 and which side were B-players. Personal identification numbers were assigned to participants (A1, A2,..., B1, B2,...) who were informed that these numbers would be used to determine pairings (one A with one B), to track decisions, and to determine payoffs.

The game played in all treatments had the same structure as in Figure 1 with the parameters  $v = 5$ ,  $c = 10$ ,  $d = 14$  and  $p = \frac{5}{6}$ . In each treatment, A-players received a sheet with two options, “In” (equivalent to “trust” in Figure 1) and “out” (equivalent to “no-trust”). B-players received a sheet with two options, “Roll” (equivalent to “cooperate”) and “Don’t Roll” (equivalent to “defect”).<sup>18</sup>

In each treatment, A-players first record their choice of “In” or “Out” and their sheets were collected. Next, B-players record their choice of “Roll” or “Don’t Roll” (a 6-sided die). B-players made this choice without knowing the actual choice of their matched A-player, but the instructions explained that a B-player’s choice would be irrelevant if his matched A-player chose Out. This “strategy method” guarantees an observation for every B-player. After the decisions of B-players were recorded, a 6-sided die was rolled (by me) for each and every B-player regardless of his or her choice, and recorded on the back of their decision sheet. This was carefully explained to the participants in advance to allow for anonymity of B-players who chose Don’t Roll. The actual resolution of the die was relevant if and 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 after a Roll choice (hence,  $p = \frac{5}{6}$ ).

Up to six treatments were run in any given session, where in each session a randomly chosen sequence of treatments was implemented. The six possible treatments were:

1. **Anonymous/No-Exposure (AN):** In this treatment participants did not know who they were matched with, nor did A-players learn why they received a payoff of zero if they did.
2. **Anonymous/Exposure (AE):** In this treatment participants did not know who they were matched with, but A-players did learn why they received a payoff of zero if they did.
3. **Matched/No-Exposure (MN):** In this treatment participants learned who they were

---

<sup>18</sup>It is customary not to use “loaded” words such as “trust” or “defect” for the actual experiment, and these are the exact terms used in Charness and Dufwenberg (2006).

matched with before they played (pairs of ID numbers were announced before the decisions and each pair acknowledged each other by standing). A-players did not learn why they received a payoff of zero if they did.

4. **Matched/Exposure (ME):** In this treatment participants learned who they were matched with before they played (as in MN) and A-players did learn why they received a payoff of zero if they did.
5. **Anonymous/Public (AP):** In this treatment participants did not know who they were matched with (as in AN and AE). After the sheets were collected for each B-player, I publicly announced what that B-player had chosen. (This is like “super” exposure to players, without knowing who they were matched with.)
6. **Matched/Public (MP):** In this treatment participants learned who they were matched with before they played (as in MN and ME). After the sheets were collected for each B-player, I publicly announced what that B-player had chosen. (again, “super” exposure but with known matches.)

It is best to start by comparing MN and ME: these two treatments exactly imitate the game described in the theoretical analysis with both no exposure and full exposure, and most closely match the experimental setting that tests hypotheses 1 and 2. The reason being that each matched pair is aware of each other, and hence each is an independent game of two players that interact and form beliefs about each other. The theory predicts that there will be more cooperation (the power of shame) and more trust (the rationality of trust) in treatment ME.

The same logic applies for AN versus AE, but now there is the “free rider” problem implied by hypothesis 3. Hence, the amount of trust and cooperation should be lower in the AN treatment as compared with the MN treatment, and similarly, the amount of trust and cooperation should be lower in the AE treatment as compared with the ME treatment. However, as described in section 3.3, it is not possible to infer the ranking of cooperation and trust between the AE and MN treatments.<sup>19</sup>

The last two treatments are an attempt to disentangle two effects: a possible “matching-effect” of having an identity of a matched partner revealed, and a “public shame effect”, of having the action of B-players announced to all the participants in the room. If shame is the primary motivator, then it is the announcement of behavior and not the identity-matching that should matter, implying that the results in AP and MP should be close to identical. Intuitively, one might expect the behavior in the public announcement treatments to be somewhat more cooperative than ME since shame is exposed to all the players in the room. However, if the beliefs of players are that once the “word is out” then it percolates throughout society, then the results of AP and MP should be similar to ME.

---

<sup>19</sup>These four treatments offer the ability to show that it is not just the change in anonymity that is playing a role. Even when anonymous, changes in the ability of others to infer behavior can play a role. For variations in anonymity alone see Soetevent (2005) and the references therein.

It is worth noting that if shame provides incentives, then the mere presence of an experimenter, or the belief of a subject that someone is observing his or her behavior, will act as a motivator for pro-social behavior. The experimental design, however, keeps the presence of the experimenter constant across treatments, thus only manipulating the exposure to other subjects.

## 5 Experimental Results

Two pilot sessions were run with one treatment each, AN and AE. This was done to ensure that the “mechanics” of the experiment were manageable and easy to implement. The next four sessions (totalling 34 pairs) had the first four treatments (AN, AE, MN and ME) in each session. Realizing the potential effects of non-anonymous matching, I added the last two treatments (AP and MP) to each of the last six sessions, so those included six treatments each (totalling 53 pairs).<sup>20</sup> Smaller sessions (with 5-6 pairs) lasted for about 30-40 minutes, and large sessions (with 10-15 pairs) lasted for close to one hour. The average payout to the participants was about \$14 which included the \$7 show-up fee.

### 5.1 The Power of Shame

Focusing on the “trustee,” or B-player, the experimental results corroborate the predictions of the theory. Table 1 presents the means and exact confidence intervals based on the binomial distribution for each of the 6 treatments. Table 1 offers two cuts of the data. On the left side are the results of the complete panel from all the sessions and all the treatments.<sup>21</sup> These results are also shown in Figure ???. It is easy to see that there is more cooperation in ME as compared to MN, as well as in AE as compared to AN. AE and MN, however, cannot be ranked, also consistent with the theoretical predictions described earlier. Interestingly, the behavior of B-players in the two public treatments AP and MP cannot be statistically distinguished, and they are also the same as in ME. This tentatively 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.

The right side of Table 1 replicates the results but only includes the first two treatments from each session. The reported results have larger confidence intervals since there are significantly fewer observations. Still, the general pattern appears consistent with the theoretical results, and this is presented to demonstrate that the results are not driven by the later sessions.<sup>22</sup>

---

<sup>20</sup>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.

<sup>21</sup>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 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 deviation of .048. Thus, adding the pilot results do not alter the conclusions, and they strengthen the confidence intervals for the AN and AE treatments.

<sup>22</sup>This can be considered as a robustness test since some scholars are concerned with the subjects “figuring out”

To consider the data from a view point of a discrete choice problem faced by the B-players between choosing ROLL or DON'T ROLL, Table 2 shows the results from a linear regression for the behavior of the B-players who played multiple treatments (i.e., excluding the pilot sessions) where the dependent variable  $y_{it}$  is the choice of player  $i$  in treatment  $t$ . With mutually exclusive treatments (which makes 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. Standard errors are clustered at the individual level and are robust Huber-White standard errors to account for the heteroskedastic variance of the  $y_i$  values.<sup>23</sup>

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 (1) and (2) yield almost identical results to the binomial analysis shown in Table 1 above. 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. The dummies are jointly significant (or, more precisely, we can reject the hypothesis, given the  $p$ -value of almost zero, that the dummies are jointly equal to zero) in the analyses in columns (1) through (3).

The more interesting  $F$ -tests that relate to the theoretical predictions are whether the coefficients on some dummy variables are statistically significantly different from those of other dummy variables. Indeed, as one might expect from a glance at Figure ??, the hypothesis that the behavior in the MN and ME treatments is the same is rejected with a  $p$ -value to close to 0 to be reported by STATA. Similarly, the data rejects that the behavior in the AN and AE is the same. However, one cannot reject the hypothesis that the behavior is similar in the ME, AP and MP treatments at conventional levels.

## 5.2 The Rationality of Trust

Turning to the “trustor,” or A-player, the experimental results again corroborate the predictions of the theory. Table 3 presents the means and exact confidence intervals for all the treatments on the right side, and for the first two on the left side. As before, these are the exact means and confidence intervals based in the binomial distribution.<sup>24</sup> The results of Table 3 are also shown in Figure 5. Table 4 shows the results from a linear regression for the behavior of the A-players who played multiple treatments in a similar way to Table 2 for the B-players.

---

what the experiment is about as it unfolds when they play through multiple treatments in the same session. I have not considered only the first session since there is too little data to offer meaningful results.

<sup>23</sup>More precisely,  $Y_i \sim \text{Bernoulli}(p = \delta_j T_j, j = 1, \dots, 6)$ . so that,

$E[(Y_i | T_j, j = 1 \dots 6) = p = \delta_j T_j, j = 1, \dots, 6)$ , where  $\text{VAR}(Y_i | T_j, j = 1 \dots 6) = p(1 - p)$ , and the errors are therefore heteroskedastic.

<sup>24</sup>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.

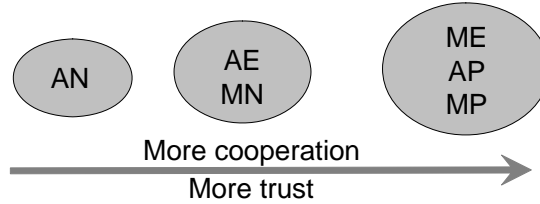


Figure 2: Individual Monotonic behavior

Similar to the B-players, here too the relevant  $F$ -tests are whether the coefficients on some dummy variables are statistically significantly different from those of other dummy variables. Once again, the stark results suggested by Figure 5 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. Similarly, the data rejects that the behavior in the AN and AE is the same. However, one cannot reject the hypothesis that the behavior is similar in the ME, AP and MP treatments at conventional levels.

### 5.3 Monotonic Individual Behavior

The results above show that aggregate behavior is consistent with the theoretical predictions. The model, however, offers a more stringent test of the theory in that considering any given player, behavior should be *monotonic*: the more extrinsic pressure there is, the higher should be the likelihood that any given B-player will choose cooperate, and that any given A-player will choose trust. This, goes a step beyond comparing means as in the tables above since it predicts monotonicity at the individual level, which can be tested using the fact that there are at least 4 observations for each individual. Figure 2 offers a simple description of the individual level monotonicity vis-à-vis the experimental treatments.

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 informational content in either AE, MN or ME as compared with AN, so if the power of shame is strong enough to induce cooperation in the AN treatment, it must be so for the other three 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 an AN treatment to an AE treatment, or when moving from a MN treatment to a ME treatment. However, things are more subtle for some of the other transitions. Namely, an A-player may learn something in the middle of the experiment when the treatment is AE or ME. That is, if an A-player trusted in the AE session and learned that he was defected against, then his posterior of 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: if player A trusted in the AE treatment and observed a defection, Bayes updating can imply that he will not trust a B-player in a later ME treatment. However, if a player A trusted in the MN (or AN) treatment, he must trust in the ME (or AE) treatment since he learns nothing in the MN (or AN) treatment.

Turning to the actual play, B-players exhibit a rather high level of monotonicity: only 12 out of 85 individuals have a violation of 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. It turns out that a positive and significant correlation occurs between A players whose trust was abused in a treatment with exposure, and the same players exhibiting a non-monotonicity after learning about their partner's defection. What's more important is that out 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. Hence, only four of the 86 A-players exhibit an irrational non-monotonicity.

An obvious question is whether a taste for reciprocity can explain this non-monotonic behavior (e.g., Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006)). That is, after being hurt, the A-player wishes to "punish" the B-player even though the likelihood of cooperation could go up. It is possible, but arguably less convincing because of the random pairings. For reciprocity to be a motive, the players should believe that the probability of re-matching is high enough, which is questionable with groups consisting of six to twelve pairs.

## 6 Discussion

### 6.1 Shame and Reputational Concerns

The premise of this paper is that individuals have extrinsic preferences over the beliefs of others, or more broadly, that they care about how they are perceived by others. Those familiar with the literature on repeated games will quickly note that when players' current actions affect the way others will reciprocate, then in equilibrium, even players who only care about material payoffs will effectively have *indirect* preferences over the beliefs of others. This is particularly true if one considers repeated games of incomplete information with imperfect monitoring (see, e.g., Mailath and Samuelson, 2006), the latter being the case when a player's digression cannot easily be detected by his counterparts.

In fact, one might reason that the players in my experiment may have treated this game as if it were really a repeated game with the other people in the room. Since exposure is a way to eliminate imperfect monitoring, this would cause reputational concerns to be more pronounced, and behavior to be more cooperative. Such an approach would offer equilibrium predictions that can rationalize many of the experimental results without resorting to direct preferences over the beliefs of others. (Note, though, that such repeated games would suffer from the well known

problem of multiple equilibria.)

I would argue, however, that three reasons make this alternative explanation less convincing than the power of shame argument of direct extrinsic preferences. First, the pool of applicants is drawn from several thousands of students and staff members at UC Berkeley, making it rather unlikely that the players can rationally perceive this to be part of a broader repeated game.<sup>25</sup> For the repeated game explanation to be convincing, either the probabilities of future interaction need to be high, or the marginal losses from the potential future interactions need to be high. Both seem unlikely given the pool of applicants.

Second, and more striking, a repeated game prediction would not easily explain the similarity of the level of cooperation between the ME, AP and MP treatments. If one takes seriously the repeated game approach of “what goes around comes around,” then exposure to more people should act as a significantly strong motivator, implying more cooperation in the public exposure treatments (AP and MP) and less in the private exposure treatment (ME).<sup>26</sup>

The relation between shame and reputation may actually be meaningful. It may be that preferences for shame are a simple reduced form mechanism that works well for the more involved repeated games that we play most of the time. Arguably, if most human contact takes the form of repeated interactions, then our preferences may have evolved to best fit these kind of games. This idea is not new, and dates back at least to the works of Frank (1987, 1988) who argues that human emotions are shaped by natural selection to increase one’s chances of survival. Behavior that appears to be irrational in a one-shot setting will promote mutually beneficial trade and, thus, increase a player’s long-run welfare.<sup>27</sup> A careful analysis of whether shame-based preference would endogenously evolve in a general class of trust games is beyond the scope of this study.

## 6.2 Beyond Trust Games

The presence of pro-social behavior has been documented in many classes of games. I will briefly discuss the application of extrinsically motivated preferences to a few well known classes of games, and discuss how the results are consistent with these preferences.

### 6.2.1 Dictator Games

Dictator games are effectively simple decision problems where a player is offered some money, say \$10, he has to decide how much to give away to another player. A variety of experiments along these lines have been executed ad nauseam, with similar results: people give away money,

---

<sup>25</sup>Indeed, casual observation of the students as they waited to be seated suggests that very few of them even knew each other before coming to the experiment.

<sup>26</sup>A repeated game approach can explain this by assuming that it is enough for one person to learn what you did, after which the word immediately spreads to all future potential partners. This seems like quite a stretch. That said, List (2007) offers results from a combination of lab and field experiments that support reputational concerns in the field that are weaker than the pro-social behavior in the lab for similar individuals.

<sup>27</sup>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.

even when they are anonymous and play only once. All of the intrinsic “guilt” type preferences mentioned in the introduction are consistent with the findings. An extrinsic “shame” mechanism is also consistent if one thinks of the experimenter being a constant observer of the game. (And, as anticipated, removing anonymity will increase the amount given.)

Recently, several experiments have results that are consistent with the implications of extrinsic preferences. Dana et. al. (2006) find that a third of participants were willing to exit a \$10 dictator game and take \$9, if the latter option ensured that the receiver never knew that a dictator game was to be played. The variety of guilt-based models cannot explain choosing the (\$9, \$0) exit outcome over the dominating \$10 dictator game, since the game includes outcomes of (\$10, \$0) and (\$9, \$1). Shame, however, is consistent with this result.<sup>28</sup> They conclude that “giving often reflects a desire not to violate others’ expectations rather than a concern for others’ welfare per se.” This statement might mean that it is not the other player’s beliefs about me that motivate me, but his beliefs about what he will get that motivates me, and I don’ want to disappoint him (much like the motivation in Charness and Dufwenberg, 2006). Note, however, that the results and interpretation of Dana et. al. are also consistent with shame: giving reflects a concern about how others perceive your type, or your behavior, and exit is a way to buy out from exposure.<sup>29</sup>

A recent paper that shares many ideas with my results is Andreoni and Bernheim (2007). They offer an explanation of the common norm in which a 50-50 division in the dictator game appears to have considerable force. Once again, they show that guilt-based preferences cannot account for this regularity, and instead argue for preferences in which people like to be perceived as fair, which shares a similar flavor to the extrinsic shame preferences introduced here. Furthermore, they too manipulate the knowledge of the recipients in a rather subtle way that corroborates the idea that so-called audience effects are consistent with a signalling, or reputational concern.<sup>30</sup>

Notice that the role of B-players in my experiment is similar to that of a dictator with two actions (though it is not “zero-sum”.) The experimental setting of this paper goes beyond establishing the presence of shame-based preferences for the B-players, and demonstrates the rational response of A-players who seem, through their behavior, to acknowledge that their

---

<sup>28</sup>Shame is also consistent with another finding in Dana et. al. (2006). They use a private game in which the receiver never knew about the game or from where any money was received. Almost no dictators exited from the private game, indicating that receivers’ beliefs are the key factor in the decision to exit.

<sup>29</sup>Lazear et. al. (2006) offer some variations on the Dana et. al. experiment and show that the propensity to choose the dictator game over exit increases as the size of the pot in the dictator game increases, keeping the exit payoff fixed. This is of course consistent with the preferences offered in section 2, where people care about shame and money. Lazear et. al. try to argue for guilt-based preferences of a certain kind, but since they do not vary the informational content like in Dana et. al. (2005), the shame-based results of Dana et. al. shed some doubt on the interpretation offered by Lazear et. al. (2006).

<sup>30</sup>A recent paper by Ellingsen and Johannesson (2007) offers a standard anonymous treatment of a dictator game and compares it to one where the recipient can send an open ended message following the division. They show that when messages are allowed, more giving occurs, and relate this finding to shame (I don’t want to get bad feedback) and pride (I want to get good feedback). Since exposure is not manipulated, the results are consistent with messages being part of the outcome space and thus supporting intrinsic preferences over these outcomes (e.g., expressed disappointment or approval).

fellow B-players are motivated by extrinsic preferences such as shame.

### 6.2.2 Voting and Public Goods

Since the description of the “Voting Paradox” by Downs (1957), there have been many attempts to offer a rational-choice explanation for the fact that many people turn out to vote despite the fact that their voting does not really matter, and that voting itself is costly. Common explanations have been similar to the conventional ideas behind pro-social behavior: people care about the “public good” through some sense of responsibility, and hence they show up at the ballots. This implies that if access to voting becomes easier, turnout should clearly not decrease, and most likely would increase.

In an interesting recent paper, Funk (2007) investigates a unique natural experiment, which generated facts that fly in the face of this conventional wisdom. She collected data from Swiss elections before and after mail-in voting was introduced. Clearly, the possibility of mail-in voting must have reduced voting costs substantially. However, not only it did overall turnout not increase on average, voter turnout decreased in the smaller communities, and somewhat increased in the larger ones. She also shows that in communities where poll station operating windows were smaller (interpreted as an increase in the cost of voting due to the lack of flexible hours), turnout decreased more.

Funk (2007) resorts to a signalling model of voters who wish to be seen as public minded. The results, however, are very much consistent with extrinsic shame-based preferences where players do not want to be perceived as shirkers (don’t care to vote). Namely, in smaller communities or in those with short poll operating windows, it is more likely that someone you know will be at the station at any given time, akin to a higher level of exposure. As such, not voting and offering a lie as to when you were there is more likely to be detected as compared with large communities and with flexible hours. By offering mail-in voting you can credibly say that you sent it in, and this noise will be the perfect cover for the cost minimizing voter. Hence, exposure was reduced.

Generally, it is easy to see how one can manipulate experimental games of public good contributions in a similar way to test whether extrinsic preferences for shame appear to play a role in these games as well.

### 6.2.3 Ultimatum Games

Ultimatum games are more complex than dictator games in that explaining the commonly observed experimental results suggests that some form of reciprocity may be present. Recall that in such games player 1, the proposer, offers to split some pot of money, say \$10, into an amount  $x_1$  for himself and  $x_2 = 10 - x_1$  for player 2. Player 2 then responds by either accepting the offer after which the sums are distributed, or rejecting it after which neither players receives money. It is well known from many experiments that proposers typically leave significant amounts to the responder, and the latter typically rejects offers of that are low (usually less than a third). See

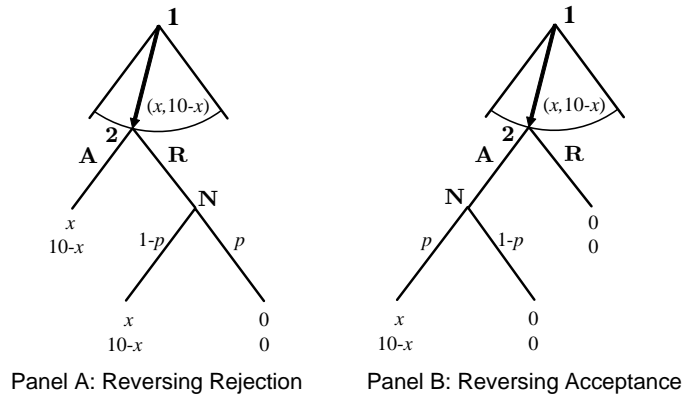


Figure 3: Variations on the Ultimatum Game

Camerer (2003) for a detailed survey.

It is possible that the interpretation of extrinsic preferences for self-presentation as a reduced form “shortcut” for reputational concerns could be consistent with both first players leaving money for responders, and responders rejecting low payments. The challenge is to devise an experiment that would distinguish between reciprocity—players 2 respond in kind to a good offer and punish a bad one—and preferences to appear strong, i.e., a sort of shame not to be perceived as weak.

A similar use of monitoring noise is possible by offering two simple variants of the ultimatum game as shown in Figure 3. In Panel A, after a responder rejects an offer there is a small probability  $p$  that Nature will reverse the rejection and make it appear as if the responder accepted the offer. Now one can manipulate whether or not the proposer observes the choices of the responder and of nature, or whether he only observes the payoffs. If rejecting a low offer is a consequence of some kind of shame-reputational concerns, then when noise about actions is introduced, responders would be more likely to accept low offers since acceptance can be misperceived as nature undoing player 2’s tough behavior. If, however, reciprocity is the concern, then behavior should not change.

The second variant would be to allocate the noise to reverse an acceptance, as shown in Panel B of Figure 3. Here, acceptance is a sign of “weakness” regardless of whether player 1 can observe player 2’s actions or not. Hence, introducing noise (no exposure) should not change the behavior of the responder if self-presentation based preferences are the driving force. The actual execution of these experiments is left for future research.

### 6.3 Implications for Policy and Strategy

The last question raised by this study is to what extent can individuals, organizations and policy-makers design incentive systems that operate in the face of extrinsic shame-based preferences? I will address this question with some anecdotes, and some thoughts on the role of shame in contractual relations.

Legal scholars have discussed the role of emotions in the law (see, e.g., Posner (2001)), and the use of shame in the public policy domain has been applied to crime deterrence for some time. To refer to a recent 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, with the intention that people will indeed be deterred from these crimes. The experiment was never truly executed because of legal battles, but it is just one example of the potential was in which shame can act as a deterrent. Hence, shame as a deterrent deserves a careful set of studies to determine its efficacy.<sup>31</sup>

The strategy of firms and organizations is also a fruitful ground for shame-based incentives to work. One common example is the use of public fund-raising in religious gathering places such as churches. In a recent study, Soetevent (2005) conducted a field experiment in thirty Dutch churches where the means by which offerings were gathered was determined by chance, using either a "closed" collection bag or an open collection baskets. When using baskets, attendants' contributions can be identified by their direct neighbors, and 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.

Turning to business cases, an interesting one dates back to the early nineteenth century regarding a novel incentive mechanism adopted by the Utopian idealist, Robert Owen, 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 and the amount produced, 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 be very much in line with shame-based preferences motivating employees to perform better.

---

<sup>31</sup>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 as a crime deterrent 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.

Motivating employees using peer exposure and social pressure may indeed be a very fruitful way for business organizations to provide incentives. In a recent study, Mas and Moretti (2008) 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. What is interesting, however, is that they find that productive workers induce a productivity increase only in workers that are *in their line-of-vision*. Hence, it seems to be exposure, as analyzed above, that is playing a central role in motivating behavior.

Even more interesting, in my view, is the potential impact that shame-based preferences have on contractual relationships and gains from trade. The selfish utility maximizing model has demonstrated that mutually beneficial economic relationships may suffer from moral hazard and opportunism when contracts cannot be fully specified and enforced. It is precisely this kind of “opportunism with guile,” as Williamson (1975) put it, that is at the heart of studying contractual relations and the role of contractual and organizational design. That said, despite many instances where such hazards exist, the way in which parties seem to do business suggests that problems of moral hazard are not rampant, and the use of external enforcement and litigation is often the exception rather than the rule.

It is possible, though not easily proven, that shame-based preferences play a role in the fact that most transactions are not reneged upon even when contractual specifications are incomplete and external enforcement is fragile. To see this, imagine the case where two agents can engage in some transaction where the outcomes can be incompletely specified in advance. Agency theory suggests that information should be gathered to support the contract only if it can help enforcement since otherwise it is useless. However, if parties can reveal information that causes exposure of poor behavior, then shame based preferences imply an important role for investing in revealing such information even when it cannot be formally used in an externally enforced transaction.<sup>32</sup> 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, once again relating the motivation from shame to the motivation from repeated interactions.<sup>33</sup>

---

<sup>32</sup>Of course, 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. Once again, the analogy between reputational concerns and emotions such as shame is apparent.

<sup>33</sup>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 people discourage malfeasance.” However, embeddedness theory acknowledges that social networks alone will not deter malfeasance. It may be that shame can offer a mechanism through which some of the ideas of embeddedness operate.

## 7 References

- Andreoni, James (1990) "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving," *Economic Journal* **100**:464-77.
- Andreoni, James and B. Douglas Bernheim (2007) "Social Image and the 50-50 Norm," mimeo.
- 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," forthcoming, *Journal of Economic Theory*.
- 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
- Bloom, Martin (2003) "Editorial—Primary Prevention and Education: An Historical Note on Robert Owen," *The Journal of Primary Prevention*, **23(3)**:275-281
- Camerer, Colin *Behavioral Game Theory: Experiments in Strategic Interaction*. 2003, Princeton University Press, Princeton, NJ.
- Charness, Gary and Martin Dufwenberg (2005) "Promises and Partnership," forthcoming *Econometrica*.
- Dana, J., D. M. Cain and R. 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 & Economic Behavior* **47**:268-98
- Ellingsen, Tore and Magnus Johannesson. (2007) "Anticipated Verbal Feedback Induces Altruistic Behavior," *Evolution and Human Behavior*, forthcoming.
- Falk, Armin and Urs Fischbacher (2006) "A Theory of Reciprocity," *Games and Economic Behavior* **54(2)**:293-315
- Fehr, Ernst and Klaus M. Schmidt (1999) "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics*, **114**: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.
- Frank R.H. *Passions Within Reason — The Strategic Role of the Emotions*, 1988, W.W. Norton & Company, New York.
- Funk, Patricia (2007) "Social Incentives and Voter Turnout: Theory and Evidence," mimeo, Universitat Pompeu Fabra.

- Geanakoplos, John, David Pearce and Ennio Stacchetti (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, 1:60–79.
- Granovetter, Mark, (1985) "Economic Action and Social Structure: the Problem of Embeddedness," *American Journal of Sociology*, **91**:481-93
- Heifetz, Aviad, Chris Shannon and Yossi Spiegel (2007) "What to maximize if you must," *Journal of Economic Theory*, **133**:31 – 57
- Kreps, David M. and Robert B. Wilson, "Sequential Equilibrium," *Econometrica*, 1982 **50(4)**:863-894
- Lazear, Edward, Ulrike Malmendier and Roberto Weber (2006) "Sorting in experiments with application to social preferences," mimeo
- Levitt, Steven D. and John A. List (2007) "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, **21(2)**:153-174.
- 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, *Repeated Games and Reputations: Long-Run Relationships*, Oxford University Press, Oxford and New York, 2006.
- Mas, Alexandre and Enrico Moretti (2008) "Peers at Work," *American Economic Review*, forthcoming.
- Niedenthal, Paula M., June Price Tangney, and Igor Gavanski (1994) "'If Only I Weren't' Versus 'If Only I Hadn't': Distinguishing Shame and Guilt in Counterfactual Thinking," *Journal of Personality and Social Psychology* **67(4)**:585-595
- Posner, Richard A., *Frontiers of Legal Theory*, Harvard University Press, Cambridge, MA, 2001.
- Rabin, Matthew (1993) "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, **83(5)**:1281-1302
- Smith, Richard H., J. Matthew Webster, W. Gerrod Parrott and Heidi L. Eyre (2002) "The Role of Public Exposure in Moral and Nonmoral Shame and Guilt," *Journal of Personality and Social Psychology*, **83(1)**:138–159
- Soetevent, Adriaan R. (2005) "Anonymity in giving in a natural context – A field experiment in 30 churches," *Journal of Public Economics* **89**: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.
- Zemack-Rugar, Yael, James R. Bettman and Gavan J. Fitzsimons (2007) "Effects of Non-consciously Priming Emotion Concepts on Behavior," *Journal of Personality and Social Psychology*, in press.

## Appendix: 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 1: 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 2: Behavior of B players

<i>Linear Probability Model, Dependent variable: Player B chose Roll (Cooperated)</i>			
	(1)	(2)	(3)
AN	0.200 (0.044)	-	0.196 (0.057)
AE	0.377 (0.053)	-	0.411 (0.070)
MN	0.400 (0.054)	-	0.431 (0.071)
ME	0.753 (0.047)	-	0.804 (0.057)
AP	-	0.725 (0.063)	0.725 (0.064)
MP	-	0.784 (0.058)	0.784 (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 3: 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 4: Behavior of A-players

<i>Linear Probability Model, Dependent variable:</i>			
<i>Player A chose In (Trust)</i>			
	(1)	(2)	(3)
AN	0.302 (0.05)	-	0.340 (0.066)
AE	0.558 (0.054)	-	0.585 (0.069)
MN	0.547 (0.054)	-	0.604 (0.068)
ME	0.744 (0.048)	-	0.660 (0.066)
AP	-	0.660 (0.066)	0.660 (0.066)
MP	-	0.736 (0.061)	0.736 (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)

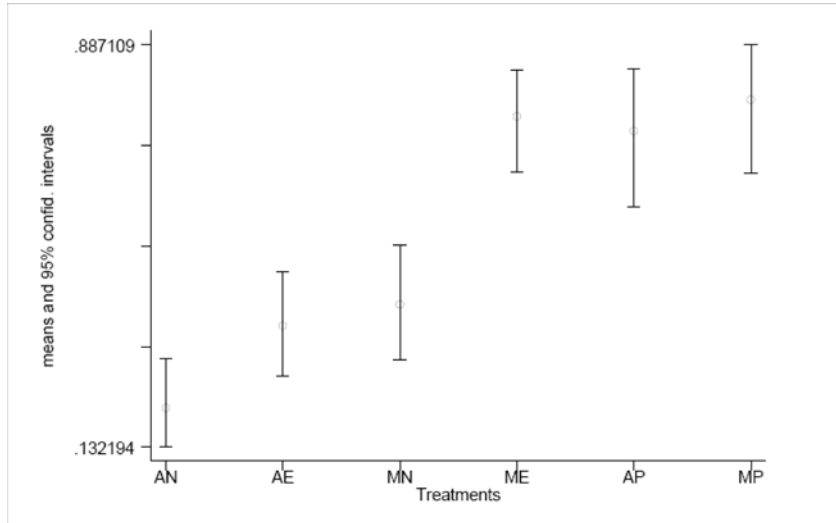


Figure 4: Percentage of B-Players Who Choose Cooperate in the Six Treatments

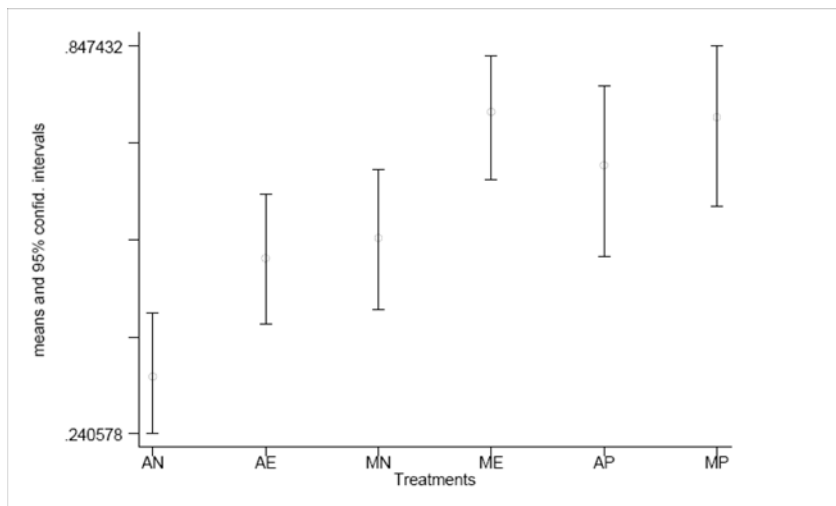


Figure 5: The Percentage of A Players Who Chose Trust in the Six Treatments