

# The Power of Shame and the Rationality of Trust

Steve Tadelis\*  
UC Berkeley  
Haas School of Business

**PRELIMINARY AND INCOMPLETE!**

November 6, 2006

## 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 “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 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 may explain observed behavior in some previously studied experiments, and imply new testable predictions. These are tested, and confirmed in a series of new laboratory experiments. Establishing that shame can be manipulated has implications for policy and strategy. *JEL* classifications

---

\*This work was inspired during the Stanford Institute for Theoretical Economic workshop in August 2004. I am grateful to Gary Charness for sharing his experimental design and his experience. I also thank Yossi Feinberg, Shachar Kariv, John Morgan and Muriel Niederle 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

To Be Completed.

- At the market level, the selfish rational choice model works surprisingly well given its simple structure and extreme assumptions. However, when human subjects are put to tests of individual decision making in simple strategic situations, their behavior departs from theoretical predictions in one systematic way. Namely, individuals seem to exhibit some form of “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.)
- The experimental results suggest that individuals must have some intrinsic incentives, or preferences, for pro-social, often welfare enhancing behavior. Many mechanisms have been proposed as such intrinsic motivators: altruism (direct and Indirect (Andreoni, 1990)); inequity Aversion (Fehr and Schmidt (1999)); preferences for fairness (Rabin (1993)), preferences for reciprocity (Falk and Fischbacher (2006)), and even more unusual motivators such as self-identity (Akerlof and Kranton (2000)).
- These motivators are primarily intrinsic in that changing the informational environment should not alter the preferences of players. As such, I loosely combine these pro-social preferences under the title of “guilt”: being selfish causes a player to experience some internal loss of utility.
- Shame can be another *distinct* motivator. As noted by a teen self-help web site, “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.”<sup>1</sup>
- This quote summarizes quite well the accepted distinction between these emotions. As Tangney (1995) writes, “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.”
- As such, preferences that incorporate shame must include a preferences over the beliefs of others, not just some intrinsic preferences over physical outcomes. For example, if players expect selfish people to act selfishly,

---

<sup>1</sup><http://www.teenhealthcentre.com/teens/mentalhealth/depression/dep03.htm>

and if one does not want to be perceived as selfish, then one ought to act in ways that decrease the belief people have over one being selfish.

- To model these kind of preferences in a game, and to apply notions of equilibrium behavior, I resort to the well established structure of ex post beliefs being formed given a prior distribution of “types,” and given beliefs over behavior and the environment. I endow some players with preferences over the beliefs of others, as in Geanakoplos, Pearce and Stacchetti (1989), and adapt a rather common notion of sequential equilibrium.
- The game I analyze and explore is a simpler variant of the well know “Trust Game,” (Berg, Dickhaut and McCabe, 1995). Player 1 can trust player 2 or exit with a safe outside option. If trusted, player 2 can return in kind by cooperating, but at a sacrifice of an additional monetary payoff that he receives if he defects. A selfish player 2 will therefore never cooperate, and anticipating this player 1 should never trust.
- The game considered has an element of imperfect monitoring as introduced by Charness and Dufwenberg (2006). Imperfect monitoring implies that following trust, player 1 cannot always be certain whether a low payoff is the result of player 2 defecting, or the result of bad luck. (**examples:** principal hires agent for a fixed wage in a hidden action setting; trader relies on carrier to deliver his goods; constituents elect politician to increase their welfare; etc.)
- Player 2 can be of two types: selfish, or thoughtful. A thoughtful type of player 2 cares about the ex post beliefs of player 1, in particular, a thoughtful player 2 would prefer that player 1 not think he is selfish. In equilibrium, given the prior beliefs of player 1 over the types and strategies of player 2, player 1 will form a posterior belief over the type of player 2, and this belief is what player 2 cares about.
- By manipulating the ability of player 1 to decipher whether a bad outcome is a result of bad luck or bad behavior of player 2, it is possible to manipulate the ex post beliefs of player 1 without changing other parameters of the game. As such, if player 2 cares about the beliefs of player 1, the theory suggests a testable implication on the behavior of player 2. Thus, shame will be a motivator for cooperative behavior by player 2, if player 1 can more easily establish whether bad outcomes correspond to bad behavior.
- Furthermore, a rational player 1 will anticipate this change in incentives for player 2, and if the informational environment implies that a thoughtful player 2 is more likely cooperate, then player 1 should be more inclined to trust player 2. Thus, trust can be explained as a rational response to the incentives provided by shame.
- The theory is then applied to a series of laboratory experiments that strongly support the hypotheses that are derived from the theoretical

analysis, implying that aside from the intrinsic motivations explored in previous studies, the motivation implied by shame is strong and robust.

- **Other applications:** dictator games; ultimatum games; tipping; public good contributions
- **Implications:** Firm strategy (observables); Public policy (e.g., “Johns” on billboards); Non-for-profits (churches, synagogues: observable donations);

## 2 A Noisy Trust Game

### 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), and 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 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.

In this game, the payoff from player 1 choosing  $N$  is  $v > 0$  to both players, 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 payment 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>2</sup>

### 2.2 Incomplete Information and Preferences over Shame

To add a dimension of pro-social behavior, I add a simple layer of incomplete information. Player 1 is assumed to be selfish in the sense that only pecuniary payoffs matter to him. Player 2, in contrast, can either be selfish, or he can

---

<sup>2</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 to the standard principal-agent model.

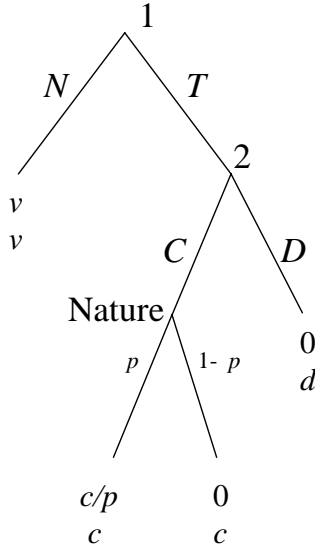


Figure 1: A Simple Trust Game

have some form of pro-social tendencies. As argued in the introduction, one can think of either guilt or shame as being separate motivators, and as such a distinction ought to be made. I formalize the difference between guilt and shame as follows: *guilt* is associated with the intrinsic cost of “cheating” player 1 after he chose to trust player 2. *Shame*, in contrast, is associated with the intrinsic cost of having player 1 *believe* that player 2’s behavior was inadequate.

Formally, a player is “Bad” with probability  $\beta$ , and “Good” with probability  $1 - \beta$ . I abuse the standard terminology by using “bad” for what we usually refer to as selfish, while “good” is an individual who potentially suffers from feelings of guilt and shame. This asymmetric introduction of types for player 2 and not for player 1 will be enough to generate some testable hypotheses.<sup>3</sup>

As an intrinsic motivator, guilt would be a hard dimension to manipulate externally.<sup>4</sup> Therefore, we introduce a type of player that is motivated solely by shame. The cost of shame is modelled as a positive cost which is proportional to

<sup>3</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 choose “trust” in a way that may cause termination of the game, so that 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>4</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 XXX.) 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.

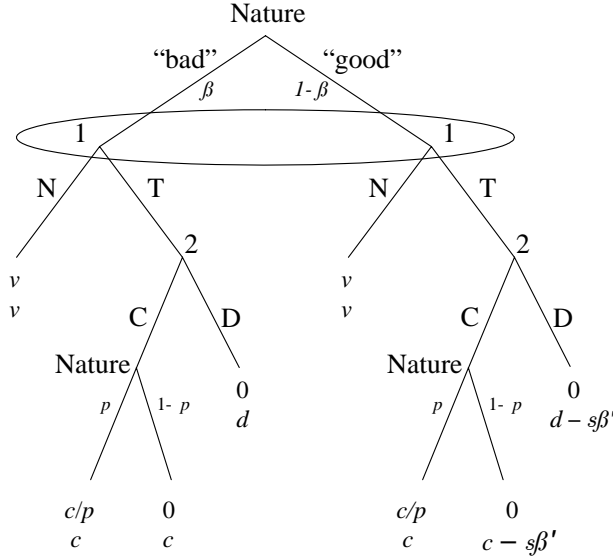


Figure 2: A Trust Game with Incomplete Information about Shame

what player 2 believes is the ex post belief of player 1 about the type of player.<sup>5</sup> Namely, if player 2 believes that player 1’s ex post belief about 2 being bad is given by  $\beta' \in [0, 1]$ , then the cost of shame is  $\beta's$  where  $s > 0$  is the intensity of shame. In a richer model with a more continuous variation of shame sensitivity, different people can have different levels of  $s$ , where  $s = 0$  would coincide with pure selfish preferences.

The full game is depicted in Figure 2. Notice that there is a departure from the standard definition of a game since 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 type of player 2 directly effects the payoffs of player 2, which is the way in which the model captures shame. This is related to the small literature on “Psychological Games,” pioneered by Geanakoplos, Pearce and Stacchetti (1989).<sup>6</sup>

### 2.3 Exposure and Inference

To complete the structure of the game we endow player 1 with an exogenous “exposure technology.” This basically refers to player 1’s ability to decipher the

<sup>5</sup>The cost of guilt can be given by some  $g > 0$  that is incurred when a good player chooses to defect. As mentioned above, I cannot reliably manipulate guilt, and as such, will ignore this potential motivator.

<sup>6</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, which is different from the notion of “shame” I introduce here.

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 “success” ( $S$ ) as the outcome in which player 1 gets a payoff of  $C/p$ , which is a consequence of player 1 choosing to trust, 2 choosing to cooperate, and nature being favorable. I refer to “failure” ( $F$ ) 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 assume that with probability  $\pi \in [0, 1]$  exposure occurs ( $e = 1$ ) and player 1 will observe the action of player 2. That is, if player 1 chooses to trust player 2, and the outcome is a failure, then with probability  $\pi$  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  $C$ . With probability  $(1 - \pi)$ , however, exposure does not occur ( $e = 0$ ) and player 1 learns nothing about the reason for failure.

To foreshadow the experimental design, it is useful to emphasize two extreme forms of detection technology, namely  $\pi = 0$  and  $\pi = 1$ . When  $\pi = 0$  exposure never happens and player 1 never learns the source of a payoff of zero, whereas when  $\pi = 1$  exposure is certain to happen. Clearly, if player 2 is exposed then he cannot “hide” behind the noise of bad luck as a source of low payoffs to player 1. As a result, a “good” player 2 will be more easily deterred from choosing  $D$ . This simple intuition is formalized and explored more carefully in the next section.

### 3 Equilibrium

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. Solving Backward, we know that a “bad” player 2 will always choose to defect. 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>7</sup>

Clearly, a “bad” player 2 will never choose to cooperate. Let  $\sigma \in [0, 1]$  be the probability that a good player 2 chooses to cooperate. If player 1 correctly anticipates  $\sigma$ , and if he trusts player 2, then following trust, his ex post conditional beliefs depend on success ( $S$ ) or failure ( $F$ ), as well as on whether he learns the action of player 2 through the exposure technology. Let  $\beta'(F, e = 0) \in [0, 1]$  be the posterior of player 1 about player 2 being a bad type following a failure, and

---

<sup>7</sup>Higher order beliefs will not play a role in the game analyzed.

no exposure of the reason for failure. It follows from Bayes' rule that,

$$\begin{aligned}\beta'(F, e = 0) &\equiv \Pr\{\text{bad}|F, e = 0\} \\ &= \frac{\beta}{\beta + (1 - \beta)[(1 - \sigma) + \sigma(1 - p)]} = \frac{\beta}{1 - p\sigma(1 - \beta)} \geq \beta.\end{aligned}$$

The expression for  $\beta'(F, e = 0)$  is easy to interpret: the beliefs of player 1 following failed trust are worse if the ex ante likelihood of bad types is larger ( $\beta$  is greater), if the good player 2 is more likely to cooperate ( $\sigma$  is greater), or if the Nature's noise is reduced ( $p$  is greater).

If, instead, exposure happens, then updating depends on what player 1 learns. If player 1 learns that player 2 chose  $D$ , then the Bayesian posterior is given by,

$$\begin{aligned}\beta'(F, e = 1, D) &\equiv \Pr\{\text{bad}|F, e = 1, D\} \\ &= \frac{\beta}{\beta + (1 - \beta)(1 - \sigma)} = \frac{\beta}{1 - \sigma(1 - \beta)} \geq \beta'(F, e = 0).\end{aligned}$$

Finally, if player 2 observes a success, or if he observes a failure but through exposure learns that player 2 chose the cooperate, then he knows that he is facing a good type, that is,  $\beta'(S) = \beta'(F, e = 1, C) = 0$ .<sup>8</sup>

I impose the following belief restriction:

**R1: Ex Post Beliefs.** The ex post beliefs of player 1 must satisfy Bayes rule as described above.

This restriction imposes beliefs for player 1 at the end of the game that are consistent with Bayes rule. The restriction is satisfied in a sequential equilibrium except for the case where player 1 chooses not to trust player 2, or if  $\sigma = 0$ . In either of these two cases, reaching a success is a zero probability event and Bayes rule does not apply. It is for these two cases that R1 imposes a restriction.

**Remark 1** *A variety of refinements would imply the belief restriction R1.*

### 3.1 No Trust/Defection Equilibrium

I first consider the possibility of an equilibrium with no trust that is supported by the correct belief that any type of player 2 would defect following trust. Given R1, if player 1 believes that  $\sigma = 0$  then a good player's utility following defection is

$$\begin{aligned}u_2^G(D, \beta') &= d - s(\pi\beta'(F, e = 1, D) + (1 - \pi)\beta'(F, e = 0)) \\ &= d - s\beta\end{aligned}$$

---

<sup>8</sup>If defection has a probability of success that is positive, then as long as this probability is lower than the probability of success from cooperation, then it must be the case that  $\beta'(S) = \beta'(F, e = 1, C) < \beta$  and  $\beta'(F, e = 1, D) > \beta'(F, e = 0) > \beta$ .

where the second equality follows from R1 and from  $\sigma = 0$ .

If instead a good player 2 would choose to cooperate (and depart from the proposed equilibrium), then his utility following cooperation is

$$\begin{aligned} u_2^G(C, \beta') &= c - sp\beta'(S) - s(1-p)(\pi\beta'(F, e = 1, C) + (1-\pi)\beta'(F, e = 0)) \\ &= c - s(1-p)(1-\pi)\beta \end{aligned}$$

where the second equality follows from R1 together with  $\sigma = 0$ .<sup>9</sup> Therefore, defection is a best response if

$$d - s\beta > c - s(1-p)(1-\pi)\beta,$$

or,

$$d > c + s\beta(\pi + p - \pi p) \tag{1}$$

This is, of course, straightforward. If shame was not a motivator, then all we need for defection to be a best response is  $d > c$ . However, since the good player is motivated by shame, there has to be a “cost of shame” premium over the pure monetary amounts. In particular,  $d$  has to exceed  $c$  by  $s\beta(\pi + p - \pi p) > 0$  for Defect to be part of an equilibrium.

Clearly, a higher sensitivity to shame (a higher  $s$ ) will cause this premium to increase. It is interesting to note two other intuitive comparative statics about this premium. First, this premium increases in  $\pi$ , the likelihood of exposure. This intuitively follows because more exposure carries more shame from defection. Second, this premium increases in  $p$ , the likelihood that nature does not cause bad luck. This intuitively follows because if nature is more “reliable”, then it appears that player 2 would bear more responsibility for failed outcomes.

I continue the analysis with the parametric assumptions that guarantee some trust in equilibrium. Namely, that (1) is violated:

**A1:**  $d < c + s\beta(\pi + p - \pi p)$

The following claim is therefore immediate:

**Claim 2** *“No trust” followed by “Defect” of all types cannot be an equilibrium if and only if A1 is satisfied.*

### 3.2 Trust/Cooperation Equilibrium

Continuing with A1, player 1’s utility from trusting player 2 is obviously increasing in  $\sigma$ , the probability that a good player 2 will choose to cooperate. Therefore, player 1’s willingness to trust is highest when he believes ex ante that  $\sigma = 1$ , in which case he would be willing to trust player 2 if  $\beta$  is not too high (the likelihood of a bad type is not too high). In particular, if  $\sigma = 1$  then player 1’s strict best response is to trust if and only if

$$\beta \times 0 + (1 - \beta)c > v$$

---

<sup>9</sup>Recall that  $\sigma = 0$  is used to determine the beliefs of player 1, despite the fact that player 2 departed from the proposed equilibrium.

or  $\beta < \frac{1}{2}$ . This of course is obvious: if the likelihood of a bad player 2 is too high, then even impeccable behavior by a good player 2 will not induce player 1 to trust player 2. The following claim is immediate:

**Claim 3** *Trust by player 1 can be part of an equilibrium if and only if A2 is satisfied.*

Thus, to allow for some trust in equilibrium I must continue with the following assumption,

**A2:**  $\beta < 1 - \frac{v}{c}$

If player 1's ex ante belief is that  $\sigma = 1$ , then his ex post belief must satisfy,

$$\begin{aligned}\beta'(F, e = 0) &= \frac{\beta}{1 - p(1 - \beta)} > \beta, \\ \beta'(F, e = 1, C) &= \beta'(S) = 0, \\ \beta'(F, e = 1, D) &= 1\end{aligned}$$

Hence, player 2's utility from cooperating is,

$$\begin{aligned}u_2^G(C, \beta') &= c - sp\beta'(S) - s(1 - p)(\pi\beta'(F, e = 1, C) + (1 - \pi)\beta'(F, e = 0)) \\ &= c - s(1 - p)(1 - \pi)\frac{\beta}{1 - p(1 - \beta)}\end{aligned}$$

If instead player 2 departs from the proposed equilibrium and chooses to defect, then posterior be

$$\begin{aligned}u_2^G(D, \beta') &= d - s(\pi\beta'(F, e = 1, D) + (1 - \pi)\beta'(F, e = 0)) \\ &= d - s\left(\pi + (1 - \pi)\frac{\beta}{1 - p(1 - \beta)}\right)\end{aligned}$$

Cooperation is therefore a best response if and only if

$$c - s(1 - p)(1 - \pi)\frac{\beta}{1 - p(1 - \beta)} > d - s\left(\pi + (1 - \pi)\frac{\beta}{1 - p(1 - \beta)}\right),$$

or,

$$d < c + s\left(\frac{\pi(1 - p) + p\beta}{1 - p(1 - \beta)}\right) \quad (2)$$

As for the no-trust analysis above, this inequality is straightforward. If the expected shame costs of defection,  $s\left(\frac{\pi(1 - p) + p\beta}{1 - p(1 - \beta)}\right)$ , are large enough then they outweigh the monetary benefit of defection,  $d - c$ .

The comparative statics on the expected shame costs of defection,  $s\left(\frac{\pi(1 - p) + p\beta}{1 - p(1 - \beta)}\right)$ , are similar to those discussed above for the shame premium for a defection equilibrium. A higher sensitivity to shame (a higher  $s$ ) will cause these costs to

increase. Also, these costs increases in  $\pi$ , the likelihood of exposure and also in  $p$ , the likelihood that nature does not cause bad luck.<sup>10</sup>

It turns out that our previous analysis of conditions for which a no-trust equilibrium will not exist is useful vis-a-vis conditions for a trust equilibrium to exist:

**Claim 4** *A1 implies (2)*

**Proof.** It suffices to show that  $s\beta(\pi + p - \pi p) < \frac{\pi(1-p)+p\beta}{1-p(1-\beta)}$ . We have,

$$\begin{aligned}
& \frac{\pi(1-p) + p\beta}{1-p(1-\beta)} - \beta(\pi + p - \pi p) \\
= & \frac{\pi(1-p) + p\beta - \beta(\pi + p - \pi p)(1-p(1-\beta))}{1-p(1-\beta)} \\
= & \frac{\pi - \pi p + p\beta - \beta\pi - \beta p + \beta\pi p + \beta\pi p + \beta p^2 - \beta\pi p^2 - \beta^2\pi p - \beta^2 p^2 + \beta^2\pi p^2}{1-p(1-\beta)} \\
= & \frac{\pi(1-\beta)(1-p) + \beta\pi p(1-\beta)(1-p) + p^2\beta(1-\beta)}{1-p(1-\beta)} > 0
\end{aligned}$$

■

Intuitively, (2) is implied by A1 because when beliefs about the action of good types are better (higher  $\sigma$ ) then the Bayes updating after receiving a payoff of zero is more severe (higher  $\beta'$ ) which creates stronger incentives for a good type to cooperate. Thus, if A1 is satisfied, this allows for the provision of incentives for a good player 2 for the worse beliefs about  $\sigma$ , which means that incentives are even stronger for higher values of  $\sigma$ .

**Remark 5** *If A1 is violated but (2) is satisfied, then there will be multiple Equilibria (both pure strategy equilibria discussed above and a mixed strategy equilibrium.) In that case the comparative statics discussed below will still be valid.*

## 4 Empirical Implications

To fix ideas, let  $\beta$  satisfy A2 and imagine that there may be a distribution of good types with variation over  $s$ . Then, given any fixed monetary payoffs of  $v$ ,  $c$  and  $d$ , it is possible to consider comparative statics on the effect of changes in  $\pi$ , the exposure technology, on the equilibrium play of the game.

### 4.1 The Power of Shame

Recall from the analysis above that the “cost of shame” premium that was identified in (1) and (2) is increasing in the likelihood of exposure,  $\pi$ . Hence, it immediately follows that,

---

<sup>10</sup>Taking the derivative of  $\left(\frac{\pi(1-p)+p\beta}{1-p(1-\beta)}\right)$  with respect to  $p$  yields  $\frac{\beta-\pi\beta}{(1-p(1-\beta))^2} > 0$ .

**Corollary 6** *As  $\pi$  increases, the set of parameters for which cooperate is part of the unique equilibrium increases, and the set of parameters for which defect is part of any equilibrium decreases.*

This can be best seen by considering condition A1. For any given set of parameters, an increase in  $\pi$  will make it more likely that A1 is satisfied. This implies the following testable hypothesis:

**Hypothesis 1: The Power of Shame.** If shame plays a role in the decision to cooperate, an increase in  $\pi$  should cause (weakly) more cooperation by players 2.

I coin this hypothesis the “power of shame” since it identifies a manipulable intervention that affects the mechanism of 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, then changes in  $\pi$  should not have an effect on the outcomes observed. This is the first prediction that differentiates the theory of shame from other social preferences.

## 4.2 The Rationality of Trust

The “power of shame” conclusion in corollary (6) has implications about the behavior of player 2 in equilibrium: a higher exposure to shame will create stronger incentives 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. More precisely, for any given set of parameters, an increase in  $\pi$  will make it more likely that A1 is satisfied. As a result, a rational player 1 will trust player 2. It immediately follows that,

**Corollary 7** *As  $\pi$  increases, the set of parameters for which trust is part of the unique equilibrium increases, and the set of parameters for which no-trust is part of any equilibrium decreases.*

This implies the following testable hypothesis:

**Hypothesis 2: The Rationality of Trust.** If shame plays a role in the decision to cooperate, an increase in  $\pi$  should cause (weakly) more cooperation by players 2.

I coin this hypothesis the “rationality of trust” since it identifies a manipulable intervention that affects the incentives to trust through the mechanism of shame. Thus, if shame is a motivator for player 2 then changes in  $\pi$  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.

## 5 Experimental Design

The experimental design follows the theoretical 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 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. No one could participate in more than one session. Average earnings were  $\$x$  (including a  $\$7$  show-up fee), and each sessions took about 45 minutes to one hour.

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 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  $v = 5$ ,  $c = 10$ , and  $d = 14$  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>11</sup>

In each treatment, first A’s recorded their choices and their sheets were collected. Next, B’s chose whether to Roll or Don’t Roll a 6-sided die. B made this choice without knowing A’s actual choice In or Out, but the instructions explained that B’s choice would be irrelevant if A chose Out. This guarantees an observation for every B player. After the decisions of B’s were recorded and collected, a 6-sided die was rolled (by me) for each B. This was carefully explained to the participants in advance, to allow for anonymity of B’s who chose Don’t Roll. This roll was relevant if and only if (In, Roll) had been chosen. The outcome corresponding to a success occurred only if the die came up 2, 3, 4, 5, or 6 after a Roll choice (hence,  $p = \frac{5}{6}$ ).

The first four sessions had four treatments each as follows:

1. **Stranger-Noise (SN):** 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 ( $\pi = 0$ ).
2. **Stranger-No-Noise (SNN):** 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 ( $\pi = 1$ ).
3. **Partner-Noise (PN):** In this treatment participants did know who they were matched with (pairs of ID numbers were announced before the deci-

---

<sup>11</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).

sions and each pair acknowledged each other by standing). A-players did not learn why they received a payoff of zero if they did ( $\pi = 0$ ).

4. **Partner-No-Noise (PN):** In this treatment participants knew who they were matched with (as in PN) and A-players did learn why they received a payoff of zero if they did ( $\pi = 1$ ).

The next six sessions included the following two treatments:

5. **Stranger-Public (SP):** In this treatment participants did not know who they were matched with. After the sheets were collected for each player B, I publically announced what that player B had chosen. (This is like  $\pi = 1$  for all player A's, without knowing who they were matched with).
6. **Partner-Public (PP):** In this treatment participants did know who they were matched with. After the sheets were collected for each player B, I publically announced what that player B had chosen.

Starting with the first four treatments, there is a clear informational ranking between treatment PN and PNN: this is exactly the game described in the theoretical analysis. Hence, the theory predicts that there will be more cooperation (the power of shame) and more trust (the rationality of trust) in treatment PNN. At some level, the same comparison should be true for SN versus SNN, but now the shame is “shared” among all the B players.

This seems to imply a kind of “free rider” problem in which the power of shame is not as severe as it would be if identities were known. As such, the theory implies that both cooperation and trust will be weaker in the stranger settings than the partner settings, other things equal. However, since other things are not equal it is not possible to infer the ranking of cooperation and trust between the SNN and PN treatments. What is known is that these will both have higher levels of cooperation and trust compared to SN, and lower levels compared to PNN.

The last two treatments are an attempt to disentangle the “partner-effect”, of having an identity of a partner revealed, and the “public shame effect”, of having the action of player 2 announced to all. If shame is the primary motivator, then it is the announcement and not the partnering that should matter, implying that the results in PNN should be similar to those in SP and PP.

## 6 Experimental Results

### 6.1 The Power of Shame

The experimental results corroborate the predictions of the theory. Table 1 shows the raw data describing the behavior of B players in the six treatments.

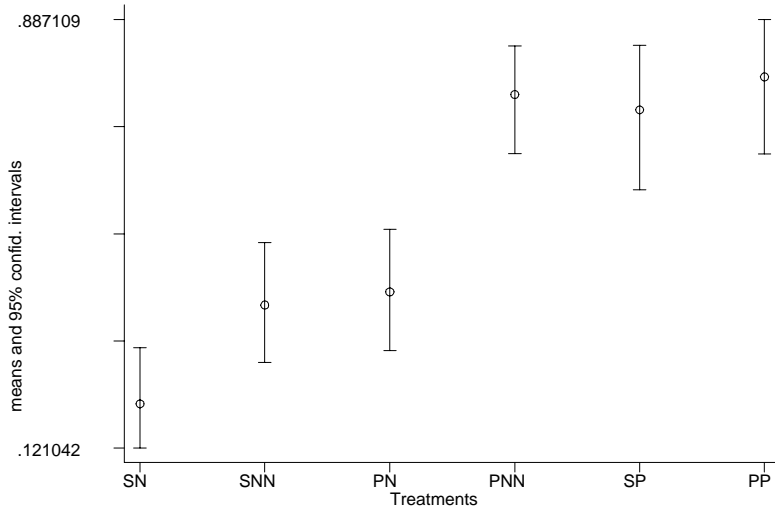


Figure 3: Percentage of B Players Who Choose Cooperate in the Six Treatments

Variable	Obs	Mean	Std. error.	Low 95%	High 95%
SN	85	.2	.0433861	.1210418	.3007931
SNN	85	.3764706	.0525514	.2736255	.4881878
PN	85	.4	.0531369	.2951936	.5119835
PNN	85	.7529412	.0467812	.647471	.8401019
SP	51	.7254902	.0624899	.5825525	.8410727
PP	51	.7843137	.0575932	.6467859	.8871094

Table 1: Percent of B players who chose Roll (Cooperate)

The results of Table 1 are also shown in Figure 3.

## 6.2 The Rationality of Trust

Table 3 shows the raw data describing the behavior of A players in the six treatments.

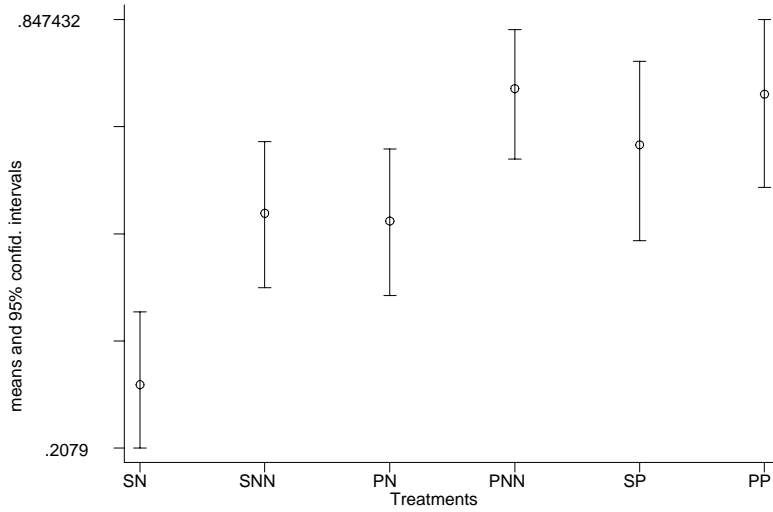


Figure 4: The Percentage of A Players Who Chose Trust in the Six Treatments

Treatment	Obs	Mean	Std. err.	Low 95%	High 95%
SN	86	.302	.050	.208	.411
SNN	86	.558	.054	.447	.665
PN	86	.547	.054	.435	.654
PNN	86	.744	.047	.639	.832
SP	53	.660	.065	.517	.785
PP	53	.736	.061	.597	.847

Table 3: Percent of A players who chose In (Trust)

The results of Table 3 are also shown in Figure 4.

### 6.3 Monotonicity and Reciprocity

TBA

## 7 Discussion

TO BE COMPLETED

- Dictator games and exit options (Dana, Cain and Dawes (2005), Lazear, Malmendier and Weber (2006)).

- How do I reconcile ultimatum games? Seems more like reciprocity... If I add two dimensions over which shame occurs: being greedy when offering or responding, and being a wimp when responding or offering, then this may be resolved. Also, there can be heterogeneity in societies if these levels are relative to the expected norm. Can get extreme giving, versus extreme selfishness. [**Richer model:** beliefs over weakness]

## 8 References

- Andreoni, James (1990) "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving," *Economic Journal* **100**:464-77.
- Battigalli, Pierpaolo and Martin Dufwenberg (2005), "Dynamic Psychological Games", mimeo.
- Berg, Joyce, John Dickhaut, and Kevin McCabe (1995) "Trust, Reciprocity, and Social History," *Games and Economic Behavior* **10**:122-142,
- Camerer, Colin *Behavioral Game Theory: Experiments in Strategic Interaction*. 2003, Princeton University Press, Princeton, NJ.
- Geanakoplos, John, David Pearce and Ennio Stacchetti (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, 1:60-79.
- Charness, Gary and Martin Dufwenberg (2005) "Promises and Partnership," forthcoming *Econometrica*.
- Dana, J., D. M. Cain and R. Dawes. (2005) "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." Forthcoming, *Organizational Behavior and Human Decision Processes*.
- 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.
- Sorting in Experiments with Application to Social Preferences
- Lazear, Edward, Ulrike Malmendier and Roberto Weber (2006) "Sorting in experiments with application to social preferences," mimeo,
- Rabin, Matthew (1993) "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, **83**:1281-1302
- Tangney, June (1995), "Recent Advances in the Empirical-Study of Shame and Guilt", *American Behavioral Science*, 38:1132-45.

## **9 Appendix**

### **9.1 Experiment Instructions**

# 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 four 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 four 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 yet to be 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 randomly 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:

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

At the end of the each decision task, person A who receives a payoff of \$0 will sometimes be told whether his matched person B chose Roll or DON'T ROLL, and sometimes will not be told.

Your final payment will be the participation fee of \$7, plus the payoff from one of the four decision pairs to be chosen randomly.

# **A1**

## Decision Sheet

My Decision is:  
(please circle one)

**IN**

**OUT**

---

# **B1**

## Decision Sheet

My Decision is:  
(please circle one)

**ROLL**

**DON'T ROLL**