On Blame and Reciprocity: An Experimental Study*

Boğaçhan Çelen† Melbourne Business School

Andrew Schotter‡ New York University

Mariana Blanco§ Universidad del Rosario

May 24, 2016

Abstract

The theory of reciprocity is predicated on the assumption that people are willing to reward kind acts and to punish unkind ones. This assumption raises the question of what *kindness* is. In this paper, we offer a novel definition of kindness based on a notion of blame. This notion states that for player j to judge whether or not player i is kind to him, player j has to put himself in the position of player i, and ask if he would act in a manner that is worse than what he believes player i does. If player j would act in a worse manner than player i, then we say that player j does not blame player i. If, however, player j would be nicer than player i, then we say that player j blames player i. We believe this notion is a natural, intuitive and empirically functional way to explain the motives of people engaging in reciprocal behavior. After developing the conceptual framework, we test this concept by using data from two laboratory experiments and find significant support for the theory.

JEL Classification Numbers: A13, C72, D63. **Keywords**: Altruism, blame, reciprocity, psychological games.

*We appreciate very useful comments of two anonymous referees, Pierpaolo Battigalli, Martin Dufwenberg, Colin Camerer, Matthew Rabin and Erkut Özbay. Mi Luo provided excellent assistance on computational exercises. We are grateful to the participants of the CESS Experimental Economics Lunchtime Seminar, 2009 North-American ESA Conference, Amsterdam Workshop on Behavioral & Experimental Economics, Cornell University Behavioral Economics Workshop, Rutgers University, Brown University Microeconomics Seminar and SfED 2012 Winter Conference, University of Birmingham, University of Exeter, Melbourne Univerity, Oxford University, Vienna University, Universitat de Barcelona, Universitat Pompeu Fabra, Universitat Autònoma de Barcelona, Freie Universität Berlin, California Institute of Technology, and UNSW for comments. We also acknowledge the partial financial support of the Center for Experimental Social Science at NYU.

†C.E.S.S., New York University, and Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton VIC 3053, Australia. E-mail:bc319@nyu.edu, url: http://bogachancelen.com/.

‡C.E.S.S. and Department of Economics, New York University, 19 W. 4th Street, New York, NY 10012, USA. E-mail: andrew.schotter@nyu.edu, url: http://homepages.nyu.edu/∼as7/.

§Faculdad de Economía, Universidad del Rosario, Calle 14 # 4-80, Oficina 207, Bogotá, Colombia. E-mail: mariana.blanco@urosario.edu.co, url: http://mbnet26.googlepages.com/home/.

1 Introduction

Recent years have witnessed a growing literature on the theory of reciprocity. Founded on the seminal work of Rabin [25] (henceforth Rabin)—further extended by Falk and Fischbacher [14], Dufwenberg and Kirschteiger [10], Battigalli and Dufwenberg [2] (henceforth B&D) and other scholars—the theory of reciprocity is predicated on the assumption that people are willing to reward kind acts and to punish unkind ones.[1] This approach raises the question of how to define *kindness*. In this paper, we offer a novel definition of kindness based on a notion of blame.

Put most simply, the notion of blame states that for player j to judge whether or not player i is kind to him, player j has to put himself in the position of player i, and ask if he would act in a manner that is worse than what he believes player i does. If player j would act in a worse manner than player i, then we say that player j does not blame player i for his behavior. If, however, player j would be nicer than player i, then we say that "player j blames player i" for his actions—i.e. player i's actions are blameworthy.

This way of viewing kindness is distinctly different from other theories in a number of ways. Following the criteria that were discussed in Schotter [26], our approach leads to an endogenous, context-dependent, and process-oriented theory. It is endogenous because players judge the actions of others by their own standards, and not by some exogenous standards imposed by the analyst. In addition, our approach allows the standards people use to judge the actions of others to differ from person to person depending on their personal norms. This is crucial, as actions that bother one person may not bother other people at all, or those actions that strike some people as being fair may be very upsetting to others. Consequently, this feature differentiates our theory from the theories that impose an exogenous norm in order to determine what is considered kind. In the current framework, blame is self-referential: It only matters what you would have done in your opponent's situation and not how the actions of others are compared to some exogenous norm.

Another important feature of our approach is that the theory is sensitive to the institutional setting. For instance, actions that are blame-free in a prison may certainly be blameworthy in civilian life. One cannot judge other people's behavior in isolation—we need to know the context they are in. This is fundamentally different than the existing theories, which assume that players' preferences are independent of the context. For example, in a leading paper Levine [23] (henceforth Levine) takes this approach to analyze experimental evidence in ultimatum, centipede, and public goods experiments. Gul and Pesendorfer [18] lay the foundations of interdependence between behavioral types, inde-

[1] For a comprehensive survey on reciprocity see Sobel [27].

pendent of the environment that decision-makers interact.

Finally, our notion of blame judges the actions of people that lead to outcomes, and not merely the outcomes themselves. Thus, it is a process-oriented theory, and it differs from those theories that are outcome-based in that respect.

In this paper, we give a formal definition of what it means to put oneself in another players strategic position, how to construct blame, how blame affects a player's preferences, and finally what an equilibrium looks like when people have blame-dependent preferences.

To provide a quick insight into how our theory differs from others, consider an ultimatum game played between two players, p (Proposer) and r (Receiver), whose preferences exhibit inequity aversion à la Fehr and Schmidt [16] (henceforth F&S) or Bolton and Ockenfels [4]. Let $u_i(x_p, x_r)$ represent the preferences of player i when the final allocation is x_p and x_r for the Proposer and the Receiver, respectively.

Given F&S preferences, in the subgame perfect equilibrium of the game, the Receiver rejects an offer (x_p, x_r) if $u_r(x_p, x_r) < u_r(0, 0)$. In our formulation of preferences that exhibit blame, the Receiver blames—hence perhaps rejects—an offer if that offer is less generous than the one he would have made had he been in the Proposer's position. In other words, according to our hypothesis, the Receiver compares the offer (x_p, x_r) to the offer he would have made if he were the Proposer, say (x_p^*, x_r^*). If the offer (x_p, x_r) is more generous than (x_p^*, x_r^*), then he accepts the offer, otherwise he blames the Proposer. If blame causes a sufficiently high disutility for the Receiver, he will reject the offer.

Recently, Gurdal et al. [19] have also studied the notion of blame. Gurdal et al. [19] provide experimental evidence that in a simple principal-agent framework a subject (the principal) tends to punish another subject (the agent) who makes a decision on his behalf, not for the decision that he makes but for the outcome of his decision. More precisely, an agent is given a task to choose between a lottery and a safe alternative whose proceedings determine the earnings of the principal. Then, a public toss of a die determines the outcome of the lottery. In the last stage, the principal decides how to allocate a fixed amount of money between the agent and a random third party. In this carefully designed experiment, Gurdal et al. [19] show that a principal is more likely to punish an agent based on the outcome of the agent's decision. In Gurdal et al. [19] the notion of blame is equivalent to punishment and the focus is on the ex ante vs ex post evaluation of the agent's performance. In contrast, our approach would determine blame by the difference of the principal's expected payoff as a result of the agent's decision and what he could have received if he were in the agent's position making the decision himself.

In Section 3.1.1 we will elaborate on the relation of our theory to those of F&S, Rabin and Levine in the context of our experiment.

1.1 Overview and Summary

As our introductory discussion indicates, the essence of the notion of blame involves the examination of a counterfactual, i.e. imagining what you would have done if you were in the position of the person whose actions you are judging. In the lab, by allowing subjects to play all the roles in a game anonymously, we can operationalize this counterfactual thought experiment.

The two experiments reported in this paper do just that. In the first experiment, we take a simple dictator game and implement it in two stages. Even though the subjects know that there were two stages, they do not know what will transpire in stage 2 until the experiment is completed. In stage 1, the subjects split 10 tokens between themselves and another anonymous subject in the lab. In stage 2, the subjects are randomly matched with another subject in the lab. After they are matched, they are offered, as Receiver, the amount the subject that they matched with sent as Sender in stage 1.[2] The Receiver does not have the option of rejecting proposals. Rather, he can punish the Sender by reducing his payoff by 1 token at no cost to himself. Note that this design places each subject both in the role of a Sender and a Receiver in stages 1 and 2, respectively. Hence, in stage 2, when a subject receives an offer, he is able to compare that offer to the offer he made as a Sender in stage 1.

According to the theory of blame, the subjects should only punish if the offer they receive is less than the offer they made in stage 1. This prediction differs from the predictions of the theories that we mentioned vide supra. Our data suggest that a significant fraction of the subjects behave consistent with our prediction. Furthermore, subjects' behavior is consistent with the theory of blame more often than the other theories.

Our second experiment uses the data generated by Carpenter, Kariv, and Schotter [7] (henceforth CKS) who run a public goods experiment with punishment. Symmetric public goods games are useful for our purposes, because in such games, at the contribution stage, each player is in fact in the strategic position of others. Consequently, they can judge whether any other player acted more kindly or less kindly than they did. The experiment varies the network that players are connected through in the punishment phase, which allows us to better identify the punishment motives of the subjects as they change from one network to another.

Note that even though we present a full equilibrium analysis, our experiments aim only to examine whether or not subjects exhibit blame preferences. In order to test the equilibrium we would need to elicit first- and second-order beliefs of all our subjects which would be cumbersome. Nonetheless, our experiments allow us unambiguously to test

[2] In the remainder of the paper, we use the term "offer" for the dictator's allocation decision despite the fact that it is not truly an offer since there is no rejection possibility.

whether our subjects use blame as the basis for reciprocity.

This paper is organized as follows. Section 2 introduces the *blame* concept in a rigorous manner and provides an appropriate equilibrium concept. We also compute the equilibria of a formal example to further elaborate the equilibrium concept. Section 3 explains our experimental designs, discusses the theoretical predictions of various theories, analyzes the data, and states our results. Section 4 concludes.

2 Blame in Games

2.1 Definitions

In order to introduce the notion of blame, we will follow the general *psychological dynamic games* framework of B&D. The literature of *psychological games* pioneered by Geanakoplos et al. [17] assumes that a player's preferences not only depend on the outcome of the game, but also on the hierarchy of beliefs about the other players' strategies and beliefs. B&D extend the theory in a number of directions, and accommodate the possibility that players update their beliefs at each decision node along a path in a sequential game. This is a particularly important extension since it allows one to address the dynamic psychological effects that can be observed during the course of a game.

Our discussion will be confined to 2-player, finite-horizon, multi-stage games with perfect information under complete information. The purpose of this section is to provide a rigorous discussion of how we formulate blame in this framework. As B&D point out, this class includes simultaneous moves games, perfect information games, and repeated games as special cases. For a discussion of the generalization to imperfectly observable actions, chance moves, and asymmetric information we refer the reader to Section 6 in B&D.

Players, Actions, and Histories. Consider a multi-stage game consisting of 2 players $N = \{1, 2\}$. The set of histories is comprised of the initial history $\varnothing$ of length 0, as well as all finite histories. A history of length ℓ is a sequence $h = (a^1, \cdots, a^\ell)$, where $a^t = (a_1^t, a_2^t)$ is a profile of actions at time $t \in \{1, \cdots, \ell\}$. We study games of perfect information, that is, a history h becomes public information as soon as it is realized.

H denotes the set of all non-terminal histories. We denote the set of finite actions available to player i at the history $h \in H$ by $A_i(h)$. If $A_i(h)$ is singleton, then player i is not active. A history is *terminal* if and only if $A_i(h)$ is empty for both players. We refer to a terminal history as an *outcome*, and denote the set of all outcomes by Z. In order to distinguish an outcome from a non-terminal history, we denote a generic outcome by z.

Each outcome z is associated with a *material payoff* for each player. The function π_i :

$Z \mapsto \mathbb{R}$ determines player i's material payoff, $\pi_i(z)$, at the outcome z. Finally, for any $h \in H \cup Z$ of length ℓ, we define the set of histories that pass through h; $H_h := \{\hat{h} \in H \cup Z : \hat{h} = (h, a^{\ell+1}, \cdots, a^{\ell+m}),$ for a profile of actions $(a^{\ell+1}, \cdots, a^{\ell+m}).\}$

Meta-Players. Our definition of blame makes reference to what a player would do in another player's position: a player puts himself in the other player's position and asks what he would do, and what beliefs he would hold, if he were in that position. More precisely, from player i's perspective, there are two games to consider: the game he plays against player j, and the *fictitious* game he plays against himself in player j's position. In order to formally address the latter game, we introduce meta-players that represent "a player who considers himself in the other player's position."

For each player $i \in N$, we define the *meta-player* $\langle ij \rangle$, for *player i in player j's position*, for $j \neq i$. Hence, given $N = \{1, 2\}$, we define the set of all players $\hat{N} := \{1, 2, \langle 12 \rangle, \langle 21 \rangle\}$. In the remainder of the paper, we use the notation i, j to refer to players in N, and we reserve the notation k, l for any player in $\hat{N}$.

Since player $\langle ij \rangle$ represents player i in player j's position, we assert that the actions that are available to player j at history h are also available to player $\langle ij \rangle$, i.e., $A_{\langle ij \rangle}(h) := A_j(h)$ for all $h \in H,\ i, j \in N$. Similarly, we define $\pi_{\langle ij \rangle}(z) := \pi_j(z)$.

Strategies. For any player $k \in \hat{N}$, we denote the set of pure strategies by S_k. A strategy of a player k is a function, $s_k : H \mapsto \cup_{h \in H} A_k(h)$, which maps a non-terminal history h to the set of actions available to him at h.

Given a strategy profile (s_i, s_k), we write $\zeta(s_i, s_k) \in Z$ to denote the outcome induced by players i's, and k's strategies, for $i \in N$, and $k \in \{j, \langle ij \rangle\}$.[3] Given a history $h \in H \cup Z$, we define $S_k(h)$ as the set of player k's strategies that allow history h.[4]

Conditional Beliefs. In order to define blame, we need two orders of beliefs. The first-order beliefs of player $k \in \hat{N}$ are about other players' strategies. We assume that players' strategies are not correlated, and write $\mu^1_{kl}(\cdot|\cdot) : 2^{S_l} \times H \mapsto [0, 1]$, $\mu^1_{kl} \in \Delta^H(S_l)$ to denote player k's conditional beliefs about player l's strategy. Player k's second-order beliefs, $\mu^2_{kl}(\cdot|\cdot) : \mathcal{B}_{\Delta^H(S_l)} \times H \mapsto [0, 1]$, are about player l's first-order beliefs about player k's strategy, where $\mathcal{B}_{\Delta^H(S_l)}$ is the Borel sigma-algebra of $\Delta^H(S_l)$.

Our definition of blame is based on the following conditional beliefs. For $i, j \in N$,

• μ^1_{ij} is player i's belief about player j's strategy,

• $\mu^1_{i\langle ij \rangle}$ is player i's belief about the strategy that he would play, if he were in player j's position,

• $\mu^1_{\langle ij \rangle i}$ is a belief that player i would hold about his own strategy in player j's position,

[3] We define the outcome for (s_i, s_k), where $i \in N$, and $k \in \{j, \langle ij \rangle\}$, because $(s_i, s_{\langle ji \rangle})$ does not induce any outcome, and $(s_{\langle ij \rangle}, s_{\langle ji \rangle})$ is not relevant in our framework.

[4] Formally, $S_i(h) := \{s_i \in S_i :$ for any $z \in H_h \cap Z,\ \zeta(s_i, s_k) = z$ for some $s_k \in S_k,\ k \in \{j, \langle ij \rangle\}\}$, and $S_{\langle ij \rangle}(h) := \{s_{\langle ij \rangle} \in S_{\langle ij \rangle} :$ for any $z \in H_h \cap Z,\ \zeta(s_i, s_{\langle ij \rangle}) = z$ for some $s_i \in S_i.\}$

• μ^2_{ij} is player i's belief about player j's belief about his strategy,
• $\mu^2_{i\langle ij\rangle}$ is player i's belief about the beliefs that he would hold in player j's position about his own strategy.

Therefore, for player $i \in N$ we write $\mu_i := (\mu^1_{ij}, \mu^1_{i\langle ij\rangle}, \mu^2_{ij}, \mu^2_{i\langle ij\rangle})$. We denote the set of analogous beliefs for player k by M_k, and define $M := \prod_{k\in\hat{N}} M_k$.

Blame and Preferences. Having discussed players' beliefs, we are now ready to define *blame*, and incorporate it into players' preferences. Let us focus on how player i judges player j, given his beliefs $\mu_i(\cdot|h) \in M_i$ at history $h \in H$. To do this, we define the expected material payoff of player i based on his beliefs $\mu_i(\cdot|h)$, when he plays against player j, and when he plays against $\langle ij\rangle$. In the former case, we have

$$\mathsf{E}_{\mu_i}\big[\pi_i\big(\zeta(s)\big)\big|h\big] := \int_{S_i\times S_j} \pi_i\big(\zeta(s)\big) d\mu^1_{ij}(s_j|h)\mu^2_{ij}(s_i|h). \tag{1}$$

Note that this is player i's expected material payoff based on his beliefs about player j's strategy, and about player j's beliefs regarding his own strategy.

In the latter case, player i plays against himself in player j's position, i.e. player $\langle ij\rangle$. Hence, the outcome that is induced by this interaction is $\zeta(s_i, s_{\langle ij\rangle})$. As a result, if player i were in player j's position, he would create an expected material payoff of

$$\mathsf{E}_{\mu_i}\big[\pi_i\big(\zeta(s_i, s_{\langle ij\rangle})\big)\big|h\big] := \int_{S_i\times S_j} \pi_i\big(\zeta(s_i, s_{\langle ij\rangle})\big) d\mu^1_{i\langle ij\rangle}(s_{\langle ij\rangle}|h)\mu^2_{i\langle ij\rangle}(s_i|h), \tag{2}$$

for himself.

The difference between player i's expected material payoffs when he plays against player j and player $\langle ij\rangle$ determines whether or not player i *blames* player j.

Definition 1. *A player $i \in N$ who holds beliefs μ_i is said to blame player $j \in N$ at history h, if*

$$\delta_{ij}(\mu_i|h) := \mathsf{E}_{\mu_i}\big[\pi_i\big(\zeta(s_i, s_{\langle ij\rangle})\big)\big|h\big] - \mathsf{E}_{\mu_i}\big[\pi_i\big(\zeta(s)\big)\big|h\big] > 0. \tag{3}$$

Let us clarify the intuition behind the definition with the following statement from player i: *I blame player j because the expected material payoff that I think he expects that I will get when I play against him, (1), is less than my expected material payoff if I played against myself in his position, (2). In other words, if I was in his position, I would be nicer to the player in my position than he is to me, $\delta_{ij}(\mu_i|h) > 0$.*

The next step is to define preferences of players. Let us define the psychological utility function of player i by $U_i(z, \mu_i) : Z \times M \mapsto \mathbb{R}$. In order to address the role of blame in players' preferences explicitly, we define the function $u_i(z, \delta_{ij})$, such that $u_i(z, \delta_{ij}(\mu_i|h)) := U_i(z, \mu_i(\cdot|h))$. We posit that blame inflicts disutility. That is, a player prefers an outcome

without blame to the same outcome with blame. In short, we assume that u_i is non-increasing in δ_{ij}.

Now, let us discuss meta-player $\langle ij \rangle$'s preferences. In order to do that, we should define what blame means for meta-players. Note that the only blame term that needs to be scrutinized is $\delta_{\langle ij \rangle i}$: how would player i blame himself when he puts himself in player j's position? Note that $\delta_{\langle ij \rangle i}$ is determined by what "player i who considers himself in player j's position" believes player i plays and what he would have done in player i's (his original) position. Such blame is determined by player $\langle ij \rangle$'s preferences and his beliefs. Clearly, when player i places himself in player j's position he is free to hold any beliefs he wishes and if those beliefs differ from the action that player i ultimately intends to take, then player $\langle ij \rangle$ may blame player i, i.e., there may be self blame. In Appendix B, under an innocuous assumption, we prove that this in fact cannot be the case in an equilibrium. In the remainder of the paper, to simplify our analysis we will assume that there is no self blame between players $\langle ij \rangle$ and i.

We are ready to define the preferences of meta-players. Since meta-players' preferences do not exhibit belief-dependence of the form that it takes psychological game theory to model, the utility function of player $k \in \hat{N} \setminus N$ with type b_k is $u_k(z) : Z\mathbb{R} \mapsto \mathbb{R}$.

Having defined players' preferences, we can now formally define a blame game.

Definition 2. *Given an extensive form with material payoffs, $\langle N, H, (\pi_i)_{i \in N} \rangle$, we define a blame game as $\langle \hat{N}, H, (u_k)_{k \in \hat{N}} \rangle$.*

Example 1. *Let us elaborate more on the preferences of player $i \in N$ with an example. Let the utility function of player i be*

$$u_i(z, \delta_{ij}(\mu_i|h)) := \pi_i(z) + \Big(b_i - \delta_{ij}(\mu_i|h)\Big)\pi_j(z).$$

This specification indicates that player i's utility is determined by the sum of his material payoff, and a proportion of the other player's material payoff. The term $\big(b_i - \delta_{ij}(\mu_i|h)\big)$ determines the weight attached to player j's material payoff. The first term b_i is the weight that player i puts on player j's material payoff without any blame consideration. However, the second term captures the idea that if player i blames player j, the overall weight decreases. If player i does not blame his opponent, he assigns a constant, non-negative weight b_i to his opponent's material payoff. As player i blames more, the weight he assigns to player j's material payoff decreases, and in fact, when it is high enough (i.e. $\delta_{ij}(\mu_i|h) \geq b_i$), player i becomes hostile to player j. In this example, player i's type is b_i, indicates his altruism towards the other player as his innate characteristic.

Now, suppose that player $\langle ij \rangle$'s utility function is as follows:

$$u_{\langle ij \rangle}(z) := \pi_j(z) + b_i\pi_i(z).$$

When player i considers himself in player j's position, his utility is still derived by the weighted sum of his material payoff (in player j's position) and sum of other players' material payoffs. Note that when he considers himself in player j's position, his preferences are not affected by any blame considerations towards himself. Moreover, in this example, player i retains the weight b_i in player j's position. In other words, when player i considers himself in player j's position, he still puts the weight b_i on the material payoff of the other player; i.e. player i's.

2.2 Equilibrium

The equilibrium concept that we will define for a blame game, $\langle \hat{N}, H, (u_k)_{k\in\hat{N}}\rangle$, is based on B&D who extend the sequential equilibrium of Kreps and Wilson [22]. The equilibrium concept requires randomized choices. In our discussion so far, we focused solely on pure strategies. The interpretation of randomized choices—as in Aumann and Brandenburger [1]—is that randomized choice of a player i is as player j's and meta-players' common first-order beliefs about player i's strategy. Therefore, we define behavioral strategies $\sigma_i = (\sigma_i(\cdot|h))_{h\in H} \in \prod_{h\in H} \Delta(A_i(h))$, as a vector of common first-order beliefs held by other players. In the tradition of Kreps and Wilson [22], an assessment is a profile $(\sigma, \mu) := (\sigma_i, \mu_i)_{i\in\hat{N}}$, where σ is a behavioral strategy profile and $\mu \in M$. We assume that behavioral strategies are independent across histories. Then, we say that first-order beliefs about player l's strategy are derived from a behavioral strategy σ_l, if for all $k \in \hat{N},\ s_l \in S_l$,

$$\mu^1_{kl}(s_l|h) := \prod_{\hat{h}\in H_h\setminus Z} \sigma_l(s_l(\hat{h})|\hat{h}).$$

Now we can define consistency of assessments.

Definition 3. *As assessment (σ, μ) is consistent if*
(a) $(\mu^1_k)_{k\in\hat{N}}$ is derived from σ,
(b) second-order beliefs in μ assign probability 1 to first-order beliefs, i.e. for all $i \in N$, $k \in \hat{N}$, $h \in H$, $\mu^2_{ik}(\cdot|h) = \mu^1_{ki}(\cdot|h)$.

In addition to consistency, sequential rationality is a requirement of equilibrium behavior. Note that for player $i \in N$, given his beliefs μ_i at history $h \in H$, a strategy s_i yields an expected utility

$$\mathsf{E}_{\mu_i}[u_i|h] := \int_{S_j} u_i\big(\zeta(s_i, s_j), \delta_{ij}(\mu_i|h)\big) d\mu^1_{ij}(s_j|h).$$

Similarly, for player $\langle ij\rangle \in \hat{N}\setminus N$, given his beliefs $\mu_{\langle ij\rangle}$ at history $h \in H$, a strategy $s_{\langle ij\rangle}$

yields

$$\mathsf{E}_{\mu_{\langle ij\rangle}}[u_{\langle ij\rangle}|h] := \int_{S_i} u_{\langle ij\rangle}\big(\zeta(s_i, s_{\langle ij\rangle})\big) d\mu^1_{\langle ij\rangle i}(s_i|h).$$

As usual, we expect that players choose their strategies to maximize their expected utilities as defined above. We are ready to define the equilibrium.

Definition 4. *An assessment (σ^*, μ^*) is a sequential equilibrium, if it is consistent, and for all $h \in H$, and for $l \in \hat{N}$, $s_l \in S_l(h)$, if $\mu^{*1}_{kl}(s_l|h) > 0$ then $s_l \in \arg\max_{s_l \in S_l(h)} \mathsf{E}_{\mu^*_l}[u_l|h]$ for all $k \in \hat{N}$.*

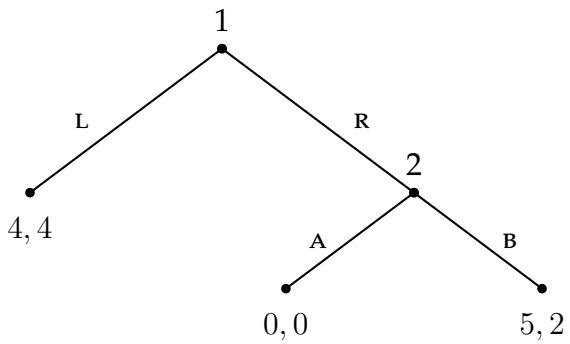
Let us summarize what happens in the equilibrium of a blame game. A player $i \in N$ judges the other player's actions by comparing what he believes player j does with what he believes he would do in their positions. This comparison defines blame. His utility function u_i is affected by blame, and in the equilibrium, given his beliefs he chooses the strategy that maximizes his expected utility. Also, when player i considers himself in player j's position, he chooses a strategy that maximizes $u_{\langle ij\rangle}$. Finally by consistency, all the beliefs are correct in the equilibrium.

2.3 A Simple Example

This section aims to illustrate the notion of blame and the equilibrium analysis by using the simple sequential game displayed in Figure 1. Throughout the analysis, for ease of exposition, we will focus on pure strategy equilibrium.

In this game, there are three outcomes (L), (R,A), and (R,B), and the material payoffs corresponding to them are given just below the terminal nodes.

Figure 1: A Two-person Extensive Form Game with Blame.



Let the utility of players i and $\langle ij\rangle$ be as in Example 1.

$$\begin{aligned} u_i(z, \delta_{ij}(\mu_i|h)) &:= \pi_i(z) + \Big(b_i - f\big(\delta_{ij}(\mu_i|h)\big)\Big)\pi_j(z), \\ u_{\langle ij\rangle}(z) &:= \pi_j(z) + b_i\pi_j(z). \end{aligned}$$

Assume that $1 \geq b_i \geq 0$, and f is a non-decreasing function of blame, such that $f(\delta_{ij}) = 0$ for all $\delta_{ij} \leq 0$. Also, we assume that $0 \leq f(\delta_{ij}) \leq 1$. Our choice of functional form suggests that if player i does not blame his opponent, he assigns a constant non-negative weight b_i to his opponent's material payoff. As player i blames more, the weight that he assigns to player j's material payoff decreases, and in fact, when it is high enough (i.e. $f\big(\delta_{ij}(\mu_i|h)\big) \geq b_i$) player i becomes hostile to player j.

Before we move into equilibrium analysis let us illustrate how blame is computed for a particular set of beliefs. Consider the case where (*i*) player 2 believes that his strategy, if he were in player 1's position, would be to play L, (*ii*) player 2 believes that player 1's strategy is to play R, and (*iii*) player 2 believes that player 1's belief about his strategy is that he plays B at the history (R). Note that if player 2 were in player 1's position, playing L would lead to a material payoff of 4 for a player in his position. On the other hand, player 2 believes that player 1 plays R, and that player 1 believes that he will actually play B. As a result, he expects a material payoff of 2. Therefore, while he would give a material payoff of 4 to a player in his position, he believes that player 1 gives him a payoff of 2. Hence, by Definition 1, he blames player 1 by a magnitude of 2.

Since there is a unique history at which each player moves, we will suppress the history and write s_k instead of $s_k(h)$. Recall that, in the equilibrium, higher order beliefs, and meta-players' beliefs have to be the same by the consistency requirement. Hence, in the equilibrium analysis that follows, we focus only on the cases where $\mu^1_{ij} = \mu^1_{\langle ji\rangle j} = \mu^2_{ji}$.

As the first step, let us determine meta-players' equilibrium behavior. In order to do that let us write their utilities.

$$u_{\langle 12\rangle}(z) = \begin{cases} 4 + 4b_1 & \text{if } z = (\text{L}) \\ 0 & \text{if } z = (\text{R,A}) \\ 2 + 5b_1 & \text{if } z = (\text{R,B}) \end{cases} \qquad u_{\langle 21\rangle}(z) = \begin{cases} 4 + 4b_2 & \text{if } z = (\text{L}) \\ 0 & \text{if } z = (\text{R,A}) \\ 5 + 2b_2 & \text{if } z = (\text{R,B}) \end{cases}$$

From $u_{\langle 12\rangle}$ we immediately observe that, by sequential rationality, $s^*_{\langle 12\rangle} = \text{B}$ for any $b_1 \geq 0$. Player $\langle 21\rangle$'s optimal behavior is determined as a response to his belief, $\mu_{\langle 21\rangle 2}$, about player 2's (his own) strategy. Note that, $\langle 21\rangle$'s best reaction to $\mu_{\langle 21\rangle 2}(\text{A}) = 1$ is L for all $b_2 \geq 0$, and his best reaction to $\mu_{\langle 21\rangle 2}(\text{B}) = 1$ is R (resp. L) if $b_2 \leq 1/2$ (resp. $b_2 \geq 1/2$). Having established meta-players optimal behavior, we move on to determining players' blame.

Tables 1a and 1b summarize players' blame at the indicated beliefs. For player i's blame δ_{ij}, the columns indicate beliefs about player j's strategy $(\mu^1_{ij}, \mu^2_{ji}, \mu^2_{j\langle ji\rangle})$, and the rows indicate beliefs about player i's strategy $(\mu^1_{ji}, \mu^2_{ij}, \mu^2_{i\langle ij\rangle})$.[5] For example, player 2's blame at history (R) given beliefs $\mu^2_{21}(\text{A}|(\text{R})) = \mu^1_{21}(\text{L}|(\text{R})) = \mu^2_{2\langle 21\rangle}(\text{A}|(\text{R})) = 0$ is $\delta_{21} = 0$ for $b_2 \leq 1/2$,

[5] We only write μ^1_{ij}, and μ^2_{ij} in the columns and rows, respectively, since other beliefs are the same in the equilibrium.

Table 1: Players' blame

(a) δ_{12}

	$\mu^1_{12}(\text{A}\|\varnothing) = 1$	$\mu^1_{12}(\text{A}\|\varnothing) = 0$
$\mu^2_{12}(\text{L}\|\varnothing) = 1$	$4-4$	$4-4$
$\mu^2_{12}(\text{L}\|\varnothing) = 0$	$5-0$	$5-5$

(b) δ_{21}

	$\mu^1_{21}(\text{L}\|(\text{R})) = 1$	$\mu^1_{21}(\text{L}\|(\text{R})) = 0$
$\mu^2_{21}(\text{A}\|(\text{R})) = 1$	$4-4$	$4-0$
$\mu^2_{21}(\text{A}\|(\text{R})) = 0$	$2-4$ if $b_2 \leq 1/2$ $4-4$ if $b_2 \geq 1/2$	$2-2$ if $b_2 \leq 1/2$ $4-2$ if $b_2 \geq 1/2$

Recall that by consistency of beliefs we impose that $\mu^1_{ij} = \mu^2_{j\langle ji\rangle} = \mu^2_{ji}$.

and $\delta_{21} = 2$ for $b_2 \geq 1/2$. This is shown at the lower right corner of Table 1b. Let us explain. Player 2 believes that other players believe that he plays B. Since he believes that player 1's strategy is R, his expected material payoff is 2. The material payoff that player 2 expects to get if he played against player $\langle 21\rangle$ depends on b_2. If $b_2 \leq 1/2$ player $\langle 21\rangle$'s optimal strategy is R, which leads to the same material payoff of 2 for player 2. However, when $b_2 \geq 1/2$, player $\langle 21\rangle$ plays L, which leads to a material payoff of 4. In summary, if $b_2 \leq 1/2$ player 2 would have followed the same strategy as player 1, and hence, he does not blame player 1. If $b_2 \geq 1/2$, contrary to player 1, player 2 would play L and would lead to a material payoff of 4, causing a difference of 2 in material payoffs. All the cases in Table 1 are computed similarly.

A word of clarification about Table 1 is in order. Recall that as players strategies unfold, in an equilibrium, their beliefs about the strategies of others' strategies and beliefs change consistently with the history of actions. In Table 1b, the first column is grayed out because at the history (R), player 2's belief that player 1's strategy is L with probability 1 is inconsistent; hence, it cannot be a part of any equilibrium.

Let us now determine player 2's equilibrium behavior at the history (R). In the equilibrium, we have $\mu^1_{21}(\text{L}|(\text{R})) = 0$. By using Table 1b, we determine player 2's utility at two outcomes (R,A) and (R,B) for different values of b_2. Table 2 states these utilities.

The shaded cells in Table 2 indicate that the outcome—hence, the strategy of player 2—is consistent with the corresponding beliefs in the columns. This helps us to determine

Table 2: $u_2(z, \delta_{21}(\mu_2|h))$ when $\mu^1_{21}(\text{L}|h) = 0$ at $h = (\text{R})$

	$\mu^2_{21}(\text{A}|h) = 1$ $0 \le b_2 \le 1$	$\mu^2_{21}(\text{A}|h) = 0$ $b_2 \le 1/2$	$\mu^2_{21}(\text{A}|h) = 0$ $b_2 \ge 1/2$
$z = (\text{R,A})$	0	0	0
$z = (\text{R,B})$	$2 + 5(b_2 - f(4))$	$2 + 5b_2$	$2 + 5(b_2 - f(2))$

player 2's optimal strategy in the equilibrium. In fact, we get

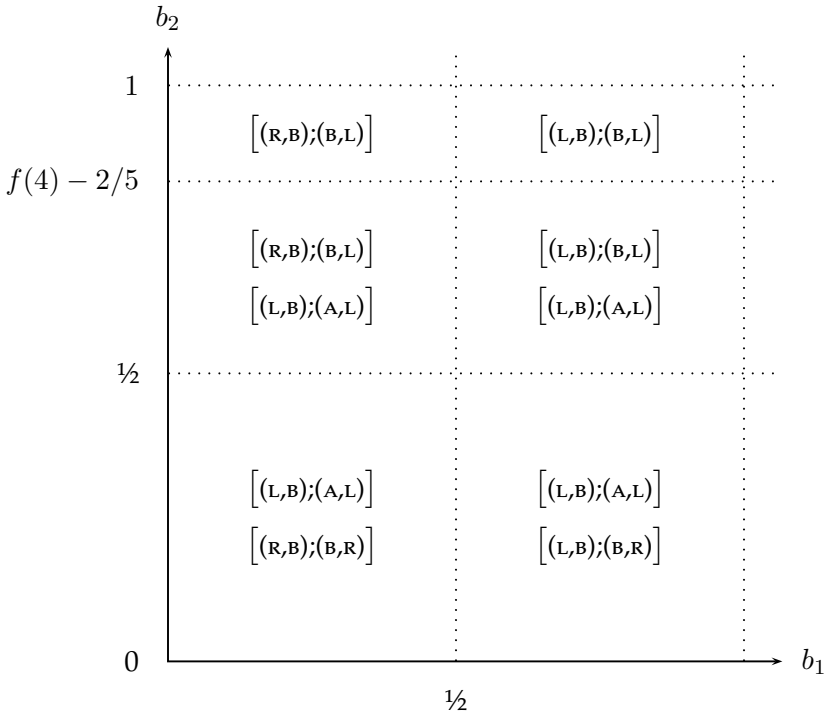
$$s_2^* = \begin{cases} \text{A} & \text{if } 0 \le b_2 \le f(4) - 2/5, \\ \text{B} & \text{if } b_2 \le 1/2, \\ \text{B} & \text{if } b_2 \ge \max\{1/2, f(2) - 2/5\}. \end{cases}$$

Finally, let us determine player 1's equilibrium behavior. Observe from Table 1a that player 1 blames player 2 ($\delta_{12} = 5$) only under the beliefs that player 2 plays A, and player 1 plays R. In that case, his utility from playing L is $4 + 4(b_1 - f(5))$, and from playing R is 0. Therefore, for any $b_1 \ge 0$ he prefers to play L. Clearly, this is not consistent with the underlying beliefs that player 1 plays R. In all other cases, player 1 does not blame player 2 and he prefers L to R if and only if $4 + 4b_1 \ge 5 + 2b_1$; i.e., $b_1 \ge 1/2$.

Since this is a game of complete information, players' types are common knowledge. Hence, the equilibrium behavior of players will potentially be different depending on players' types. In fact, Figure 2 depicts different regions of the b_1, b_2 space, each determining a different equilibrium. The horizontal axis is b_1 and the vertical axis is b_2. We list the equilibria as $\Big[\big(s_1(\varnothing), s_{\langle 12\rangle}(\text{R})\big); \big(s_2(\text{R}), s_{\langle 21\rangle}(\varnothing)\big)\Big]$. We omit the beliefs since they are consistent with strategies in the equilibrium.

Note that there is multiplicity of equilibria in some regions. For example in the region where $0 \le b_1 \le 1/2$ and $1/2 \le b_2 \le f(4) - 2/5$, there are two equilibria, namely $\big[(\text{R,B});(\text{B,L})\big]$ and $\big[(\text{L,B});(\text{A,L})\big]$. In the former equilibria, player 2 is nice enough that he would play L if he were in player 1's position. Therefore, he blames player 1 who plays R by $\delta_{21} = 2$. However, player 2's blame is not strong enough to make him play A. As a result, player 1 who is relatively more selfish ($b_1 < 1/2$) plays R, and player 2 plays B. The latter equilibria is completely different. When player 2's strategy is A, he blames player 1 even more ($\delta_{21} = 4$) if player 1 plays R. This blame is strong enough that player 2 indeed prefers to play A in case player 1 plays R. In order to avoid this, player 1 plays L.

Figure 2: Characterization of Sequential Equilibria of the Game in Figure 1.



For each region, the type of equilibria that appear in the region are labeled as $\left[\left(s_1(\varnothing), s_{\langle 12\rangle}(\text{R})\right); \left(s_2(\text{R}), s_{\langle 21\rangle}(\varnothing)\right)\right]$. This figure represents the case where $f(4) - 2/5 > 1/2 > f(2) - 2/5$.

3 Experiments

In what follows, we will discuss the two experiments conducted to test our notion of blame. The first is a modified version of a dictator game with punishment, and the second is an experiment that investigates a public goods game with punishment, performed by CKS. Since the data produced by the CKS are richer than our dictator game data, we will be able to perform a more elaborate empirical analysis including a structural estimation. Taken together, these experiments offer significant support for our notion of blame.

Note that these experiments are not direct tests of our equilibrium concept but rather aim to test reduced form versions of our model concentrating only on the actions taken by our subjects in all positions and their punishment behavior. These actions are all the data we need to test the premises of our theory. In fact, our dictator-game experiment is not a game but rather an exercise performed to test whether or not the preferences of subjects are consistent with the notion of blame.

Put differently, our theory has two components: first one describes the preferences of players when involved in the counterfactual thought experiment concerning blame, and

the second one delineates the strategic aspects of interaction, and its associated equilibrium. In both experiments we discuss here, we are exclusively interested in whether or not our subjects perform the blame thought experiment, and put themselves in others' situations. In each experiment, the action a subject may take can be unambiguously ranked in terms of kindness. For instance, giving more in the Dictator game is kinder than giving less. (This is also the case for contributions to the public goods in the second experiment.) In this regard, the blame thought experiment is straightforward since in stage 1 all subjects are, by definition, in the position of each other. Then, in stage 2, according to the definition of blame, they are expected to compare how much they gave in stage 1 to how much they are offered by the subject they are matched with. If a subject gave more in stage 1, he blames the subject he is matched with.

3.1 Dictator Experiment

The aim of our dictator experiment was to test the basic elements of our theory of blame focusing on subjects' preferences and motivations for punishment. In order to do this, we designed an experiment that did not involve any strategic interaction. Rather, the experiment was a direct test of preferences involving blame. This design had the obvious advantage of holding strategic considerations absent for our subjects, and not having them confound observed behavior. Later in this section, we also present the analysis of an experiment—which involves a strategic situation—run by CKS.

For our experiment, which was conducted at the experimental lab of the Center for Experimental Social Science, 120 subjects were recruited from the undergraduate population of New York University. They engaged in one round of the experiment that lasted about 25 minutes. The show up fee was $8, and subjects earned on average $16.72.

The task that our subjects faced was extremely simple. It was composed of two stages. Subjects were not informed about the content of stage 2 before they completed stage 1. However, at the beginning of the session, they were informed that there would be a two stages.

In stage 1, the subjects played a dictator game. That is, they were asked to split 10 tokens (convertible to dollars at the rate of 1 token = 1$) between themselves and an anonymous subject in the room. After all the subjects made their choices they were randomly divided into two equally sized groups called Senders and Receivers, and then randomly matched in pairs. If a subject was chosen to be a Sender, his payoff was equal to what he decided to keep for himself of the 10 tokens. If he was chosen to be a Receiver, his payoff was equal to the amount given to him by the Sender whom he was matched with. Subjects were not told their payoff from stage 1 until after both stages of the experiment were completed.

After stage 1 was completed, subjects moved on to stage 2, where they were randomly matched with another subject in the lab. After they were matched, they were offered, as a Receiver, the amount that the subject whom they were matched with sent as a Sender in stage 1. In other words, if a subject was matched with a person who gave 3 tokens in stage 1, he was offered 3 tokens in the stage 2 as his payoff while the other subject received 7. Subjects did not have the option of rejecting proposals; however, they could, if they wished, punish the Sender that they were matched with by reducing his payoff by 1 token at no cost to themselves. In order to elicit their response, we used the strategy method: rather than having the subjects punish directly, we asked them to set a cutoff before seeing the offer.[6] If the offer was equal to or less than the cutoff, 1 token would be removed from the other subject's payoff. For example, say that a subject chose a cutoff of 4. If the subject was matched with a Sender who offered 3 tokens in stage 1, and kept 7 for himself, then the computer would reduce the Sender's payoff by 1 token from 7 to 6. If the Sender gave 5 tokens and kept 5 tokens, his tokens would not be reduced, and his payoff would stay at 5.

The stage 2 payoff was determined as follows. After subjects made their choices in stage 2, the computer randomly determined whether they were a Sender or a Receiver. If a subject was designated to be a Receiver, the subject whom he was matched with was designated to be a Sender, and vice versa. A Sender's payoff was equal to what he decided to keep for himself in stage 1 minus 1-token reduction if there was a punishment.

The total payoff for the subjects in the experiment was equal to the sum of earnings in stage 1 and 2 plus the show up fee.

The data of interest generated by the experiment are the stage 1 offers and the stage 2 cutoffs since those are the basic ingredients both for our theory and other theories.

In the next section, we will determine the predictions of four theories that could be applied in this experimental environment. We will then analyze the data in the light of these predictions.

3.1.1 Theoretical Predictions

If theories are to be of interest, they must generate predictions that are distinctly different from those of other competing theories, and be testable using a simple and transparent experiment. The experiment outlined above, we believe, is an excellent testing ground for our theory of blame because it is straightforward and makes considerably dif-

[6] While some investigators have suggested that behavior may be different when subject behavior is elicited using the strategy method, Brandts and Charness [6], in a review paper, found limited support for this conjecture especially when the number of contingencies elicited in the strategy is small. A similar point is made in Brandts and Charness [5] where they compared the "hot" and "cold" elicitation methods and found no significant difference.

ferent predictions from all of the other leading theories that can be used to explain the data it generates—i.e., the F&S's Inequity Aversion Model, Rabin's Theory of Reciprocity, Levine's Theory of Interdependent Preferences, and our Theory of Blame.

Our goal in this section is to describe the behavior predicted by these theories for the particular experiment we have just discussed. Throughout this section, we assume that a subject assigns a utility $v(x)$ to a material payoff x. For expositional ease, let us assume that v is increasing, continuous, and concave.

Since our experiment asks subjects to first divide 10 experimental tokens, and then to decide on a punishment to be directed at those whose division they think is worthy of punishment, we will investigate each contending theory to determine their predictions on this division and punishment behavior. More precisely, we will discuss a subject's optimal allocation between himself, $10 - x^*$, and another subject, x^*, and also determine whether the same individual is willing to punish a subject who gives him y when he is in the Receiver position.

Blame

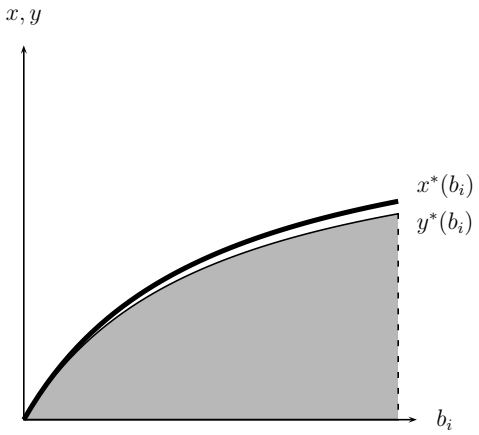
Let us begin with the predictions of our theory. Note that since the Sender's problem in stage 1 is simply choosing an optimal allocation, a subject i chooses $x \in [0, 10]$ to maximize $v(10-x)+\big[b_i - f\big(\delta_{ij}(\mu_i|h)\big)\big]x$, where f is a non-negative and increasing function such that $f(0) = 0$ for all $\delta_{ij} \leq 0$. In stage 1, since subjects make a simple allocation decision, there is no room for them to blame anyone. Hence, in stage 1, we assume that $\delta_{ij}(\mu_i|\varnothing) = 0$ for any subject i. It is therefore straightforward to see that the optimal amount x^* a subject will send is a non-decreasing function of b_i.

The amount $x^*(b_i)$ is what subject i with b_i will optimally offer to an anonymous subject when he is asked to allocate a material payoff of 10 as a Sender.

Now suppose that the subject is in the Receiver's position in stage 2 and he is offered an amount y. Based on the premise of our theory, he evaluates this offer based on what he would have offered as a Sender. Since, he was in the Sender's position in stage 1, we assume that he updates his preferences by incorporating blame $\delta_{ij}(\mu_i|h) = x^*(b_i) - y$. As a result, the Receiver may find it optimal to punish the sender for the offer y if $y < y^*(b_i) := x^*(b_i) - f^{-1}(b_i)$.

Figure 3 summarizes these observations. The bold line indicates the optimal offer of the Sender. Since the choice of f is arbitrary the punishment cutoff $y^*(b_i)$ is below $x^*(b_i)$ for any b_i. The important feature is that both the offer, $x^*(b_i)$, and the punishment cutoff, $y^*(b_i)$, are increasing in a subject's b_i. Moreover, $x^*(b_i) \geq y^*(b_i)$ for any b_i.

Figure 3: Prediction of the Theory of Blame



The optimal offer of a subject with b_i is $x^*(b_i)$. The subject can punish any offer below $y^*(b_i)$, i.e. shaded region.

Inequity Aversion

The view that individuals can exhibit inequity aversion is studied by F&S.[7] According to this theory, although a person likes more material payoff, he gets disutility from being far off from others. Based on F&S's formulation of such preferences, in our problem, a subject chooses an offer x to maximize

$$v(10-x) - \alpha \max\left\{2x-10, 0\right\} - \beta \max\left\{10-2x, 0\right\}$$

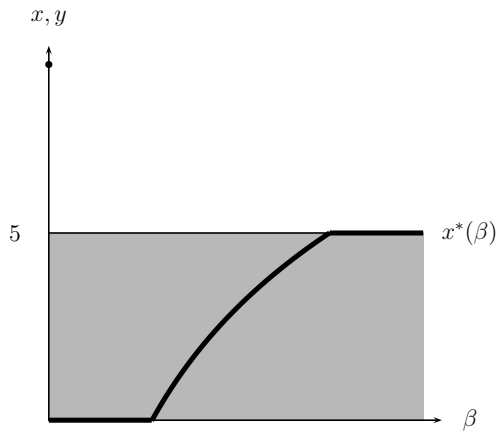
where $\alpha > \beta \geq 0$, and v is a concave and continuous function.[8] Note that the larger a subject's β parameter, the nicer he is when acting as a Sender since he dislikes having more than the Receiver. In that sense, it is directly comparable to our b_i. Also note that the α parameter is relevant when the subject receives less than 5 (allocates more than 5 to the Receiver) while the β parameter is relevant when the subject receives more than 5 (allocates less than 5 to the Receiver.)

Since it is never optimal for a subject to allocate more than 5 to the Receiver (given $\alpha > 0$), it will be the β parameter that will determine a subject's allocation when acting as a dictator. Hence, $x^*(\beta)$, which is the optimal amount a subject with parameter β will

[7] Bolton and Ockenfels [4] study the same idea with a different formulation of preferences. Since the predictions of the two papers are the same, we do not provide a separate discussion.

[8] In the original formulation of F&S, the utility derived from the material payoff is $10 - x$. Therefore, depending on the value of β, Sender will offer either 0 or 5. The concavity of the v function makes the giving behavior smoother and provide a better match to data.

Figure 4: Prediction of the Theory of Inequity Aversion



The optimal offer of a subject with β is $x^*(\beta)$. Any subject can punish any offer below 5, i.e. shaded region.

offer, increases in β. However, it never exceeds the most equitable allocation 5, because beyond 5 the subject bears a disutility from inequity aversion, and receives less material payoff. The bold line in Figure 4 illustrates the optimal offer.

Now, we turn to the question of when a subject punishes a Sender in stage 2. Let y be the offer that a Receiver gets. Any offer $y < 5$ creates a disutility since $\alpha > 0$. Therefore, the Receiver will find it optimal to punish any offer $y < 5$ since punishment reduces the existing inequality at zero cost. Also note that a Receiver who offered $x^* > 0$ when he is in the Sender position must necessarily have $\beta > 0$. Therefore, for any offer $y \geq 5$, since the punishment further increases the inequality, the Receiver does not find it optimal to punish the Sender. Only when $x^* = 0$ it is possible that $\beta = 0$, and hence, the subject is indifferent between punishing any offer y.

The Figure 4 depicts this behavior. The bold line indicates the optimal offer $x^*(\beta)$ of a subject as a Sender, and the gray area indicates the offers that will be punished. Note that the behavior under F&S differs qualitatively from that of our blame theory because, while in our theory the punishment cutoff is an increasing function of a subject's b_i, under F&S, the punishment threshold is a constant of 5, and independent of a subject's α. This provides a very strong prediction for the F&S's theory, which is that all offers below 5 will be punished. This prediction is distinctly different than the prediction of the theory of blame.

Kindness

Most reciprocity theories rely on an exogenous norm of kindness in order to determine how kind people are.[9] For instance, in the theory of Rabin, "[...] players have a shared notion of kindness and fairness and that they apply these standards symmetrically." The main postulate of this theory is that people return kindness with kind acts, and unkindness with mean acts. Precisely, Rabin defines kindness with reference to a predetermined allocation (e.g. equitable allocation) on the Pareto frontier of possible payoffs. If a player believes that his opponent's strategy leads to a payoff that is less than this predetermined allocation, then the player finds his opponent's act unkind. Although Rabin shows that the results are valid for a large class of such reference points, they are nevertheless exogenously determined.

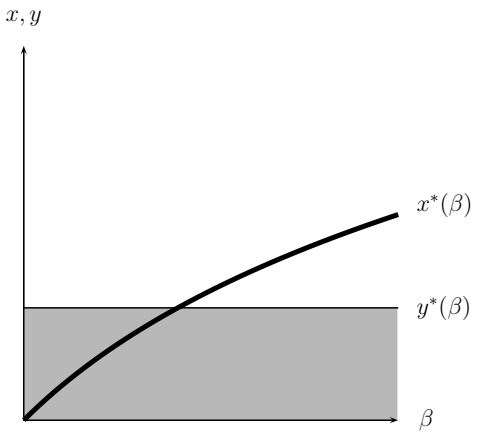
In Rabin's model, there is no room for reciprocity unless there is strategic interaction among players. Therefore, according to Rabin's original formulation subjects offer zero in stage 1 of our experiment. However, one can reasonably generalize this approach, and capture how nice a subject is towards an anonymous person in the absence of reciprocity by including a parameter, say β, in the utility function. Given such a generalization, we would expect a player to give more in stage 1 as β increases as in Figure 5. Note, however, that according to Rabin, the behavior of a subject in stage 2 of our experiment is independent of his offer in stage 1. This is the case because in stage 2 the kindness of a Dictator is judged strictly against the "split-the-difference" norm imposed by Rabin. In other words, it is independent of the behavior of a subject in stage 1. If the Dictator gives more than 5 he is kind, and otherwise he is unkind. In fact, no matter which norm is used in Rabin's model, stage 2 punishments are not correlated with stage 1 offers. This logic is similar to that of the F&S, where offers are determined by a player's β parameter, while punishments are governed by α. More precisely, as long as α is greater than zero, the punishment cutoff is always 5 and independent of stage 1 behavior. In Rabin's model, offers may be determined by a player's altruism while punishments by the kindness norm players adhere to.

In stage 2 of the experiment, we assume that the punishment behavior is determined by how kind a player thinks the Sender's offer is in stage 1. Precisely, the offer is unkind if it is below the predetermined norm, and kind otherwise. Therefore, the theory predicts that a Sender is punished if the offer he makes is below the norm, which is arbitrary.

Figure 5 summarizes our discussion.

[9] See [3, 11–14] for leading examples.

Figure 5: Prediction of the Theory of Rabin



The optimal offer of a subject is $x^*(\beta)$. Any subject can punish an offer below $y^*(\beta)$, i.e. the shaded region.

Interdependent Preferences

Levine takes a novel approach in formulating altruism and reciprocity by using orthodox game theoretical tools.[10] In particular, in order to analyze experimental data, Levine models the underlying game as a Bayesian game, where the types determine how altruistic or kind a subject is towards another subject. A simplified version of the preferences are as follows.

$$v(10-x)+\frac{a_i+\lambda a_j}{1+\lambda}x.$$

That is, each subject i has a type $a_i \in (-1,1)$, which determines the weight they assign to the other player's material payoff. Players are also sensitive to the type of who they are playing against. In particular, the utility function posits that the weight is higher if the opponent is nicer; i.e. higher a_j. Also, the intensity of the sensitivity is captured by a parameter λ. Clearly, if $\lambda = 0$, player i is not affected by whom he plays against. As $\lambda > 0$ increases, the player becomes more sensitive. Therefore, for $\lambda > 0$, player i is nicer against a nice player.

In our experimental setup, in stage 1, since a Sender does not know whom he plays with, he chooses x to maximize

$$v(10-x)+\int \frac{a_i+\lambda a_j}{1+\lambda}dF(a_j)x$$

[10]For a more general approach we refer the reader to Gul and Pesendorfer [18].

where F is the distribution of types.

It is easy to determine the optimal offer $x^*(a_i)$ of player i. Since the weight is an increasing function of a_i, the optimal offer is non-decreasing in a_i. Recall that in our experiment we keep stages 1 and 2 strategically independent. Therefore, the type of a player is directly revealed through his offer y, which he chose as a Sender in stage 1.

In stage 2, whether or not a Receiver punishes a Sender depends on his and the Sender's type. In fact, if $a_i < -\lambda a_j$ the Receiver punishes. This in fact suggests that there is an inverse relation between what an agent offers and his punishment threshold. The intuition is simple. According to Levine's formulation, how much a player cares about the material payoff of another player increases both in the other player's kindness (type) as well as his own kindness. Hence, a spiteful player may possibly punish another player despite his kind offer. Similarly, a very nice agent does not punish anybody except maybe a very spiteful people.

Finally, note that Levine's theory differs from ours: in the theory of blame if a player is nicer he expects more from others because he is judging them by his own standards; for Levine, however, nicer people are more tolerant to the spiteful behavior of others because, in some sense, the two types are averaged in determining how much a subject cares about his opponent.

As a result, Levine's theory predicts punishment behavior that is opposite of ours. In our theory, punishment cutoffs are increasing in how nice subjects are, and nice people demand nice behavior. Yet in Levine's theory, punishment cutoffs are decreasing in how nice a subject is.

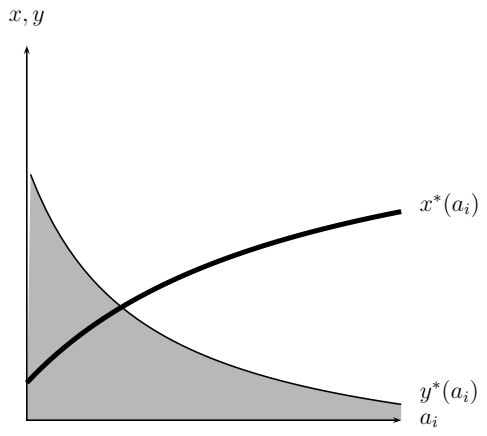
Figure 6 illustrates the discussion.

3.1.2 Experimental Results

In light of the predictions discussed above, we will first present the results of our experiment by concentrating on the relationship between the offers made in stage 1 and the cutoffs stated in stage 2. It is important to note that, the theories we discussed made stark and qualitatively different predictions. While the F&S (and other inequality averse models like Bolton and Ockenfels [4]) predict that cutoffs will be 5 no matter what the offer a subject made in stage 1, the theory of blame predicts that cutoffs increase in the offers made. On the other hand, Levine suggests the opposite: cutoffs are a decreasing function of offers. Therefore, the only theory suggesting a positive relationship between the offers and the cutoffs is the theory of blame, and below we will also demonstrate that our data is consistent with such a positive relationship.

It is important to note that we are not interested in having a horse race among theories. We believe that no single theory can explain the varieties of behaviors that people exhibit,

Figure 6: Prediction of the Theory of Levine



The optimal offer of a subject with a_i is $x^*(a_i)$. Any subject can punish any offer below $y^*(a_i)$, i.e. shaded region.

either in experiments, or in life. Therefore, numerous theories are needed to explain observed behavior. What is of potential interest, however, is how the population of people is distributed across these theories. To this end, we present another calculation where we look at the conformity of four theories to the data for each offer made in stage 1. What we demonstrate is that while our theory of blame does a good job at organizing the data, it is not the only theory with explanatory power.

First, we observe that 36 out of a total of 120 subjects offered 0 in stage 1. This is not surprising given the results of typical dictator games in the literature. Of these 36 subjects, 12 also set 0 as their cutoff, while 24 set a positive cutoff. Such behavior is consistent with subjects being either selfish or spiteful. However, trying to test various theories using the data generated by the subjects who offered 0 is not very satisfactory. For example, if we allow subjects to be spiteful, and have a non-positive weight, $b_i \leq 0$, on the material payoff of others, our theory can explain all those observations. Restricting subjects to care about the payoff of others, $b_i > 0$, implies that, since 0 is at the bottom of the support of offers and cutoffs, our theory can be supported only if those subjects who offered 0 in stage 1 also set cutoffs at 0 in stage 2 (which, as mentioned above, 12 out of 36 subjects did.) Levine's theory cannot fail because it predicts an inverse relation between the offers and the cutoffs; hence an offer of 0 is not in conflict with a non-negative cutoff. (Setting 0 in both stages is a boundary case of that theory, yet, when allowed, makes any observation consistent with it.) F&S is satisfied when a subject offers 0 in stage 1, only if he then sets a cutoff of 5 in stage 2 (of which 11 subjects did.) Such a subject would prove himself to be selfish in stage 1 ($\beta = 0$), and envious in stage 2 ($\alpha > 0$.) Rabin's theory, like F&S, is also consistent with

offering 0 in stage 1, and setting a cutoff of 5 in stage 2 given his formulation of kindness, and assuming that the only kind offer in a dictator game is 5 (the midpoint of the Pareto frontier.) However, if one substitutes a different norm in his kindness definition, then any positive cutoff—which we will denote by κ—can satisfy his theory for subjects who offer 0 in stage 1.

Given this discussion, a better place to look for confirmation of the theory of blame is in the interior of the offer set where people make positive offers. Not only does our theory separate here more naturally but also it makes most sense in a situation where people are not completely selfish as those people are the type who would care enough to put themselves in the place of others. On the other hand, selfish or spiteful types are less likely to engage in a thought experiment involving blame.

As stated above, the most direct confirmation of our theory will be derived from establishing that the cutoffs that subjects set are increasing functions of their stage 1 offers. Since only our theory posits an increasing relationship between these two variables, if such a relationship is established, it will provide support exclusively for our theory.

In order to investigate this relationship, we use a simple OLS regression, where we regress the cutoff set by our subjects as a function of the offers they made in stage 1. The results of this regression are presented in Table 3 below.

Table 3: Relationship Between Cutoffs and Offers

Positive offers

	Coefficient	$\|t\|$	$p > \|t\|$
offer	.284	2.04	.045
	(.139)		
constant	3.080	6.180	.000
	(.499)		
	$N = 83$, Adj $R^2 = 0.370$		

All offers

	Coefficient	$\|t\|$	$p > \|t\|$
offer	.200	1.76	.080
	(.113)		
constant	3.416	10.13	.000
	(.337)		
	$N = 120$, Adj $R^2 = 0.018$		

Standard errors in parentheses.

When this regression is run on the full data set (including both 0 and positive offers),

we find that the offer variable is positive and marginally significant ($p < .080$.) When we remove those subjects who made 0 offers, however, the coefficient is significant ($p < .045$) and positive. The fact that the relationship between offers and cutoffs is positive lends support to our theory exclusively, since ours is the only one that posits a positive relationship.

These regressions, while supportive of our theory, do not capture the subtleties of the predictions made by the four theories both when subjects make zero or positive offers in stage 1. In order to give a quick look at how these predictions differ when we move from the set of subjects who chose 0 in stage 1 to those who made positive offers, we present Table 4.

One quick point about Table 4 is in order. First, note that the predictions of theory of blame, Rabin, and F&S are rather straightforward. For example, our theory of blame predicts that when subjects make positive offers, they set cutoffs below their offer. Therefore, a subject's behavior complies with the theory of blame, if his cutoff is not higher than his offer. For the Rabin and F&S theories, we categorize subjects as satisfying those theories if in stage 2, they set a cutoff of κ, and 5, respectively. In Rabin's case, we perform the test for each $\kappa \in \{1, 2, 3, 4, 5\}$ and choose the κ that yields the highest compliance.

Table 4: Predictions for Zero and Positive Offers

	Theoretical Prediction: Stage 1 offer $= 0$			Theoretical Prediction: Stage 1 offer > 0	
	Offer (x)	Cutoff (y)		Offer (x)	Cutoff (y)
Blame	$x = 0$	$y = 0$, if $b_i \geq 0$ $y > 0$, if $b_i < 0$	Blame	$x > 0$	$y \leq x$, $dy/dx \geq 0$
F&S	$x = 0$	$y = 5$	F&S	$x > 0$	$y = 5$
Levine	$x = 0$	$y \geq 0$	Levine	$x > 0$	$dy/dx \leq 0$
Rabin	$x = 0$	$y = \kappa$	Rabin	$x > 0$	$y = \kappa$

Deriving the predictions of Levine's theory is more complicated for two reasons. First, his theory only suggests that the offers and cutoffs are inversely related. However, since the subjects make a single decision, we can not test Levine's theory at the individual level. (Our regression results above already indicate that it lacks support on the aggregate level.) Second, Levine's theory is embedded in a game of incomplete information, where the cutoffs that subjects set depend on their inference about the type of subjects they are playing against, given that subject's offer. Hence, in order to know if subjects are acting consistently with his theory, we need to derive the function relating a subject's offer to his type, and then investigate whether his decision is a best response to this information.

In order to do this, we perform a simple calibration exercise, which we will explain

next. Suppose that a subject i's utility function is, $u(x) = ln(10 - x_i) + \frac{a_i + \lambda a_j}{1+\lambda}$.[11] In stage 1, subject i does not know the type of the subject he plays against when he makes an offer; therefore he responds to the mean type $\bar{a}_{-i}$.[12] Given λ, subject i chooses x_i that maximizes his utility. Although the first order condition determines a_i as a function of $\bar{a}$ for those who make a positive offer, we can pin down a similar relationship for those who offer zero. To do that, we fix an $\bar{a}$ and compute a_i for those subjects who make positive offers. For those who offer 0, we use their stage 2 punishment behavior to determine their type.[13] This exercise determines a new $\bar{a}$, which may not be consistent with the $\bar{a}$ assumed at the start of the exercise. Hence, in order to provide consistency, we iterate on $\bar{a}$ and λ until we converge to a fixed point.

The upshot of this exercise is that subjects who make positive offers in stage 1 imply that they have positive types ($a_i > 0$); hence, will only punish negative types. However, since the only types that are identified as negative types are those who offered 0 in stage 1, our calibration exercise implies that the subjects making a positive offer in stage 1 must set a zero cutoff in stage 2. Therefore, Levine's theory would predict that any subject offering positive amounts in stage 1 will set a zero cutoff in stage 2.

Figure 7: Compliance with Theories

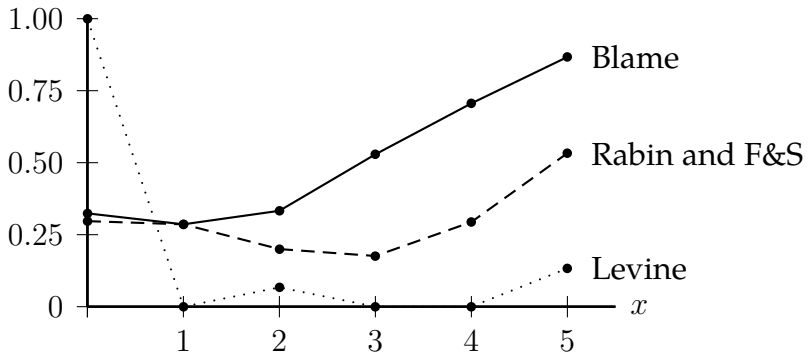


Figure 7 presents fraction of subjects whose behavior is consistent with the Rabin, F&S, Levine, and the theory of blame for each x sent at stage 1, based on our calculations.[14] For Rabin's theory, as we explained above, we compute the fraction of subjects who adhere to

[11] Our exercise is robust to any concave utility function for a subject's material payoff. Using the log function is done for reasons of simplicity.

[12] We denote the mean of the sample excluding player i by $\bar{a}_{-i}$

[13] For $x_i > 0$ the first order condition is $\frac{a_i + \lambda \bar{a}_{-i}}{1+\lambda} = \frac{1}{10-x_i}$ from which we can derive a subjects type as $a_i = \left[\frac{1+\lambda}{10-x_i} - \lambda \frac{n}{n-1}\bar{a}\right] \frac{n-1}{n-1-\lambda}$. When $x_i = 0$, $a_i < \left[\frac{1+\lambda}{10-x_i} - \lambda \frac{n}{n-1}\bar{a}\right] \frac{n-1}{n-1-\lambda}$. However, we can pin down a_i when we observe subject i's punishment behavior in stage 2 since he will punish if $a_i + \lambda a_j < 0$.

[14] We only include offers made up to 5 since there were too few observations for offers of 6 or more.

the theory for each κ, and report the fractions for κ that yields the highest compliance. In fact, we find that $\kappa = 5$ leads to the best adherence to Rabin's theory. Hence, we report it together with F&S.

Figure 7 is interesting because it indicates that as stage 1 offer increases, the consistency of our data with various theories changes. For example, Levine's theory is trivially satisfied when the offer is 0 no matter what the cutoff is. As a result, we see that 100% of the observations are consistent with them. For the theory of blame, when stage 1 offer is 0, we find that the consistency rate is 32.4%.[15]

For F&S, when the offer is 0 the cutoff must be 5 which is true for 29.7% of the observations. This is a curious result because while their stage 1 behavior indicates purely selfish motives, their stage 2 behavior indicates strict inequality aversion.

As offers are positive and increasing, we see an increase in the percentage of observations consistent with the theory of blame, a fall in the predictive powers of the Levine theory, and an increase in F&S and Rabin theories.

In summary, the data generated by our experiment lend support to the idea that the subjects used elements of blame in their punishment behavior. That is not to say that there is not support for other theories. As discussed, when the offers are positive, the predictive ability of the theories changes. It is important to note that our theory's predictive is consistently higher for subjects who make higher offers, whereas the predictive power of the alternative theories we investigate changes. Also, the fact that we are able to establish a positive relationship between offers and cutoffs tends to diminish support for F&S, and Rabin, both of whom suggest that there is no relationship between offers and cutoffs, and Levine who posits a zero cutoff for positive offers in our experiment.

3.2 Public Goods Experiment

3.2.1 Experimental Design

In order to provide more evidence supporting our theory of blame, we should be able to detect it in other experiments as long as they have two main features: *(a)* subjects are allowed to punish others (or reward them) and, *(b)* subjects are placed in identical symmetric situations so they know how they would behave if they were in another person's shoes.[16]

Such requirements are easily satisfied by public goods experiments with punishments where—as in Fehr and Gächter [15]—subjects are allowed to punish others. In such sym-

[15] For theory of blame, if we allow b_i's to be negative, in fact 100% of the subjects would be consistent when their offer was zero. In our computations, we restricted b_i to be non-negative.

[16] While Iriberri and Ray-Biel [21] demonstrate that revealed subject preferences can change when subjects are placed in an experiment with role uncertainty, we observe no such effects in our experiments.

metric games, a subject can engage in the blame thought experiment since he knows how much he contributed to the public goods, and can therefore compare his behavior to others, and see if he wants to blame and punish them.

However, such data must be used with caution, since most, if not all, of these experiments are embedded in a complete network, where all subjects can see the contributions of all others, and punish as many people as they wish. This network structure is not the best to use in assessing the motives for subject punishment, since it is steeped with free riding, and other coordination problems, which easily mask the motives for punishment. For example, suppose that subject i is in such a public goods game with three others, and he observes the contributions of all three. In addition, say that it is common knowledge that each person can observe and punish whomever he wants. Whether or not subject i punishes, then, becomes a function of his beliefs about whether or not the others in his group will punish, and he might easily decide to free ride on their punishments. Furthermore, who subject i punishes is also a complex coordination game since he obviously would want to punish the subject who has contributed the least, yet as such a person is more likely to be punished by others, subject i might decide to punish the second lowest, etc. The point is that subject i's punishment behavior may be a poor indicator of his preferences because it is confounded by these strategic considerations.

A better way to identify a subject's true preferences or motives for punishment would be to look at behavior in a different network where such coordination and free riding problems are absent. Such an exercise was done by CKS, where they look at four-player public goods games with punishments under a variety of network architectures. The best of these architectures for our purposes is called the *directed circle*, where subjects are arranged in a circle, in which Subject 1 can observe and punish Subject 2, Subject 2 can observe and punish Subject 3, Subject 3 can observe and punish Subject 4, and finally Subject 4 can observe and punish Subject 1. In this case, each person has the sole responsibility to observe and punish just one person; hence there is no free riding or coordination problems, and therefore, the motivation for punishing is clearer.

In their experiment, the subjects first play a standard Voluntary Contribution Game in two stages. In stage 1, they contribute and receive feedback regarding their payoffs, and the mean contribution of the others in their group. In stage 2, after observing their stage 1 information, they decide whether or not to punish, and by how much. Punishment points directed at another subject reduce the target's payoff by 10% for every one point used by the subject.

3.2.2 Results

There are a number of theories that can explain punishment in such public goods games. One theory, investigated by De Quervain et al. [9], is a norm-based theory that says people punish others if they violate a group contribution norm. This norm is set exogenously, possibly based on some commonly held theory of fairness. A similar notion of social norm is also mentioned in Charness and Rabin [8]. In Fehr and Gächter [15], the norm is that those who fail to contribute the mean amount or more are candidates for punishment. According to the theory of blame, however, the standard used to determine punishments is not an exogenously defined norm, but rather a personal and individual standard; based on how the person himself would behave in the situation being examined.

For the public goods game, the subjects who subscribe to the theory of blame, should punish only if the other subject contributed less than they did, and never punish those who contributed more, whether or not those people contributed more than the average of the group.

These different theories of punishment can easily be investigated using the CKS data, and running a simple random effects probit regression, where the probability of punishment is a function of the difference between a subject's contribution and that of the other subject whom he observes (the target), the difference between the target's contribution and the group's mean, and the mean itself. Let pc_i be the public goods contribution of subject i , pc_{-i} be the target's contribution, and m be the mean of the group's contribution in a given period. More formally the regression run is

$$\Pr(\text{punishment}) = \alpha + \beta_1 \Delta^{+}_{other} + \beta_2 \Delta^{-}_{other} + \beta_3 (pc_{-i} - m) + \beta_4 (m) + \epsilon_i,$$

where $\Delta^{+}_{other} := pc_i - pc_{-i}$ if $pc_i - pc_{-i} > 0$, and 0 otherwise; and $\Delta^{-}_{other} := pc_i - pc_{-i}$ if $pc_i - pc_{-i} < 0$, and 0 otherwise. If the theory of blame is responsible for punishment behavior, we would expect that coefficient β_1 to be positive and significant, while all other coefficients should be zero since all that should matter for punishment is whether or not a subject's contribution was more than the contribution of the target, which is captured by Δ^{+}_{other}. The regression results are presented in Table 5.

As we see in Table 5, consistent with our expectations, the only variable with a statistically significant coefficient is Δ^{+}_{other}. The positive coefficient means that a subject is more likely to be punished the further his contribution falls below that of the person who observes him. The relationship of the target's contribution to the mean is insignificant. The coefficient in front of the mean variable itself is negative but again insignificant. The negativity of the coefficient makes sense because, as the mean increases, being below it is less of a transgression.

Table 5: Punishment Behavior in CKS: Directed Circle

	Coefficient	Z	$p > \|Z\|$
Δ^{+}_{other}	.153	4.86	.000
	(.032)		
Δ^{-}_{other}	.046	1.45	.148
	(.032)		
$(pc_{-i} - m)$	-.038	-1.16	.246
	(.030)		
mean	-.014	-.44	.663
	(.033)		
constant	-1.422	-2.24	.025
	(.636)		

$N = 240$, Wald $\chi^2(4) = 45.92$, $\Pr > \chi^2 = .000$.
Robust z-statistics are reported in parentheses (clustering at the subject level.)

This simple regression lends support for the theory of blame, and disconfirms the idea that punishment behavior is determined by the observed behavior of a subject in relation to some exogenously determined norm. What matters is how much a subject's contribution differs from that of the person who is watching him—a personal norm.

There are more ways of looking at the CKS data which allows us to better explore the subtleties of our theory. For example, the decision to punish in a public goods game is a two-tiered decision, where first a subject decides whether or not to punish, and then, how much to punish. In order to take this into account we estimate a hurdle model. The hurdle regression consists of two steps. First, we fit a logit regression on whether or not to punish (the binary choice on the extensive margin) using the same set of explanatory variables mentioned above, (i.e., the difference between subject i's public contribution and that of his target allowing for non-linear effects for positive and negative differences, the difference between the target's contribution and the group's mean, and the mean itself.) Next, we fit a Poisson regression of the punishment level using the same set of explanatory variables

This hurdle model is estimated via maximum likelihood method to jointly estimate the two probabilities described above on the CKS data both for the directed circle and the complete networks. We do this in order to emphasize our point that, when testing theories of punishment for public goods games instead of previously used complete networks, the directed circles are the most relevant as issues of coordination and free riding on the punishments of others are absent. The results of this estimation are presented in Table 6.

As we established in the previous exercise, the key driver of both the decision to punish and its magnitude is the difference between the punisher and the target's contribution

Table 6: Hurdle Model Estimation

	(1)	(2)	(3)	(4)
	Directed Circle		Complete Network	
Punishment Margin	Extensive	Intensive	Extensive	Intensive
$(pc_i - pc_{-i})^+$	$.212^{***}$	$.034^{*}$	$.054^{***}$	$.001$
	(.040)	(.018)	(.020)	(.010)
$(pc_i - pc_{-i})^-$	$-.007$	$-.041$	$.160^{***}$	$-.049^{***}$
	(.033)	(.027)	(.019)	(.018)
$(pc_i - \text{mean})$	.040	$-.019$	$.128^{***}$	$-.008$
	(.040)	(.019)	(.021)	(.014)
mean	.004	$-.010$	$-.018$	$-.076^{***}$
	(.043)	(.019)	(.018)	(.013)
constant	-1.885^{***}	1.607^{***}	$-.132$	2.055^{***}
	(.762)	(.427)	(.281)	(.196)
Observations	238	238	716	716

Robust standard errors in parentheses.
$^{***}\ p < .01,\ ^{**}\ p < .05,\ ^{*}\ p < .10,$

levels (see columns (1) and (2) in Table 6.) Note, however, that this difference only matters when it is positive or when the target contributes less than the punisher. This is consistent with our theory since blame only exists when this difference is positive. Again, we observe that any variable referring to the mean contribution of the group is not significant in explaining neither the decision to punish nor its magnitude.

These results change when we move to the complete network (columns (3) and (4)), where subjects can punish any of the three other subjects in their group. Here we see that even though a positive difference in the contribution level of the punisher and the target is a significant determinant of the decision to punish or not, it is not significant for the magnitude of the punishment. More interestingly, however, a negative difference between the contribution of the punisher and the target is significant both for the punishment decision and its magnitude. This means that in the complete networks there seems to be a significant amount of what has been called "anti-social punishment" as people are punishing those who contribute more than they do.

Such anti-social punishment has been a source of great interest (see Herrmann, Thöni and Gächter [20], Thöni [28], Monin [24] amongst others), and many explanations have been offered. Fehr and Gächter [15] list random error, improvement of the relative position (status preferences), and revenge as possible explanations. Thöni [28] explains anti-social punishments by assuming that subjects are inequality averse. However, it is important to note that, when tested using a complete network public goods game, the question of

anti-social punishment is confounded by matters of preference, i.e., inequality aversion, do-gooder degradation, revenge, as well as strategic issues of coordinated punishment and free riding. The directed circle design eliminates most of these strategic issues as each subject is only responsible for punishing only one other person, and cannot even see the contributions of others. In other words, the directed circle allows for isolating the pure preference for punishment, which cannot be done using a complete network.

For example, take the revenge motive. In a complete network, when a low-contributing subject is punished and is able to observe the contributions of all the subjects, he might very well assume that he is punished by a high-contributing subject. In the next round, he may seek revenge and punish that subject. However, in the directed circle, there is no room for revenge because a subject can neither see the contribution of the person who punished him nor is he able to punish.

Also, note that only in the complete network, variables that are connected to the mean contribution of the group are significant. This is a direct artifact of the feature of this network which allows each subject to see the contributions of all the subjects, and as a result, punishers are more likely to punish those who are the most egregious free riders. What is missing in this design is the counterfactual of whether or not these subjects would punish others whose contribution were above the mean but below theirs. The directed circle is well suited for answering that question.

Finally, we use the CKS data to do a more structural estimation exercise. The rationale behind this exercise is derived from what determines punishment in our model. According to our model, in a public goods game punishment is a combination of the weight that subjects place on the material payoffs of others, and on how sensitive they are to blame. The subject who place more weight on the other's payoff, and who is not likely to be bothered when he is exploited, is unlikely to punish. By using the CKS data can we can identify the interaction between these two effects.

Let us look at the problem faced by our subjects in stage 2 of the experiment after they have made their contributions in stage 1. In stage 2, they know the amount they contributed in stage 1, and depending on the network they play in, they are able to see either the contributions of all the other three subjects (complete network) or that of the one person they are monitoring (directed circle). Assume that the utility of a subject in stage 2 is given as follows (for the complete network):

$$(B - g_i + \gamma G - c\sum_{j \neq i} k_j) - e^{-\sum_{j \neq i} \beta_i (B - g_j + \gamma G - k_j)}$$

where B (equal to 25 in the experiment) is the endowment of subject i, g_i is subject i's contribution to the public goods, γ is the marginal per capita return to investment in the

public goods (0.4 in the experiment), k_j is subject i's punishment of j, $c \in [0, 1)$ is the cost of punishment, ($c = 0.5$ in the experiment), and $G = \sum_i g_i$ is the total contribution of subjects to the public goods in stage 1. Finally, the blame-adjusted altruism parameter is,

$$\beta_i = b_i - \max\left\{\left[\frac{\pi_{ii} - \pi_{ij}}{\pi^{\max} - \pi^{\min}}\right]^{\alpha}, 0\right\},$$

where b_i is subject i's specific component that captures his degree of altruism towards the opponent, which we assume follows a normal distribution $b_i \sim N(\mu, \sigma^2)$. Inside the blame term, π_{ii} denotes what i's payoff (before punishment) would have been if j were to make a contribution in stage 1 equal to what i made, instead of the contribution j actually made, while π_{ij} is i's payoff (again before punishment) using j's actual stage 1 contribution. π^{max} and π^{min} are the maximum and minimum possible payoffs for a given player—in this case the difference in i's payoff between contributing nothing and contributing everything to the public goods given what the other $n-1$ subjects have contributed.[17] The denominator thus serves as a normalization, and will always be $(1-\gamma)B = 15$.[18] Finally, α captures the measure of subjects' blame-sensitivity. The objective of the subject in stage 2 is to choose a vector of punishments, k_j's, to maximize his utility.

The first term of the utility function captures the stage 1 payoff of subject net of the cost of punishments he makes, while the second term is the sum of the payoffs of others (net of the punishments he imposes on them) weighted by his blame adjusted altruism term.

The utility function above is defined for the complete network. In a directed circle, player i can only monitor one specific player j when making his punishment decision. Utility is therefore,

$$(B - g_i + \gamma G - ck_{j^*}) - e^{-\beta_i(B - g_j + \gamma G - k_j)},$$

where it is assumed that player i monitors player j^* while player j monitors player i.

By the first-order conditions, we have the optimal punishment decisions in the complete network (a very similar condition holds for the directed circle):

$$k_j = \frac{\ln(-c\beta_i^{-1})}{\beta_i} + 3B - \sum_{j \neq i} g_j - 3\gamma G - \sum_{l \neq i,j} k_l.$$

We develop a Method of Simulated Moments (MSM) estimator for the structural estimation. The parameters μ, σ, α are: The mean and standard deviation of the b_i distribution, and the power parameter inside the blame term. The aim of the estimation is to minimize

[17] This is independent of what the other subjects have contributed.

[18] Note that this normalization determines the units by which we measure the sensitivity to blame parameter, α.

a distance measure between model-implied moments and their empirical counterparts.

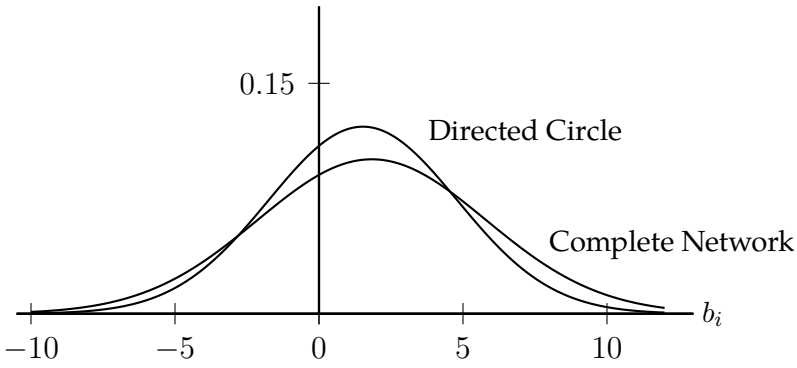
The results of this estimation for both the complete network and directed circle are presented in Table 7 with the associated b_i distributions in Figure 8.

Table 7: Structural Estimation: Punishment

	Directed Circle	Complete Network
α	1.231^{***}	1.002^{***}
	(.001)	(.001)
μ	1.525^{***}	1.833^{***}
	(.000)	(.000)
σ	3.276^{***}	3.968^{***}
	(.003)	(.000)

Robust standard errors in parentheses.
*** $p < .01$, ** $p < .05$, * $p < .10$,

Figure 8: Estimated Distribution of b_i's.



These estimates confirm the results of the hurdle model in the sense that the blame parameter, α, is greater in the directed circle than in the complete network. This difference is significant at the 1% level ($p < .000$). The distribution of b_i's indicates that the subjects tend to be only mildly altruistic with mean b_i's of 1.525 and 1.833 for the directed circle and complete network, respectively (significantly different at the 1% level ($p < 000$)) and with a healthy fraction of the mass of the b_i distribution taking on negative values. This is not surprising because we only consider the data from the punishment stage of the experiment where the subjects have little scope to show their altruistic concerns for others.

One note of caution regarding these estimates is that we do not claim that these specific values of α, μ,and σ are parameter values that we consider relevant for other games

or situations. As we stated before, we consider blame to be a context-dependent theory where acts that seem blame free in one context can be emotionally charged in another. As a result, the blame parameter estimated here (especially since it is unit sensitive given our normalization) may not be transferrable to other games or situations. (Note that α varies across our two networks indicating that even within the context of a public goods game, the blame parameter differ depending on the network used.)

Despite this caveat, these estimates are intuitively appealing. First, given that many subjects free ride, it is not surprising that our parameter estimates indicate only mild altruistic sentiments among our subjects. This is reinforced by the fact that subjects do punish relatively frequently indicating that α should be significantly greater than zero, which is in our estimates using both our directed circle and complete networks.

4 Conclusions

In this paper, we have proposed and tested a theory of kindness that is an essential ingredient to any theory of reciprocity. Simply put, our theory of kindness states that in judging whether or not player i has been kind to player j, player j would have to put himself in the strategic position of player i, and ask himself how he would have acted under identical circumstances. If j would have acted in a worse manner than i acted, then we say that j does not blame i for his behavior. If, however, j would have been nicer than i was, then we say that "j blames i" for his actions (i's actions were blameworthy.) After presenting a formal definition of this concept, we investigated how it can be applied to the analysis of a dynamic psychological extensive form games, and then developed the notion of a Sequential Blame Equilibrium.

Using a simple modified dictator game experiment and a public goods game with punishment, we demonstrated that our theory has substantial amount of explanatory power, especially when compared to the other competing models.

The theory of blame we present has some features that are both intuitively and empirically appealing. First, as we have argued, theories of kindness or fairness should be endogenous or at least be based on the individual tastes and preferences of the person making an assessment. One-size-fits-all theories that impose the same norm (i.e., equity) on all decision makers fail to capture the individual differences in kindness assessments.

References

[1] Robert Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63:1161–1180, 1995.

[2] Pierpaolo Battigalli and Martin Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35, 2009.

[3] Sally Blount. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2):131–144, 1995.

[4] Gary E. Bolton and Axel Ockenfels. Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, 2000.

[5] Jordi Brandts and Gary Charness. Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, 2:227–238, 2000.

[6] Jordi Brandts and Gary Charness. The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, 14:375–398, 2011.

[7] Jeffrey Carpenter, Shachar Kariv, and Andrew Schotter. Network architecture and mutual monitoring in public goods experiments. *Review of Economic Design*, 12(2):93–118, 2012.

[8] Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869, 2002.

[9] Dominique J.-F. De Quervain, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. The neural basis of altruistic punishment. *Science*, 305(5688):1254–1258, 2004.

[10] Martin Dufwenberg and Georg Kirchsteiger. A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298, 2004.

[11] Armin Falk, Ernst Fehr, and Urs Fischbacher. On the nature of fair behavior. *Economic Inquiry*, 41(1):20–26, 2003.

[12] Armin Falk, Ernst Fehr, and Urs Fischbacher. Testing theories of fairness-intentions matter. *Games and Economic Behavior*, 62(1):287–303, 2008.

[13] Armin Falk, Ernst Fehr, and Christian Zehnder. Fairness perceptions and reservation wages—the behavioral effects of minimum wage laws. *Quarterly Journal of Economics*, 121(4):1347–1381, 2006.

[14] Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315, 2006.

[15] Ernst Fehr and Simon Gächter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

[16] Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, 1999.

[17] John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1:60–79, 1989.

[18] Faruk Gul and Wolfgang Pesendorfer. Interdependent preference models as a theory of intentions. Princeton University, mimeographed,, 2011.

[19] Mehmet Y. Gurdal, Joshua B. Miller, and Aldo Rustichini. Why blame? *Journal of Political Economy*, 121(6):1205–1247, 2013.

[20] Benedikt Herrmann, Christian Thöni, and Simon Gächter. Antisocial punishment across societies. *Science*, 319(5868):1362–1367, 2014.

[21] Nagore Iriberri and Pedro Rey-Biel. The role of role uncertainty in modified dictator games. *Experimental Economics*, 14:160–180, 2011.

[22] David Kreps and Robert Wilson. Sequential equilibrium. *Econometrica*, 50:863–894, 1982.

[23] David K. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1:593–622, 1998.

[24] Benoît Monin. Holier than me? threatening social comparison in the moral domain. *International Review of Social Psychology*, 20(1):53–68, 2007.

[25] Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302, 1993.

[26] Andrew Schotter. *Free Market Economics: A Critical Appraisal*. Blackwell Publishers, New York, NY, 2nd edition, 1990.

[27] Joel Sobel. Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2):392–436, 2005.

[28] Christian Thöni. Inequality aversion and antisocial punishment. *Theory and Decision*, 76:529–545, 2014.

Appendices

Appendix A: Experimental Instructions

This is an experiment in economic decision-making. If you make good decisions you may be able to earn a good payment, which will be given to you at the end of the experiment.

Experimental Procedures

The experiment is composed of two stages. We will hand out instructions for Stage 2 after we are finished with Stage 1.

Stage 1

In Stage 1 of the experiment you will be given 10 tokens. Your task will be to divide this 10 tokens between you and an anonymous other person in this room. That means you will be asked to state how much of the 10 tokens you want to give to the other person, and therefore how much you would retain for yourself.

After you have made your choice, you will be randomly divided into two equally sized groups called *Senders* and *Receivers* and you will be matched in pairs. If you are chosen to be a *Sender* your payoff will be equal to what you have decided to keep for yourself of the 10 tokens. If you are chosen to be a *Receiver*, your payoff will be the amount given to you by the *Sender* you matched with.

You will not be told your payoff from Stage 1 until after the entire experiment is completed.

Note: The number you will be asked to enter in the screen will be the amount you want to give to another anonymous person in the room paired with you.

Stage 2

In Stage 2 of the experiment you will be randomly matched with another subject in the lab. This random matching will be *independent* of the one performed in Stage 1 so there is very little chance you will be matched with the same person. After you are matched you will be offered, as a *Receiver*, the amount your matched pair member sent as a *Sender* in Stage 1. In other words, if your matched pair member was a person who gave 3 tokens in Stage 1, you will be offered 3 tokens in Stage 2 as your payoff and your pair member will get 7.

Although you cannot change the amount you are given, you have the option to reduce the amount your match pair member keeps for himself/herself by deciding whether to reduce his/her payoff by 1 token. You can do this is a slightly indirect manner. Rather than stating whether you want to reduce your match pair member's tokens or not after you see the amount given to you, we will ask you **before** you see the offer to state an amount below which you will decide to reduce the pair member's payoff by one token. Call this

number your cutoff and assume for illustrative purposes that you entered a cutoff of 4 into the computer. If your matched pair member gave you 3 tokens and kept 7 tokens for himself/herself, since 3 is less than cutoff of 4, the computer will reduce your matched pair member's payoff by 1 token from 7 to 6. If the other person gave you 5 tokens and kept 5 tokens, his/her tokens will not be reduced and will stay 5.

Your final payoff will be determined as follow. After you have made your choice, the computer will randomly determine whether you are a *Sender* or a *Receiver*. Clearly, if you are a *Receiver* then your match pair member is a *Sender* and vice versa. If you are chosen to be a *Sender* your payoff will be equal to what you have decided to keep for yourself of the 10 tokens in Stage 1 minus any 1-token reduction by your pair member. If your math pair member decided not to take a token away, you will receive your payoff undiminished. Otherwise, you will receive one token less than what you kept for yourself. If you are chosen to be a *Receiver*, your payoff will be the amount given to you by your match.

Final Payoff Your final payoff from the two stages of the experiment will simply be the sum of your payoffs in each Stage. You will be paid $1 for each token you have at the end of the experiment.

Appendix B: Self Blame

So far, we based our discussion on the assumption that player i does not blame himself when he puts himself in player j's position. In this section, we will show that this assumption in fact holds as a result in any equilibrium. In order to do that, we have to extend the framework to include higher-order meta-players and their beliefs. Note that whenever our current notations and concepts naturally extend to this general framework, we do not introduce them again.

Let $N^0 := N$, $N^1 := \{\langle ij\rangle : i,j \in N^0\}$, $N^n := \{\langle ki\rangle : k \in N^{n-1}, i \in N^0\}$, and $N^* = \bigcup_{n=0}^{\infty} N^n$. In addition, if $k = \langle li\rangle$ for some $l \in N^*$, then we write $\langle ki\rangle = k$, i.e. we rule out a player putting himself in the same position twice in a row, as it is irrelevant in our theory. Note that a player $k \in N^* \setminus N$ takes the form $k = \langle \cdots \langle\langle ij\rangle i\rangle j\rangle \cdots \rangle$. In this case, we denote the *original player* i by $\phi(k) := i$.

We maintain the assumption that the material payoff and the actions available to player $\langle ki\rangle \in N^*$ are the same as those of player $i \in N$; i.e. $A_{\langle ki\rangle} = A_i$ and $\pi_{\langle ki\rangle} = \pi_i$, for $i \in N$, and $k \in N^*$. Furthermore, we introduce the following "self consistency" assumption.

Assumption (Self consistency)**.** *For any $\langle ki\rangle \in N^*$, if $\phi(k) = i$ then $u_{\langle ki\rangle} = u_i$.*

This means that when the original player $\phi(k)$ eventually considers himself back in his own position, he retains the same preferences. Put differently, when a player puts himself in his own position while he considers being in the the other player's position his preferences are consistent.

For any player $k \in N^*$, strategies and conditional beliefs are defined exactly as before, and hence we do not repeat the definitions here. In order to define player k's blame we need to write the expected material payoff when he plays against an opponent in N, and against himself in the opponent's position. Let $k = \langle li\rangle$ for some $l \in N^*$, then

$$\begin{aligned}
\mathsf{E}_{\mu_k}\big[\pi_i\big(\zeta(s_k,s_j)\big)\big|h\big] &:= \int_{S_k\times S_j} \pi_i\big(\zeta(s_k,s_j)\big)\, d\mu^1_{kj}(s_j|h)\mu^2_{kj}(s_k|h),\\
\mathsf{E}_{\mu_k}\big[\pi_i\big(\zeta(s_k,s_{\langle kj\rangle})\big)\big|h\big] &:= \int_{S_k\times S_j} \pi_i\big(\zeta(s_k,s_j)\big)\, d\mu^1_{k\langle kj\rangle}(s_{\langle kj\rangle}|h)\mu^2_{k\langle kj\rangle}(s_k|h).
\end{aligned}$$

Therefore, a player $k = \langle li\rangle \in N^*$ who holds beliefs μ_k blames player $j \in N$ at history h by

$$\delta_{kj}(\mu_k|h) := \mathsf{E}_{\mu_k}\big[\pi_i\big(\zeta(s_k,s_{\langle kj\rangle})\big)\big|h\big] - \mathsf{E}_{\mu_k}\big[\pi_i\big(\zeta(s_k,s_j)\big)\big|h\big].$$

This allows us to define player k's psychological utility function $U_k(z,\mu_k(\cdot|h))$. As before, for player $k = \langle li\rangle$, we define the function $u_k(z,\delta_{kj}(\mu_k|h))$ such that $u_k(z,\delta_{kj}(\mu_k|h)) := U_k(z,\mu_k(\cdot|h))$, and assume that u_k is non-increasing in δ_{kj}.

Behavioral strategies, players' assessments, and consistency of an assessment are defined exactly as before in Section 2.2. Consequently, (the sequential equilibrium) Definition 4 applies to the blame game $\langle N^*, H, (u_k)_{k\in N^*}\rangle$. We are ready to state the result, which we call no self-blame.

Proposition 1 (No self-blame)**.** *Let the assessment (σ^*, μ^*) be a sequential equilibrium of the blame game $\langle N^*, H, (u_k)_{k\in N^*}\rangle$. For any $k \in N^* \setminus N$, such that $\phi(k) = i$ and $\langle kj\rangle = k$, we have $\delta_{ki} = 0$.*

Proof. Suppose that assessment (σ^*, μ^*) is a sequential equilibrium. Also, let $k \in N^* \setminus N$ such that $\phi(k) = i$ and $\langle kj\rangle = j$, then for any $h \in N$ we have

$$\delta_{ki}(\mu_k^*|h) = \mathsf{E}_{\mu_k^*}\big[\pi_i\big(\zeta(s_k, s_{\langle ki\rangle})\big)\big|h\big] - \mathsf{E}_{\mu_k^*}\big[\pi_i\big(\zeta(s_k, s_i)\big)\big|h\big].$$

First observe that $\mu^{2*}_{ki} = \mu^{1*}_{ik}$, and $\mu^{2*}_{k\langle ki\rangle} = \mu^{1*}_{\langle ki\rangle k}$ by consistency of beliefs (Definition 3 *(b)*). Moreover, $\mu^{1*}_{\langle ki\rangle k} = \mu^{1*}_{ik}$ by consistency of beliefs (Definition 3 *(a)*); hence, $\mu^{2*}_{ki} = \mu^{2*}_{k\langle ki\rangle}$.

Observe that

$$\begin{aligned}
\mathsf{E}_{\mu_i^*}[u_i|h] &= \int_{S_j} u_i\big(\zeta(s_i, s_j), \delta_{ij}(\mu_i^*|h); b_i\big)\, d\mu^{1*}_{ij}(s_j|h) \text{ and,}\\
\mathsf{E}_{\mu^*_{\langle ki\rangle}}[u_{\langle ki\rangle}|h] &= \int_{S_j} u_{\langle ki\rangle}\big(\zeta(s_{\langle ki\rangle}, s_j), \delta_{\langle ki\rangle j}(\mu^*_{\langle ki\rangle}|h); b_i\big)\, d\mu^{1*}_{\langle ki\rangle j}(s_{\langle ki\rangle}|h).
\end{aligned}$$

Since $\mu^{1*}_{ij} = \mu^{1*}_{\langle ki\rangle j}$ by consistency, and $u_i = u_{\langle ki\rangle}$ by self consistency, we have $\sigma^*_i = \sigma^*_{\langle ki\rangle}$. Consequently, in the equilibrium, $\mu^{1*}_{ki} = \mu^{1*}_{k\langle ki\rangle}$. This immediately implies $\mathsf{E}_{\mu_k^*}\big[\pi_i\big(\zeta(s_k, s_{\langle ki\rangle})\big)\big|h\big] = \mathsf{E}_{\mu_k^*}\big[\pi_i\big(\zeta(s_k, s_i)\big)\big|h\big]$. Thus, $\delta_{ki}(\mu_k^*|h) = 0$. □

The critical assumption behind this result is self consistency. In order to further elaborate suppose that $k = \langle ij\rangle$. Then, in the equilibrium the meta-player $\langle ij\rangle$ does not blame i (himself) because what he believes i plays is the same as what he would have done in that position simply because i's preferences are exactly the same as "player i who considers himself in player j's position" puts himself in player i's position.