ON BLAME-FREENESS AND RECIPROCITY: AN EXPERIMENTAL STUDY

Mariana Blanco
Bogaçhan Çelen
Andrew Schotter

SERIE DOCUMENTOS DE TRABAJO

No. 85

Julio 2010

On Blame-freeness and Reciprocity: An Experimental Study*

Mariana Blanco† Boğaçhan Çelen‡ Andrew Schotter§

Universidad del Rosario Columbia University New York University

June 28, 2010

Abstract

The theory of reciprocity is predicated on the assumption that people are willing to reward nice or kind acts and to punish unkind ones. This assumption raises the question as to how to define *kindness*. In this paper we offer a new definition of kindness that we call "blame-freeness." Put most simply, blame-freeness states that in judging whether player i has been kind or unkind to player j in a social situation, player j would have to put himself in the strategic position of player i, while retaining his preferences, and ask if he would have acted in a manner that was worse than i did under identical circumstances. If j would have acted in a more unkind 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). We consider this notion a natural, intuitive and empirically relevant way to explain the motives of people engaged in reciprocal behavior. After developing the conceptual framework, we then test this concept in a laboratory experiment involving tournaments and find significant support for the theory.

JEL Classification Numbers: A13, C72, D63.

Keywords: Altruism, blame, reciprocity.

*We are grateful to participants in the CESS Experimental Economics Lunchtime Seminar, 2009 North-American ESA Conference, the Amsterdam Workshop on Behavioral & Experimental Economics, and seminars at Cornell University and the Brown University for comments. We also acknowledge the partial financial support of the Center for Experimental Social Science at NYU. The findings, recommendations, interpretations and conclusions expressed in this paper are those of the authors and not necessarily reflect the view of the Department of Economics of the Universidad del Rosario.

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

‡C.E.S.S., New York University, and Graduate School of Business, Columbia University, 3022 Broadway, 602 Uris Hall, New York, NY 10027. E-mail:bc2132@columbia.edu, url: http://celen.gsb.columbia.edu/.

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

1 Introduction

Recent years have witnessed a growing literature on the theory of reciprocity. Founded on the seminal work of Rabin [18]—further extended by Falk and Fischbacher [10], Dufwenberg and Kirschteiger [5] and others, and generalized by Sobel and Segal [21]—the theory of reciprocity is predicated on the assumption that people are willing to reward nice or kind acts and to punish unkind ones.[1] This assumption raises the question of as to how to define "kindness." In this paper we offer a definition of kindness that we call blame-freeness. The notion that we propose is a natural, intuitive and empirically relevant way to explain the motives of people engaged in reciprocal behavior. We develop the conceptual framework and then test it in a laboratory experiment involving tournaments.

Put most simply, blame-freeness states that in judging whether player i has been kind or unkind to player j in a social situation, player j would have to put himself in the strategic position of player i, while retaining intrinsic characteristics of his preferences (i.e., j does not become i but simply takes his strategic position), and ask if he would have acted in a manner that was worse than i did under identical circumstances. If j would have acted in a more unkind 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.e. i's actions were blameworthy. Furthermore, if blame is a source of disutility, j may have the motivation to punish i whenever possible even if the punishment is costly for him. Note that blameworthiness is only a necessary condition for punishment while blame-freeness is a sufficient condition for non-punishment. In other words, in the strong form of the theory, we should never observe any player punishing those whose actions are judged blame-free, i.e., actions which they themselves would have taken if they were in the same situation as the other player.

This way of viewing kindness is distinctly different from others in a number of ways. First, as stated in Schotter [19], blame-free justice is an endogenous, process-oriented theory in which people judge the actions of others by their own standards and personal norms but not by some exogenous standard imposed on them by the analyst.[2] Blame-freeness allows the standards that people use to judge the actions of others to differ from person to person depending on their personal norms and background. Indeed, actions that bother you may not bother other people at all, and things that strike you to be fair may be very upsetting to others. In addition, the theory is not independent

[1] For a comprehensive survey on reciprocity see Sobel [23].

[2] The most notable exogenous norm is egalitarianism or inequality aversion studied by Bolton, Fehr and Schmidt [12] or Bolton and Ockenfels [2].

of the context.[3] For instance, actions that are blame-free in a prison (or a college dorm) may certainly be blameworthy in normal civilian life. One cannot judge the behavior of people in isolation—we need to know the institutional setting they are in. Finally, as stated above, blame-freeness judges the actions of people that lead to outcomes and not merely the outcomes themselves. So, it is also a process-oriented theory. This is counter to those theories that are outcome-based. Finally, our theory is distinct from intention-based theories [1, 7, 10, 9, 8] since blame-freeness is totally self-referential: it only matters what you would have done in your opponent's situation and not what he intended to do.

To put some flesh on this notion of blame-freeness and to differentiate it from other theories reciprocity, let us consider a few examples of how our analysis differs from those of other scholars.

1.1 Rabin-Charness, Kindness, and Blame

In Rabin [18]'s theory of fairness, person j judges person i's actions as being unkind if they lead to a payoff for j that is less than $1/2$ of the total payoff available along person j's payoff frontier, given j's beliefs about what i thinks j will do. In other words, there is a split-the-difference ethic imposed by Rabin that is supposed to define kindness no matter what situation is under investigation.[4] But what if i's action led to a payoff for j that was only $1/3$ of the total available but j, if he was in i's position, would have given his opponent even less, say $1/6$. Under what circumstance should j be upset with i While he may not like his payoff, he certainly understands j's actions and in fact, compared to what he would have done, he must even consider i to be generous. The point, therefore, is that feelings of justice, fairness and kindness are subjective and must emanate from the person doing the evaluation himself. They should not be imposed from the outside using some other (e.g. egalitarian) standard.

In a related paper, Charness and Rabin [4] define a "demerit parameter" which captures how a player feels towards his opponent. This parameter is determined by comparing the behavior of an opponent to what a "decent person" would do in his circumstances. In this approach, therefore, an opponent's action is considered in relation to an exogenously determined social norm, while in ours, the standard used to judge behavior is endogenous and defined by our blame-free norm.

[3] Gul and Pesendorfer [14] lay the foundations of interdependence between behavioral types, independent of the environment decision-makers interact.

[4] It is important to point out that we are not criticizing Rabin's specific kindness function here as much as the use of any exogenously imposed kindness function.

1.2 Fehr-Schmidt, Inequality Aversion, and Blame

Mr. Nasty and Ms. Very Nasty play an Ultimatum game. Mr. Nasty offers Ms. Very Nasty \$1 out of \$100 when placed in the proposer's position. Ms. Very Nasty, on the other hand, would offer only \$ .50 if she were in that position. Ms. Very Nasty accepts Mr. Nasty's \$1 offer. Why? Clearly from Ms. Very Nasty's point of view she receives an amount that is higher than the amount she would offer Mr. Nasty if she were the proposer. In fact, Ms. Very Nasty might even gloat that Mr. Nasty offered her far more than she would have offered him if she had been in his position.

To be more precise, consider an Ultimatum game played between two Fehr-Schmidt (Fehr and Schmidt [12]) players, p (Proposer) and r (Receiver), each endowed with a utility function of the form,

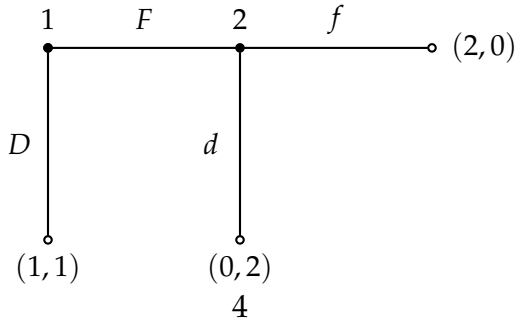
$$\begin{aligned} u_p(x_p, x_r) &= x_p - \alpha_p \max\{x_r - x_p, 0\} - \beta_p \max\{x_p - x_r, 0\}, \\ u_r(x_p, x_r) &= x_r - \alpha_r \max\{x_p - x_r, 0\} - \beta_r \max\{x_r - x_p, 0\}, \end{aligned}$$

where $\alpha_i \geq \beta_i$, and $1 > \beta_i \geq 0$ for $i = p, r$. In such a 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 blame-free theory the utility functions of players are not restricted to be of any form. What is required is that a 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 position. Hence, under blame-free hypothesis, the Receiver compares the offer (x_p, x_r) to what offer he would have made if he were the Proposer, say (x_p^*, x_r^*). If $u_r(x_p, x_r) \geq u_r(x_p^*, x_r^*)$, he accepts, while if $u_r(x_p, x_r) < u_r(x_p^*, x_r^*)$, he blames. If blame causes a disutility for the receiver, he may even reject the offer.

1.3 Dufwenberg-Kirchsteiger, Intensions, and Blame

One strand of reciprocity theory considers intensions as its focal point. Consider the following game of Dufwenberg and Kirchsteiger [5] (DK) presented in Figure 1. that they use to motivate the relevance of intentions in modeling reciprocity.

Figure 1: Game Γ_1 in DK.



In this game Player 1 moves first and can either play F or D. The question that DK ask is under what circumstance can Player 1's playing F be considered as kind. Their answer is that whether Player 1's action is kind or not depends on what his beliefs are about what Player 2 is going to play. If Player 1 believes that Player 2 will play f, then Player 1's action of F will be considered unkind since it will reduce Player 2's payoff from 1 to 0. However, F will be considered a kind move if Player 1 believes that Player 2 will play d at his node. Hence, in their model while players' beliefs about each other are the center of attention, in our analysis we focus on the preferences of players when they are placed in each others strategic position.

To elaborate on this distinction let us return to Figure 1. According to the theory of blame-freeness in order to define whether a move of F is kind or unkind Player 2 must place himself in the strategic position of Player 1 and ask whether he would have chosen F if he were in Player 1's position. For example, say that Player 2 is a strict egalitarian (Fehr-Schmnidt) player who would prefer the outcome $(1, 1)$ to any other outcome in the game. From his perspective, if Player 1 were to move F, he would immediately consider such a move blameworthy (i.e., unkind) since he would never have chosen to do that. Note that, unlike DK, no beliefs are required in order to make this judgment.

1.4 Levine, Altruism and Blame

Another popular theory of reciprocity is that of Levine [16] where the utility that a player receives from his actions depend on his own and his opponents' types. Since the types are private information and drawn from a commonly known distribution, the game is modeled as a Bayesian game. For Levine [16] there is a one-to-one mapping from the "niceness" of one's opponent's actions to his type. The utility that a player receives from an outcome is a function of the player's "direct" and "adjusted" utility where the direct utility (u_i) is simply the player's material payoff while the adjusted utility (v_i) takes into account his assessment of how nice his opponent is. More precisely Levine [16] posits a generalized version of the following utility function:

$$v_i = u_i + \frac{a_i + a_j}{2} u_j$$

$-1 < a_i, a_j < 1$ are the coefficients of altruism (types) of players i and j respectively. Note that given i's altruism coefficient, player i's utility, v_i, is an increasing function of his assessment of j's type, meaning that the nicer player i thinks that player j has been the more he cares about his payoff. Under our theory, however, this judgement is a relative one. Player i perceives player j as nice only if player j has taken an action which is blame-free i.e., nicer than the action he would have taken if he were in j's position.

This distinction is meaningful. For example, say that $a_i = 0.9$ while $a_j = 0.8$; that is both players i and j are "nice" but i is nicer than j. In the context of blame-free theory i would blame j for not being as nice as he would be in j's position. However in Levine [16], player j is considered nice regardless of i's altruism parameter. Put differently in our model niceness is a relative concept, while in Levine [16] it is an absolute concept.

1.5 Overview and Summary

As we can see from the examples above, the essence of blame-freeness involves the examination of a counter factual, i.e. imagining what you would have done if you were in the position of the person whose actions you are judging. Although real world data does not lend itself to such observations, in the lab it is possible to allow subjects to play all roles in a game anonymously and then test to see if their behavior is consistent with the blame-free hypothesis. The experiment demonstrated in this paper test this hypothesis.

In the experiment, subjects engage in an asymmetric tournament identical to the one used by Schotter and Weigelt [20] (hereafter SW), where players have different costs of effort. In each round a subject plays in two tournaments; one in the role of the advantaged player (low cost) against a disadvantaged player (high cost), and another in the role of the disadvantaged player against an advantaged player. In other words, subjects play in both roles simultaneously in two tournaments with two different opponents who are in the opposite roles. This experiment was used because it was noted in SW as well as Kräkel [15] that asymmetry in tournaments lead to strong emotional responses to the behavior of advantaged subjects whose impact we are attempting to capture here. In addition to tournament stage, in our experiment, we have a punishment stage. In the tournament stage of the game, subjects choose an effort level and then move to the punishment stage where they can punish their opponent.

Since we are able to observe what a subject does when he is in the advantaged or disadvantaged role, we can check if a subject punishes according to the predictions of the blame-free theory—e.g. whether a disadvantaged subject punishes his advantaged opponent for choosing an effort level that is above his own effort level in the advantaged role. Note however, that although under the blame-free hypothesis while choosing a blame-free effort in the tournament is a sufficient condition for non-punishment, choosing a blameworthy effort level is only a necessary condition for punishment. This is so because whether a subject ultimately punishes his opponent depends not only on his blaming him for his actions but also on the cost of punishment and how much blame exists.

The results of our experiment support the view that people consider blame as part of their notion of kindness. Remembering that blame-freeness is a sufficient condition for non punishment, and focusing on disadvantaged subjects who are most likely to blame their advantaged opponents for taking advantage of their positions, we see behavior consistent with this view 81% of the time. In other words, when disadvantaged subjects face advantaged opponents who choose lower effort levels than they do when placed in their position, they fail to punish them 81% of the time. On the other hand, blameworthy acts only constitute a necessary condition for punishment. Of the sub-sample that assigned punishment points to their opponents, 73% punish blameworthy acts at least 50% of the time.

The actual adherence to blame-freeness cannot be determined from these numbers since, as we have said, blameworthiness is only a necessary condition for punishment (many factors mitigate its influence such as the cost of punishment, the sensitivity of blame etc.). Consequently, these statistics are a lower bound on adherence to the theory. Some probit regressions run also indicate that it is the actions of opponents that spark blame and not the final payoffs since adding payoffs as a dummy variable in the probit regression is not significant once actions are accounted for. This means subjects were not focusing on the payoff distribution of the tournament game, as Fehr and Schmidt [12] would predict, but both their payoffs and the strategies determining them. Intention-based theories such as [18] also seems to have less explanatory power.

In summation, the evidence presented here suggests that a non-negligible part of the population behave according to the prescriptions of blame-freeness. We believe that this notion of justice has many advantages with respect to other fairness theories in the literature. Not only does it allow us to relax the assumption that the most preferred distribution is the egalitarian one, but it can also rationalize experimental data that has not yet been explained by other fairness theories.

The paper is organized as follows. Section 2 introduces the *blame-freeness* concept in a more rigorous manner and provides a formal example of how to compute blame and equilibria in games exhibiting blame. Section 3 presents the theory of tournaments used in our experiments and demonstrates how the inclusion of blame alters the standard theoretical results for such tournaments. In Section 4 we explain the experimental design. Section 5 presents our results and Section 6 concludes.

2 Blame in Extensive Games

Before we can introduce considerations of blame into the tournament game that serves as the basis of our experiment, it is necessary to pause and discuss the concept more

abstractly so that we can establish exactly what we have in mind. Accordingly, in this section, we discuss a finite extensive form game with complete and perfect information where blame is a motivating factor. After we do this, we then integrate the concepts we develop into our tournament model.

Players, actions, histories and preferences. Consider a game consisting of two players $i = 1, 2$.[5] The set of histories $\mathcal{H}$ is composed of the initial history $\varnothing$ as well as finite sequences of players' actions. We say that a history h is terminal if there is no action profile a such that $(h, a) \in \mathcal{H}$. We call a terminal history an *outcome* and denote the set of all outcomes by H, and the space of lotteries over the outcomes by $\Delta(H)$. Associated with each terminal node is a *prize* for each player. Hence a prize function $\pi_i : H \mapsto \mathbb{R}$ determines the prize $\pi_i(h)$ of player i associated with the outcome $h \in H$. We posit that players have *preferences* over the space of lotteries over outcomes that are represented by *utility function* $v_i : \Delta(H) \mapsto \mathbb{R}$.[6]

In order to get a meaningful answer to the question of what a player would do if he were in the position of his opponent, we need to allow our subjects to differ in a meaningful way since otherwise all players would act identically in every position they found themselves and no blame could exist. While there are many ways to introduce this difference, for simplicity, we assume that people differ according to their "caring parameter"; b_i indicates how much weight they place on their opponent's payoff in their utility function and the function is written as:

$$v_i(h; b_i) := \pi_i(h) + b_i \pi_j(h).$$

$v_i(h)$ depicts the fact that player i's utility is determined as the sum of his prize and a proportion of player j's prize. The fraction $b_i \in [0, 1]$ is the weight attached to other player's prize and we say that b_i is player i's caring parameter that describes his altruism towards the other player. We stress, however that nothing in the theory of blame-freeness depends upon this functional form assumption.

As a first step, to answer what it means to put oneself in another's strategic position we assume that a player i is endowed with a function $v_{ij} : \Delta(H) \to \mathbb{R}$ that represents his preferences over the space of lotteries over outcomes if he were in player j's position. Since these preferences are parameterized by player i's caring parameter we will write $v_{ij}(\cdot\,; b_i)$. It is important to note that when player i puts himself in player j's position

[5] Although the notions that we discuss extend to the case of $n \geq 2$ players with some effort, for expositional and notational ease we focus on two-player games.

[6] Segal and Sobel [21] focuses on the representation of strategic preferences and the existence of Nash equilibrium when the players are strategic utility maximizers. Although there are minor differences in our definition of strategic preferences, a careful modification of their theorems also apply to our framework. Since the scope of the present paper aims to emphasize different issues, we omit this exercise.

he retains his own preferences over outcomes—i.e. his own caring parameter—so that while he is in player j's strategic position he views it from his own perspective.

Strategies. A strategy of a player i is a map σ_i that determines an action for each non-terminal history $h \in \mathcal{H} \setminus H$ whenever it is player i's turn to move. We write Σ_i to denote the set of all strategies for player i and we write $\Sigma := \Sigma_1 \times \Sigma_2$. If a strategy profile $\sigma = (\sigma_1, \sigma_2)$ leads to an outcome $h \in H$, we denote it by h_σ whenever we want to refer to the strategy profile explicitly.

Strategic preferences. Since players blame others for their behavior that lead to final outcomes we need to include this behavior in a player's utility function. To do this we follow Sobel and Segal [21, 22] who demonstrate that under appropriate assumptions, one can characterize a subject's evaluation of the kindness of his opponent by a utility function that increases a player's caring parameter for opponents who behave kindly and decreases it when an opponent behaves unkindly. In other words, reciprocity occurs because a kind action by one's opponent leads one to care more about him and therefore take actions that are better for him. Conversely, an unkind action has the opposite effect. In this paper the kindness function will be defined via our notion of blame.[7] To formalize this we assume that in strategic environments, players have preferences over the strategy profiles $\Sigma = \Sigma_1 \times \Sigma_2$, which we call *strategic preferences*. These preferences are represented by a *strategic utility function* $u_i : \Sigma \to \mathbb{R}$.

As in Sobel and Segal [21, 22] this strategic utility function allows us to incorporate the behavior of a player leading to an outcome as an argument in a player's utility function along with the outcome itself. In particular, we assume that players maximize their strategic utilities when there is an explicit reference to the strategy that leads to an outcome. While this assumption allows us to impose more structure on behavior, it does not rule out the standard approach. Indeed, if a player i is indifferent between any two outcome-equivalent strategies, his strategic preferences are equivalent to his preferences over outcomes.

Blame. Our exposition so far is general. In what follows, we discuss our key blame concept that defines a class of strategic preferences that characterize a player i who *evaluates* an outcome reached through a strategy profile (σ_i, σ_j) by asking himself:

What would I do if I were in player j's position playing against strategy σ_i?

The function v_{ij} is key to answering this question. Being endowed with functions (v_i, v_{ij}), player i can judge an outcome from player j's perspective. As a matter of fact,

[7] Sobel and Segal [21, 22] do not tie themselves to any specific notion of kindness since their analysis is on a level of generality higher than ours. They do, however, suggest a few possibilities none of which include blame.

we assume that given σ_i player i chooses $\sigma_{ij} \in \Sigma_j$ to maximize $v_{ij}\big(h_{(\sigma_{ij},\sigma_i)}; b_i\big)$. Hence player i's answer to the previous question is:

> If I were in player j's position playing against strategy σ_i, I would play $\sigma_{ij} = \arg\max_{s\in\Sigma_j} v_{ij}\big(h_{(s,\sigma_i)}; b_i\big)$.

This leads to our definition of *blame*: In a strategy profile σ, player i is said to blame player j if

$$\delta_i^\sigma := v_i\big(h_{(\sigma_{ij},\sigma_i)}; b_i\big) - v_i\big(h_\sigma; b_i\big) > 0.^{8}$$

Let us explain this definition. If player i would play σ_{ij} as a response to σ_i in player j's position, then player i's utility would be $v_i\big(h_{(\sigma_{ij},\sigma_i)}; b_i\big)$. This is the utility he would get if he played against someone like himself (in fact this is the utility he would receive if he actually did play against himself). His utility when playing against his actual opponent is $v_i\big(h_\sigma; b_i\big)$. So player i blames player j at σ when the utility he could enjoy if player j was behaving exactly like he would is more than the utility he actually enjoys against his true opponent. Note that player i does not blame j if j's strategy led to a utility for i that is larger than the utility i would have achieved if j had chosen strategy σ_{ij}—i.e. if he does not blame player j if he is nicer to i than he would have been to *himself*.

In order to capture the strength or intensity of this blame we define player i's *blame function* as $f_i(\delta_i^\sigma)$. For tractability we assume that f_i is non-negative, continuous, non-decreasing in δ_i^σ and zero when $\delta_i^\sigma \leq 0$.

Now we can define the strategic utility function for player i of type b_i with reference to blame in the following way:

$$u_i(\sigma; b_i) := v_i\Big(h_\sigma; \beta_i(\sigma)\Big)$$

where $\beta_i(\sigma) := b_i - f_i(\delta_i^\sigma)$. That is, at the strategy profile σ, player i's utility of the outcome h_σ depends on σ by altering his caring parameter from b_i to $\beta_i(\sigma)$. So in our previous example, the strategic-utility of player i is

$$u_i(\sigma) = \pi_i(h_\sigma) + \beta_i(\sigma)\pi_j(h_\sigma).$$

Note that when $f_i(\delta_i^\sigma) > 0$ player i blames player j for his actions under σ and when $f_i(\delta_i^\sigma) > b_i$, the blame is so significant that player i actually receives disutility from player j's positive prize. It is in such a case that i may take actions that diminish j's payoff in an

[8] Clearly $\arg\max_{s\in\Sigma_j} v_{ij}\big(h_{(s,\sigma_i)}; b_i\big)$ need not be a singleton. In that case we can revise the definition as $\delta_i^\sigma := \Big(\min_{\sigma_{ij}\in\arg\max_{s\in\Sigma_j} v_{ij}\left(h_{(s,\sigma_i)}; b_i\right)} v_i\big(h_{(\sigma_{ij},\sigma_i)}; b_i\big)\Big) - v_i\big(h_\sigma; b_i\big)$.

effort to restore his utility level. Also note that if $\delta_i(\delta_i^\sigma) < 0$ then, since player i does not blame the other player, we have $u_i(\sigma; b_i) = v_i(h_\sigma; \beta_i(\sigma))$ and player j's actions are blame-free. When j's actions are blameworthy, he is likely to be punished as long as the cost of punishment is not too large. Hence, blameworthiness is only a necessary condition for punishment while blame-freeness is a sufficient condition for lack or punishment.

2.1 Equilibrium

Now we are ready to define the relevant equilibrium concepts in our context. Let $\Gamma := (\mathcal{H}, v_i, v_{ij}, u_i, f_i)_{i=1,2}$ denote the game we defined in Section 2.

Definition 1. *A strategy profile $\sigma^* \in \Sigma$ is a Nash equilibrium of the game Γ if for all i, $u_i(\sigma^*; b_i) \geq u_i(\sigma_i', \sigma_j^*; b_i)$ for all $\sigma_i' \in \Sigma_i$.*

Given that the players are strategic utility maximizers, the equilibrium concept is standard. Note, however, that the Nash equilibrium is defined with respect to a player's strategic utility function $u_i(\sigma^*; b_i)$ and not $v_i(h_{\sigma^*}; b_i)$. This means that the utility of players includes the entire strategy profile (and hence blame if any exists) as an argument.

In order to incorporate sequential rationality in the solution concept we aim to refine the equilibrium in that direction. However, subgame perfect refinement requires more care since the strategic preferences depend on the blame factor at each history of the game. In other words, at each subgame, a player will question why he is at that subgame to begin with. We will discuss this issue in detail in what follows.

Let us write $\Gamma(\tilde{h}) := \left(\mathcal{H}|_{\tilde{h}}, v_i|_{\tilde{h}}, v_{ij}|_{\tilde{h}}, u_i|_{\tilde{h}}, f_i\right)_{i=1,2}$ for the subgame of Γ that succeeds history $\tilde{h} \in \mathcal{H}$. The definitions of the constituents of a subgame, except for the strategic preferences, are standard. $\mathcal{H}|_{\tilde{h}}$ is the set of sequences of actions such that $(\tilde{h}, h) \in \mathcal{H}$ for any $h \in \mathcal{H}|_{\tilde{h}}$. The utility functions are

$$v_i|_{\tilde{h}}(h; b_i) := v_i\big((\tilde{h}, h); b_i\big), \text{ and } v_{ij}|_{\tilde{h}}(h; b_i) := v_{ij}\big((\tilde{h}, h); b_i\big)$$

where $(\tilde{h}, h) \in H$. For a strategy σ_i, we write $\sigma_i|_{\tilde{h}}$ for the strategy that projects σ_i in $\Gamma(\tilde{h})$. That is $\sigma_i|_{\tilde{h}}(h) := \sigma_i(\tilde{h}, h)$ for all $h \in \mathcal{H}|_{\tilde{h}}$.

The issue of defining blame in a subgame is less straightforward. In a given subgame, one can define blame by focusing only on the projection of the strategies in the subgame. However, this definition would not address the question of why players are supposed to play in that particular subgame. Therefore, a reasonable definition of blame should be able to question the strategy profile that takes players to a given subgame. In order to accomplish this goal, in a subgame $\Gamma(\tilde{h})$, at a strategy profile $\sigma|_{\tilde{h}}$, we define $\delta_i^{\sigma|_{\tilde{h}}} := \delta_i^\sigma$, and hence $\beta_i(\sigma|_{\tilde{h}}) = \beta_i(\sigma)$. In other words, the projection of a strategy profile in

any subgame yields the same blame term as it does in the entire game. Given our definition of blame, we can define strategic preferences as before. That is $u_i|_{\tilde{h}}(\sigma|_{\tilde{h}}; b_i) = v_i|_{\tilde{h}}\Big(h_{\sigma|_{\tilde{h}}}; \beta_i(\sigma)\Big)$.

Since our definition of a subgame is complete, we are ready to define subgame perfect equilibrium of a game Γ.

DEFINITION 2. *A strategy profile σ^* is a subgame perfect equilibrium of the game Γ if*
(i) $\sigma^*|_{\tilde{h}}$ is a Nash equilibrium of the game $\Gamma(\tilde{h})$ for all $\tilde{h} \in \mathcal{H} \setminus H$, and
(ii) (no self-threat) for each $i = 1, 2$, $v_i|_{\tilde{h}}\Big(\tilde{h}_{(\sigma_i^*|_{\tilde{h}}, \sigma_{ij}^*|_{\tilde{h}})}; b_i\Big) \geq v_i|_{\tilde{h}}\Big(h_{(\sigma_i|_{\tilde{h}}, \sigma_{ij}^*|_h)}; b_i\Big)$ for all $\sigma_i \in \Sigma_i$, for all $\tilde{h} \in \mathcal{H} \setminus H$.

The subgame perfect equilibrium has two requirements. The first is a standard condition: no player has an incentive to deviate from the equilibrium strategy at each subgame of the game. The second requirement asserts that if $f_i(\sigma) = 0$ for all σ then $\sigma_i^*|_h$ is a best response to $\sigma_{ij}^*|_h$ in each subgame $\Gamma(h)$ for all $i = 1, 2$. Put differently, if we modify the game Γ to a game Γ^i where the utility function of player j is $v_j = v_{ij}$, then $(\sigma_i^*|_h, \sigma_{ij}^*|_h)$ is a Nash equilibrium of subgame $\Gamma^i(h)$ for all $h \in \mathcal{H}$.

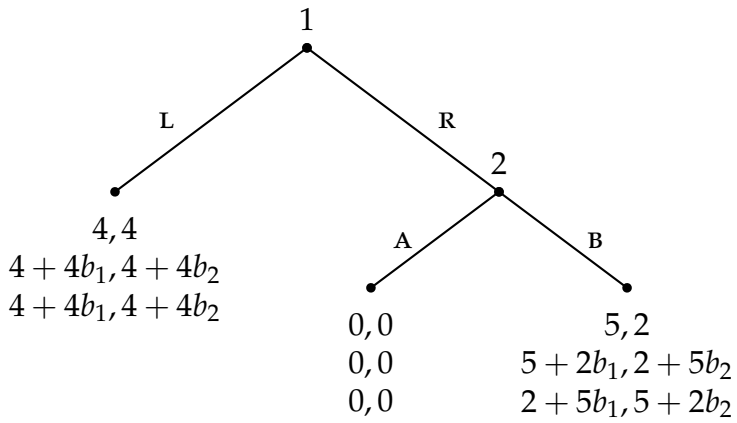
Let us elaborate more on the role of the second requirement. If a player i were playing the game against himself,—i.e. in the position of player j with the utility function v_{ij}—there should not be any blame involved in the strategic relationship. In other words, a player should not *blame* himself in the equilibrium of the game if he were playing against himself.

Let us illustrate the role of this condition with a simple example of ultimatum game, where the players allocate a surplus of size 10. Suppose that regardless of whether he is a responder or proposer, player i's preferences are such that he prefers an allocation that gives him more than 5 to any allocation that gives him less than 5. Also an allocation that gives him 0 is always the least preferred allocation. Suppose that player i is the responder and his strategy is to reject any offer that gives him less than 5. Note that if he were in the proposer's position his best response would be to offer 5 against that strategy. Therefore, any offer that gives him less than 5 will make him blame the proposer. Observe that blame originates from his own strategy that rejects any offer less than 5. Although this strategy is not credible, it makes him blame a proposer who offers him less than 5. This is the point where condition 2 becomes critical by requiring credibility of his strategy in the hypothetical game where he plays against himself.

2.2 A SIMPLE EXAMPLE

To illustrate these ideas in a simple example, consider the following sequential game.

FIGURE 2: A TWO-PERSON EXTENSIVE GAME WITH BLAME.



The first line at the terminal histories is the prizes, the second line is v_i, and the third line is v_{ij}.

We will demonstrate that while the only subgame perfect equilibrium of this example for selfish rational players ($b_i = 0$) results in outcome (R,B), when strategic preferences are a function of blame, for some caring parameters, the only subgame perfect equilibrium outcome is L. This is true because in the presence of blame following the history R, player 2 can credibly play A if R is chosen. Note that at each terminal node we have three sets of payoffs. The first set of payoffs are the prize payoffs for each player at each terminal node, $\pi_i(h)$, $i = 1, 2$. However, these are not the utility payoffs at these nodes since we have assumed that our players care about the prizes received by their opponents through their caring parameters. Hence according to a player's preferences these outcomes are evaluated as follows:

$$v_i(h; b_i) = \pi_i(h) + b_i \pi_j(h) \text{ for } i, j = 1, 2,$$

where $b_i \geq 0$, for $i = 1, 2$. This yields the second set of payoff vectors in Figure 2.

For our analysis of blame we also need to know what each player would do if he were in the strategic position of his opponent. Thus, we need to define the utility of each player for all histories if that player were in the role of his opponent. These payoffs are defined as

$$v_{ij}(h; b_i) = \pi_j(h) + b_i \pi_i(h) \text{ for } i, j = 1, 2.$$

In order to analyze the equilibrium of this game for players who maximize their strategic utilities, we should determine players' strategic preferences over all strategy profiles. Let us first start with player 1 and analyze his strategic preferences. We first observe that

$v_1\big((\text{R,B}); b_1\big) \geq v_1\big(\text{L}; b_1\big)$ if and only if $b_1 \leq 1/2$. Furthermore the outcome (R,A) yields the least utility for player 1 for any b_1.

When player 1 plays R, he blames player 2 for playing A, because he would never play A if he were in player 2's position. This immediately follows from $v_{12}\big((\text{R,B}); b_1\big) = 2 + 5b_1 > v_{12}\big((\text{R,A}); b_1\big) = 0$. Hence player 1 would never choose A after history R in player 2's position.

Since player 1 blames player 2 at strategy profile (R,A), we need to understand how player 1's blame affects his strategic utility. Note that $\beta_1(\text{R,A}) = b_1 - f_1(5 + 2b_1)$ since $\delta_1^{(\text{R,A})} = ((5 + 2b_1) - 0)$. For simplicity, let us assume that $f_i(\delta_i^{(\text{R,A})}) = \delta_i^{(\text{R,A})}$ for all $\delta_i \geq 0$, and zero otherwise, for $i = 1, 2$. Then, for $b_1 \leq 1/2$, player 1's strategic utility from the strategy profile (R,A) is $u_1\big((\text{R,A}); b_1\big) = \pi_i(h_{(\text{R,A})}) + \beta_1((\text{R,A}))\pi_j(h_{(\text{R,A})}) = 0 + (b_1 - (5 + 2b_1)) \times 0 = 0$. In contrast if player 2's strategy is to choose action B at the history R, player 1's strategic utility is $u_1\big((\text{R,B}); b_1\big) = v_1\big((\text{R,B}); b_1\big)$ since B is a blame-free action.

Also observe that if player 1 chooses L, he does not blame player 2 for any strategy since what player 2 does does not affect his payoff; hence $u_1(\text{L}; b_1) = 4 + 4b_1$. If $b_1 \geq 1/2$, player 1's most preferred outcome is L. But player 1 does not blame player 2 for either playing A or B when he plays L.

The analysis of player 2's strategic preferences is more interesting. Note that if player 2 were in player 1's position his utility from outcome L is $4 + 4b_2$ while the outcome (R,B) is $5 + 2b_2$. Thus, for any $b_2 \geq 1/2$, player 2 would prefer L over (R,B) if he were in player 1's position. Consequently, player 2 blames player 1's strategy R only if $b_2 \geq 1/2$; so let us suppose that this is the case. In order to understand player 2's thought experiment assume that player 1 plays R. Player 2's line of reasoning is as follows. If I were in player 1's position I would play L. That would result in a payoff of $4 + 4b_2$ as opposed to $2 + 5b_2$ assuming I would react by B. Therefore, at strategy profile (R,B) I blame him by $\delta_2^{(\text{R,B})} = (4 + 4b_2) - (2 + 5b_2) = 2 - b_2$, which results in strategic utility

$$u_2\big((\text{R,B}); b_2\big) = 2 + 5(2b_2 - 2).$$

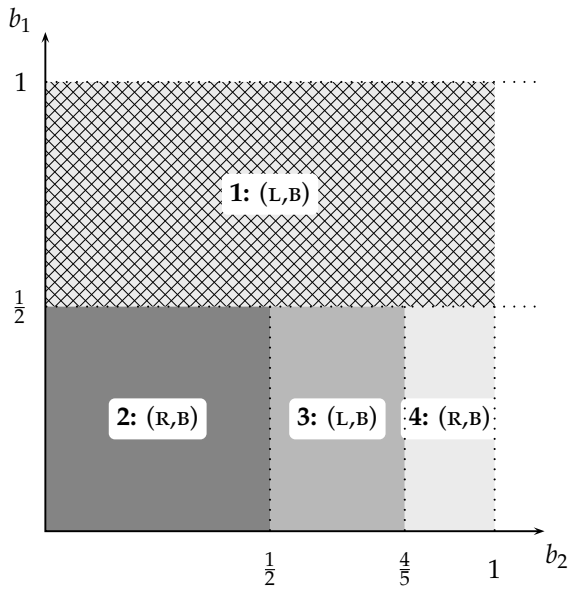
Hence, if $b < 4/5$, player 2 credibly plays A as a reaction to R since $u_2\big((\text{R,B}); b_2\big) < 0 = u_2\big((\text{R,A}); b_2\big)$.

Simple algebra shows that player 2 with a caring parameter $b_2 \in [1/2, 4/5)$ will play A as a response to R, and plays B as a response to L. However, for $b_2 \geq 4/5$ player 2 will play B in response to player 1's action R. This is true because while player 2 still blames player 1 for his choice of R, he cares so much about him that he refuses to punish him for playing R.

To illustrate these ideas consider Figure 3.

Figure 3 depicts equilibrium of games with different combinations of caring param-

FIGURE 3: CHARACTERIZATION OF SUBGAME PERFECT EQUILIBRIUM OF THE GAME IN FIGURE 2.



eters. The horizontal axis is b_2 and the vertical axis is b_1. In region 1, player 1 plays his dominant strategy L, and all types of player 2 plays B in the subgame. Hence the subgame perfect equilibrium of the game is (L,B) and it does not involve any blame. In region 2, in the subgame perfect equilibrium, player 1 plays R and player 2 plays B. This equilibrium does not involve any blame either simply because player 2 would have done the same thing if here in player 1's position. In region 3 player 2 blames player 1 if his action is R. As a result, player 2 prefers to play A in response to R, and b in response to L. Since player 1 prefers L over (R,A), in the equilibrium he plays L. In that scenario observe that there is no blame in the equilibrium. But it is the "blame" that makes player 1 play L. Finally in region 4, player 2 is altruistic enough ($b_2 > 4/5$) that he does not punish player 1's action of R, even though he actually blames player 1.

3 TOURNAMENTS: A SIMPLE MODEL

In this section we investigate the equilibrium effort level predictions of the tournament game that serves as the basis of our experimental design. We investigate the case where strategic utilities are determined by considerations of blame and players are strategic utility maximizers. We discuss and state our results in the body of the text and present the intuition of our argument, but do not burden the reader with formal proofs since it is punishment behavior and not effort levels that is the main focus of our blame-free hypothesis.

3.1 Uneven Tournaments without Blame

The experiment used to test our blame-free theory is one involving uneven tournaments with punishments. In the standard treatment of such tournaments, no punishments are used in the equilibrium since such punishments are not credible in the subgame perfect equilibrium. We then introduce blame into the analysis and demonstrate how the results predicted by the standard theory change.

An uneven tournament is a rank-ordered tournament where the cost of the exerted effort is different for at least one of the agents. Under our design, subjects have the chance to express their discontent with the outcomes of the tournament by reducing the payoff of the other player. For the punishment stage of the game we follow the linear punishment mechanism implemented by Fehr and Gaechter [11], Carpenter [3] and Nikiforakis and Normann [17], among others.

More precisely, the experiment involved two stages. In Stage 1 the subjects played an uneven tournament game identical to that used by SW. After the results of this tournament were known, they moved on to the punishment stage where they could use some punishment points, D, given to them to reduce the payoff of their opponent. Such punishments were costly to the subjects since any punishment points not used could be kept and converted into U.S. dollars. As is usual in these two-stage games with punishment, a sub-game perfect equilibrium does not involve any punishment.

In Stage 1 each player i chooses an effort level $e_i \in [0, \bar{e}]$, which generates an observable output

$$y_i = e_i + \epsilon_i,$$

where ϵ_i is the realization of a uniform random variable whose support is $[-a, a]$ for some $a > 0$. We assume that the random variables are identical and independent for the two players. Exerting effort is costly for the players: For player 1 the cost of effort level e_1 is $(e_1)^2/c$, whereas for player 2, the cost of effort e_2 is $\alpha(e_2)^2/c$, where $c > 0, \alpha > 1$. The output levels determine the payoffs of the tournament. If $y_i > y_j$ then player i receives M while player j receives $m < M$. Since the case $y_1 = y_2$ is a zero-probability event we omit this case.

In the second stage of the game, players are given some information about what occurred in Stage 1. In the experiment, the information given to them varies depending on the treatment, which we will explain later. The players are endowed with *D punishment points*. Once they receive the information about what happened Stage 1, they can use the punishment points to decrease the payoff of their opponent. Each point that i assigns to j (denoted as d_i^j) costs him one point and reduces j's payoff by h points.

Given an effort profile (e_1, e_2), player 1 wins the tournament if $\epsilon_1 - \epsilon_2 > e_1 - e_2$. We

denote this probability by $p(e_1, e_2)$ and given our uniform distribution assumption for ϵ_i we compute it as follows.

$$p(e_1,e_2) = \begin{cases} \frac{1}{2} + \frac{e_1-e_2}{2a} + \frac{(e_1-e_2)^2}{8a^2} & \text{if } -2a \leq e_1 - e_2 \leq 0, \\ \frac{1}{2} + \frac{e_1-e_2}{2a} - \frac{(e_1-e_2)^2}{8a^2} & \text{if } 0 \leq e_1 - e_2 \leq 2a. \end{cases}$$

We can readily write the expected payoff of an outcome where the effort levels are e_1, e_2, and punishments are d_1^2, d_2^1 as:

$$\begin{aligned} \pi_1\big((e_1,e_2),d\big) &= p(e_1,e_2)\mu + m + D - \left(\frac{(e_1)^2}{c} + d_1^2 + hd_2^1\right), \\ \pi_2\big((e_1,e_2),d\big) &= M - p(e_1,e_2)\mu + D - \left(\alpha\frac{(e_2)^2}{c} + d_2^1 + hd_1^2\right), \end{aligned}$$

where $\mu := M - m$.

We assume that players' utility functions are

$$v_i(h;b_i) = \pi_i(h) + b_i\pi_j(h), \qquad v_{ij}(h;b_i) = \pi_j(h) + b_i\pi_j(h).$$

Critically, since in this section we assume that the players are not motivated by considerations of blame, the strategic preferences are $u_i(\sigma;b_i) = v_i\big(h_\sigma;b_i\big)$ for $i = 1,2$, where $0 \leq b_i \leq 1$ is player i's caring parameter, reflecting his altruism for the other player. The equilibrium analysis of the game is quite straightforward. Let us denote $\phi_1 := \frac{c\mu}{8a^2}(1-b_1)$ and $\phi_2 := \frac{c\mu}{8\alpha a^2}(1-b_2)$ and state the equilibrium in the next Proposition:

PROPOSITION 1. *If strategic utility functions do not involve blame, then in the subgame perfect equilibrium of the game $d_i^j = 0$ for $i,j = 1,2$ and the effort levels are*

$$(e_1,e_2) = \begin{cases} \left(\frac{\phi_1}{1-\phi_1+\phi_2}2a, \frac{\phi_2}{1-\phi_1+\phi_2}2a\right) & \textit{when } \phi_1 \geq \phi_2, \\ \left(\frac{\phi_1}{1-\phi_2+\phi_1}2a, \frac{\phi_2}{1-\phi_2+\phi_1}2a\right) & \textit{when } \phi_1 \leq \phi_2. \end{cases}$$

Proof. See Fain [6] who treats that case where $b_i = 0, i = 1,2$. □

3.2 UNEVEN TOURNAMENTS WITH BLAME

When considerations of blame exist the analysis of the tournament becomes slightly more complicated and interesting. To start we assume that players' utility functions are

$$v_i(h;b_i) = \pi_i(h) + b_i\pi_j(h), \qquad v_{ij}(h;b_i) = \pi_j(h) + b_i\pi_j(h)$$

and the strategic preferences are $u_i(\sigma; b_i) = v_i\big(h_\sigma; \beta_i(\sigma)\big)$ for $i = 1, 2$, where $b_i \geq 0$, as before, is player i's caring parameter.

In our experiment we run treatments where only the disadvantaged subject can punish as well as ones where both the advantaged and disadvantaged subjects can punish. In the following two sections we analyze these cases.

3.2.1 One-Sided Punishment

In order to characterize the subgame perfect equilibrium for the one-sided punishment case we first go to the punishment subgame, look at the punishment behavior of player 2, and then incorporate player 2's punishment strategy into the effort choices of players in Stage 1. Since the game we investigate is one of complete information, player 2's punishment strategy can be anticipated and player 1 may decide to choose a lower effort level in an attempt to avoid punishment. Given the lowered effort of player 1, player 2 can be expected to increase his effort level in an attempt to increase his chances of winning. This is indeed the behavior in the equilibrium of the game when player 2 is more caring than player 1. This logic is summarized by the following Proposition:

Proposition 2. *The effort level of the advantaged player in a tournament with one-sided punishment stage is weakly less than the effort level of the same player in a tournament without punishment stage, while the effort level of his disadvantaged player is weakly greater.*

The intuition behind this result is clear. Consider the equilibrium defined in Proposition 1 where no blame exists and hence no punishment is forthcoming. At that equilibrium the players equate the marginal costs and benefits of increasing their effort levels and find an effort level that equates them. When blame and punishment are introduced, either that original equilibrium effort level for the advantaged subject calls forth punishment or it does not. If it does, then while the marginal cost of effort at the old equilibrium remains the same, the marginal benefit has decreased dramatically because every unit of effort exerted now calls forth a punishment. Given this decreased marginal benefit, the only way the advantaged subject can restore equilibrium is to decrease his effort (or at least not increase it) which is what Proposition 2 predicts. (If no punishment is forthcoming, then his actions will remain unchanged). For the disadvantaged subject, in the case where he punishes, either the advantaged subject decreases his effort or he does not. If he does not decrease his effort then the marginal costs and benefits of effort remain the same and he will not alter his effort away from no-blame no-punishment equilibrium. However, if punishment does alter the effort chosen by the advantaged subject in a downward direction, then the marginal benefit of increased effort (at the old equilibrium) has increased since his marginal probability of winning has increased (with

the marginal cost unchanged) and this will lead to an increase in the effort choice of the disadvantaged.

3.2.2 Two-Sided Punishment

The characterization of equilibrium efforts in the two-sided case is a generalization of Proposition 2.

Proposition 3. *When the equilibrium effort levels of players in a tournament with two sided punishment to one sided punishment are compared, either one of the following cases applies:*
1. the effort level of the advantaged player goes up while the effort level of the disadvantaged goes down,
2. the effort level of the disadvantaged player goes up while the effort level of the advantaged goes down,
3. both effort levels remain the same.

Although the main argument of the proof is similar to Proposition 2, the analysis is more tedious. The critical step in the equilibrium analysis is to understand the sources of blame. There are two reasons a player may blame his opponent. The first reason is the effort choices in the tournament stage of the game: if player i exerts more effort than player j would have exerted in the position of player i, then player j may blame player i and punish him. The second source of blame is *punishment*: a player may blame the other because of his punishment level and retaliate by punishing him back.

Note that given the utility functions that we assume, there is at most one player who blames the effort choice of his opponent. That is, if player i blames player j due to his effort choice, it cannot be the case that player j blames player i 's effort choice as well. Therefore, in any equilibrium there is at most one player who blames the other due to effort choices. This observation restricts the number of cases that can be observed in equilibrium. Consider the following three scenarios:

1. **Player i punishes player j for his effort but player j does not retaliate.** This scenario can take place for two reasons. Either player j does not blame player i's punishment, or although he blames player i's punishment, it is too costly for him to retaliate.

2. **Player i punishes j for his effort, player j retaliates.** Player i punishes player j. Player j finds player i's punishment blame-worthy and his blame justifies the cost of punishing player i back.

3. **Neither player punishes the other.** This scenario can be observed for four reasons: (a) neither player blames the other for his effort choice, (b) neither player's blame

does not justify the cost of punishment, (*c*) although player *i*'s blame is high enough to punish player *j*, he prefers not to punish him to avoid *j*'s retaliation, (*d*) a player adjusts his effort level to avoid punishment, hence the equilibrium does not exhibit any punishment.

Scenario 1 is identical to the one-sided punishment case covered by Proposition 2 since it does not matter whether *j* is not allowed to punish (as in the one-sided case) or simply chooses not to. Note, however, that player *j* can be either an advantaged or a disadvantaged player. If he is advantaged, then we get the same result as Proposition 2 where the advantaged players efforts weakly decrease and the disadvantage weakly increase. If *j* is disadvantaged, then the opposite result holds.

In Scenario 2 assume that player *j* is punished for his effort level. This will reduce his effort below what it would be in the no punishment case. Now say *j* retaliates against *i* and this retaliation reduces *i*'s punishment. This will increase *j*'s effort but his effort can never increase past the no-punishment level. Depending on whether *i* was advantaged or disadvantaged, the effort levels of *i* and *j* will either increase or decrease.

In Scenario 3 if reasons (*a*)-(*c*) hold then we are simply back in the no-punishment equilibrium. If reason (*d*) holds, then that player who anticipates punishment will lower his effort level to avoid it. This will create an equilibrium where his effort level is lower while the other player increases his effort, yet no punishment will be observed.

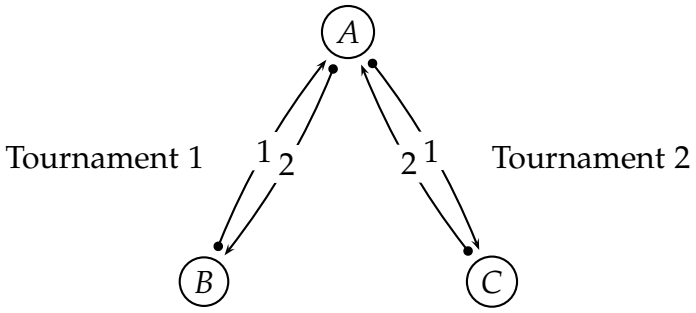
4 Experimental Design

In order to assess whether individuals judge other individuals' behavior using their own behavior as a reference (i.e., whether they use the blame-free thought experiment) we need to know what each subject would do in both possible roles. A novel feature of our design is that each subject plays in both roles simultaneously. Hence, subjects played in both roles in each round in two different and independent tournaments.

Let us clarify the setup. Let *A*, *B* and *C* be three subjects, and player 1 and player 2 be two different roles in the tournament. Subject *A* plays one tournament (Tournament 1) in player 2 role when he is matched with subject *B* who is in player 1 role, and another one (Tournament 2) in player 1 role when he is matched with subjects *C* who is in player 2 role. Figure 4 depicts this scenario. All the treatments in our experiment are run following this structure.

The basic structure of the games implemented in the experiments follows from Subsections 3.2.1 and 3.2.2. Recall that player 1 and player 2 differ in two dimensions. First, they have different costs of effort in such a way that player 1 is *advantaged* and player

FIGURE 4: MATCHING SCHEME IN A TYPICAL ROUND.



2 is *disadvantaged*. Second, disadvantaged player can punish advantaged player in the one-sided case (Subsection 3.2.1), while the opposite is not true. In the two-sided case (Subsection 3.2.2) however, both players can punish each other.

As we explained before (Figure 4), in each round of our experiment, each subject is matched with two different subjects (opponents), therefore being involved in two independent tournaments: one where he is in the advantaged role and the other where he is in the disadvantaged role. Subjects are supposed to make their decisions for both tournaments *simultaneously*.

An experimental session lasts 34 rounds under different matching protocols, namely fixed matching (F) and random matching (R). Under R subjects are randomly and independently matched at the beginning of each round, whereas under F subjects are matched with other subjects and this match stays fixed.

In some treatments the first half (rounds 1-17) of the session is under F (R), while second half (rounds 18-34) is under R (F). In other treatments the entire session (rounds 1-34) is either F or R. In sum we have four matching schemes: FR, RF, FF, and RR, where, for instance, FR means rounds 1-17 are F and rounds 18-34 are R.

In treatments where matching scheme is FR or RF, when subjects enter the lab they are told that the experiment is divided in two different parts, with the first one lasting for 17 rounds. Only when the first part is over do the subjects receive instructions for the second part of the session.

The final defining factor of our design concerns the information that the subjects are given between the tournament and punishment stages of the game. We have two information regimes. In the *low information* regime (*l*), after the subjects choose effort in the tournament stage but before they make their punishment decision, they are given the information only about the effort choices of their opponents in the tournament. We also have an *high information* regime (*h*), where the subjects are given all the information about the tournament before the punishment stage. That is, in high information treatments subjects are told the winner of the tournament, payoffs earned in the tournament as well

as the effort levels. This choice of design allows us to test our hypothesis cleanly since we can compare subjects' responses to their opponent's effort choices as well as their responses to outcomes that involve random factors beside effort choices.

Overall, we have four different treatments: 1-FR-*ll*, 1-RF-*ll*, 2-FF-*lh*, and 2-RR-*lh*. The notation is self explanatory. For instance 2-FF-*lh* means that the treatment involves two-sided punishment, the matching protocol is fixed in both halves of the session and, while in the first part of the experiment the subjects observed efforts before the punishment stage, in the second part of the experiment they observed the outcome payoffs of the tournament as well as the efforts.

We chose the tournament's parameter values following SW. Particularly, subjects' effort levels were limited to the interval $[0, 100]$ ($\bar{e} = 100$.) We set $\alpha = 2$ and we say that a subject is in the disadvantaged role if the cost of effort is given by $c(e) = \alpha e^2/150$ ($c = 150$), otherwise we say that subject is in the advantaged role. The random shocks that determine the outcome of the tournament lie in the range $(-60, 60)$ ($a = 60$). The prizes of the tournaments are $M = 204$ and $m = 86$. Additionally, we endowed each player with $D = 68$ punishment points when they were in the disadvantaged role. Every punishment point assigned to his opponent reduced his payoff by 3 punishment points. Every punishment point he kept was converted into US Dollars at the rate 1 point=\$0.0015, while the exchange rate for each point earned in the tournament was \$1/322 points ($h \approx 1.45$ by exchange cross-rate.)

All the sessions were conducted at the Center for Experimental Social Science lab at New York University with a total of 68 participants, all students at that university. The experiment was computerized using the software z-Tree (Fischbacher [13]). Sessions lasted for about one hour and a half and participants received an average payment of \$24 dollars. All parameter values and procedures were common-knowledge. Table 1 presents our experimental design.

5 Results

In this section we will present our results by answering a set of questions generated by our theory.

5.1 Question 1: Do Subjects Punish when They Should According to the Theory of Blame-Freeness?

In the context of our experiments, blame-freeness is, first and foremost, a theory of punishment or at least reciprocity. Hence, we will begin by investigating punishments. As

Table 1: Experimental design.

Treatment	Matching Protocol	Information Regime
1-FR-*ll*	Fixed: rounds 1-17	Low: rounds 1-17
	Random: rounds 18-34	Low: rounds 18-34
1-RF-*ll*	Random: rounds 1-17	Low: rounds 1-17
	Fixed: rounds 18-34	Low: rounds 18-34
2-FF-*lh*	Fixed: rounds 1-17	Low: rounds 1-17
	Fixed: rounds 18-34	High: rounds 18-34
2-RR-*lh*	Random: rounds 1-17	Low: rounds 1-17
	Random: rounds 18-34	High: rounds 18-34

we have stated before, while blame-freeness is a sufficient condition for no punishment, blameworthiness is only a necessary condition. Put differently, when a subject chooses more effort in the advantaged role than his advantaged opponent does against him while he is in the disadvantaged role, then our theory predicts unambiguously that we should see no punishment forthcoming. However, when a subject in the same position chooses less effort, then whether we see punishment occurring will depend on the cost of punishment, its impact as a deterrent to the advantaged subjects, and the caring parameters of the subjects' utility functions. If the cost of punishment is sufficiently high, the blame sufficiently small or the benefits sufficiently low in the eyes of the subjects, then we would not expect to observe any punishments.

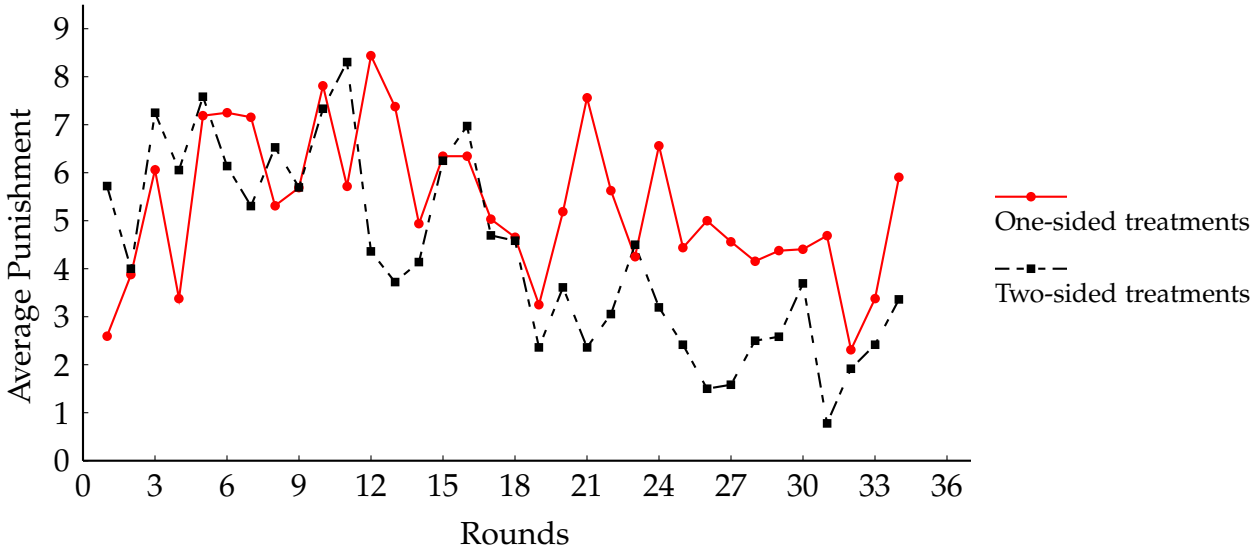
In most of what we do below we will concentrate on the punishing behavior of the disadvantaged subjects. This makes more sense since strategically they are the ones for whom blame and punishment is most natural since they can blame their advantaged cohort for exploiting their advantaged position by choosing too high an effort level. In our one-sided punishment treatment, only the disadvantaged were allowed to punish so our focus on disadvantaged subjects is certainly natural there. While the advantaged subjects in our two-sided punishment treatment can also blame their disadvantaged cohorts by comparing what the disadvantaged subjects chose with what he would have chosen if he were disadvantaged, we think that comparison is less interesting. Still, at the end of this section we will investigate the punishing behavior of the advantaged subjects as well where indeed we will find some surprising results.

5.1.1 Punishment Levels: Disadvantaged Subjects

Before we look into how closely the punishing behavior of our disadvantaged subjects conformed to our theory, let us describe the amount of punishment existing in the exper-

iment in general. Figure 5 presents the average levels of punishment made by subjects in the disadvantaged role in both the one-sided and two-sided punishment treatments over their 34 rounds history. What is worth noting here is that mean punishment levels are clearly above zero so that punishment was an alternative employed by at least some disadvantaged subjects. Further, note that these are the punishments sent by the punisher and that the punishments received were three times this amount.

FIGURE 5: AVERAGE PUNISHMENT IN ONE-SIDED AND TWO-SIDED TREATMENTS BY THE DISADVANTAGED PLAYERS.



5.1.2 ADHERENCE TO THE BLAME-FREE THEORY

Testing for adherence to our theory, as we have indicated above, is slightly tricky since blameworthiness is only a necessary condition for punishment. However, when one's advantaged opponent acts in a blame-free manner, then that is sufficient for non-punishment and this is what we will look at first. To this end, empirically, we say that a disadvantaged subject blames his advantaged opponent if his opponent's effort level is higher than the effort level he actually chooses when he is in the advantaged position in the other tournament. With respect to blame-free behavior of one's advantaged opponent, we see punishment behavior consistent with our theory 81% of the time. In other words, when disadvantaged subjects faced advantaged opponents who chose lower effort levels than they did when placed in their position, they decided not to punish them 81% of the time. More precisely, of the 1,215 observations where there shouldn't be punishment,

we actually observe no punishment in 983 cases.[9] This clearly implies that subjects refrained from punishing their advantaged opponents when those opponents acted in a more kind manner than they would if they were in their position.

Taking punishment behavior into account and not just non-punishing behavior, we see that of the subset of subjects who punished at least once, 73% of them exhibited behavior that did not violate the blame-free theory at least 50% of the time. This means they punished only when there was blame and failed to do so when there was none. More precisely, of the 68 subjects in our experiment 31 never punished, leaving 37 who did at least once. Such subsets of non-punishers are seen in almost all experiments where punishing exists since subjects are hesitant to punish their fellow subjects in almost all laboratory experiments (see, for example, Fehr and Gaechter [11] where punishment levels are quite low). In the end-of-session questionnaires distributed subjects report two main reasons for this behavior: the fact that punishment is costly and their unwillingness to hurt their opponents. Both of these reasons are consistent with our theory since the existence of blame is only a necessary condition for punishment and the cost of punishment is a reason not to do so. However, caring about one's opponent means that subjects have high caring parameters, b_i's, and hence it takes a lot of blame to force them to punish. In fact, it is not clear how much punishment we should expect in our experiments since subjects with low caring parameters, self-centered individuals, are unlikely to blame their fellow competitors and therefore unlikely to punish them despite their low b_i in their strategic preference function, while those with high caring parameters are likely to blame their opponents but, since they care so much about them, unlikely to punish them. Hence, the level of punishment activity cannot easily be predicted but, as we have done, we can check for its theoretical consistency when it exists.

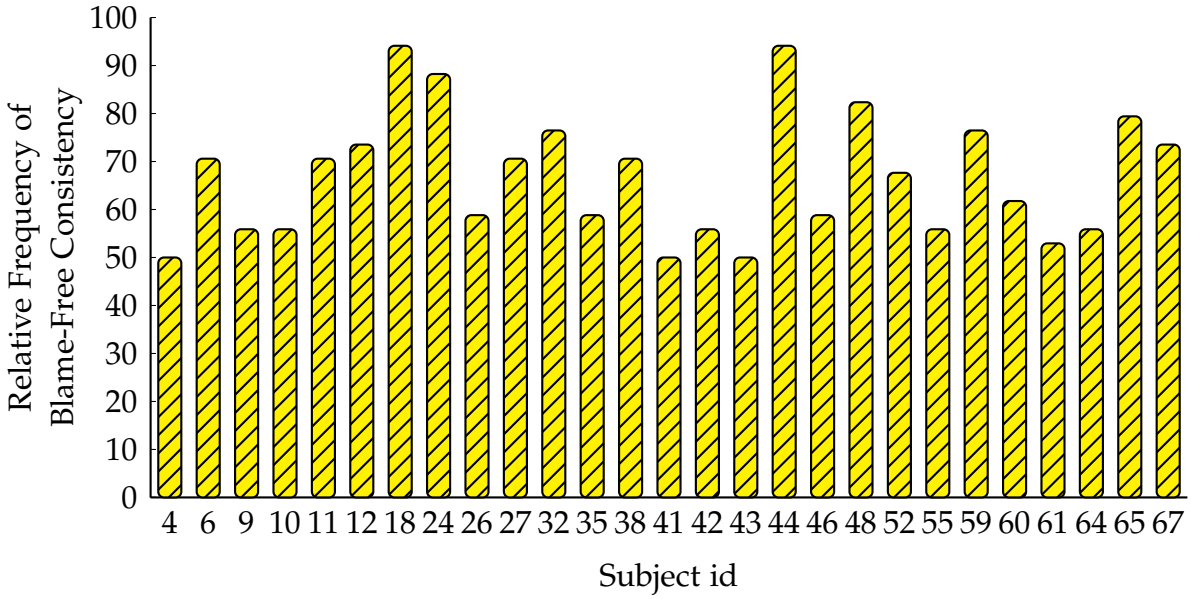
Figure 6 indicates the consistency of subject behavior to our theory by that subset of 37 subjects who punished at least once and who were consistent with our theory at least 50% of the time.

Notice that the mean adherence is 68% with two subjects adhering as much as 94% of the time.

One cannot judge exactly whether this is strong or weak support for our theory because not punishing blameworthy opponents may still be consistent with blame-freeness as long as the caring parameter, b_i, is high enough and the perceived costs are large enough. Still, our data indicates that there is a non negligible part of the sample that be-

[9] There are subjects who never punish. If we look only at those who punish at least once, then that figure drops to 63% (635 observations where there was no reason to punish and 403 observations with no punishment). Still, subjects who never punish are consistent with our theory since they may simply have very low (or very high) caring parameters or punishment may be costly. Hence the 81% figure may be more accurate.

FIGURE 6: RELATIVE FREQUENCY OF BLAME-FREE CONSISTENCY FOR PUNISHERS WHO ARE CONSISTENT MORE THAN 50% OF THE TIME: DISADVANTAGED SUBJECTS.



haves according to blame-free principles when it comes to judge other person's behavior.

To get a better insight into whether disadvantaged subjects punished when they were supposed to according to the theory of blame-freeness, note that the observed behavior of subjects would support it if subject j in the disadvantaged role punishes his opponent subject i whenever j's effort in the advantaged position is lower than i's effort in that same role. Of course these circumstances are only necessary conditions for punishment so requiring punishment when blame exists is a very strict test. Despite this fact, ceteris paribus, if our blame-free theory has teeth, it should be true that the probability of punishment is increasing in the amount of blame and that is a consequence we can estimate. To do this we define the variable Δ as difference between subject j's effort when he is advantaged and i's effort when he is advantaged (while j is disadvantaged). Since blame is decreasing in this difference, we would expect that the probability of punishment would decrease with the variable Δ and hence the coefficients associated with this variable should be negative.

The first column in Table 2 shows the probit regression for the likelihood of punishing in the disadvantaged role as a function of Δ. We denote the dummy variable that takes value 1 when a disadvantaged subject punishes by $D_{\text{p\_dis}}$. Note that the coefficient is negative and significant, suggesting that the larger the deviation of i's behavior with respect to j's behavior (the more negative the variable Δ is) the more likely j will punish i. In other words, as i's behavior gets closer to j's behavior (in i's role), it is the less likely that j will punish i. Given that we are not interested in the cases where Δ is positive,

to check the robustness of our results we ran a probit using a variable Δ^+, which is the variable Δ truncated at zero to the right. The second column reports this regression. The previous result holds, though significance level drops to 10%.[10]

TABLE 2: PROBIT REGRESSIONS FOR PUNISHMENT IN DISADVANTAGED ROLE.

	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$
Δ	$-.004$	-	-
	$(.002)^{**}$		
Δ^+	-	$-.005$	-
		$(.003)^{*}$	
Δ_{past}	-	-	$-.007$
			$(.002)^{***}$
constant	$-.741$	$-.808$	$-.752$
	$(.169)^{***}$	$(.177)^{***}$	$(.175)^{***}$
N	2,312	2,312	2,244

Robust standard errors in parenthesis, clustered by group in the Fix Matching rounds and by Session in the Random Matching ones.
* Significance at 10% level. ** Significance at 5% level. *** Significance at 1% level.

For a further test we define a weaker notion of blame-free justice, which assumes that if subject j was in i's position in the past, he will use his own past behavior when he judges i's actions. Thus, we construct the variable Δ_{past} that takes the difference between j's average effort in the advantaged role in all previous rounds and i's effort in the current round. As shown in the third column of Table 2, this variable performs better than the previous ones in explaining the likelihood of observing punishment.[11] This result indicates that subjects may not be comparing the effort of their opponent in a round to their effort level in that round, but rather to the mean level of their effort in all rounds up to the current round. This slight generalization yields better econometric

[10]Unless otherwise indicated, significance levels of the regressors used in this paper are calculated with robust standard errors, where the clustering group is the relevant group of interaction. The Z-tree program matched subjects into groups of 4 participants and then assigned them to different tournaments. So even when subjects did not know it, they were interacting in groups of 4. For the fixed matching rounds, the variable group is used as the relevant group of interaction. For the random matching rounds or sessions the session is considered as the relevant group of interaction for the clustering.

Reported results are robust to cluster by "session" in the session that started with random matching and then switched to fixed matching protocol. We did this in order to control for possible behavioral spill-overs from the first to the second half of the session.

[11]In this regression we have a lower number of observations because the variable Δ_{past} uses at least one lag, then the first round observation is lost. In order to check if the increase in the significance level of the variable is due to this fact we ran a probit regression using the variable Δ as a regressor for the observations of rounds 2 to 34. The significance level of the variable Δ does not change with respect to the regression reported in the first column of the table.

results.

5.2 Question 2: Does Blame-Freeness Explain Our Data Better Than Competing Theories?

Blame-freeness is different from some other theories of justice (e.g. Fehr and Schmidt [12] or Bolton and Ockenfels [2], utilitarianism, Rawlsian justice) because it is a process theory as opposed to an end-state theory. Put differently, blame-freeness cares about the actions taken to determine an outcome and not merely the outcome itself. (Remember a subject's utility depends on his opponents strategy as well as his material payoff at the terminal node of the game). Hence, two outcomes with the same payoff vector can yield a different utility to a player if these outcomes were determined by different strategies, i.e., strategies that differed in their blameworthiness.

In this subsection we explore whether the prevalent fairness theories in the literature can explain punishment behavior better than the blame-free hypothesis. The inequality aversion theories assume that individuals resist outcomes that deviate from the equal split of the surplus without evaluating the actions that led to the resulting distribution. In Fehr and Schmidt [12]'s version of this theory, the way that subjects evaluate the final distribution is by comparing the payoffs that each individual got with their own payoff. Loosely speaking, in the game implemented in this paper, a Fehr and Schmidt [12] individual would mainly care about the net payoff that each agent got from the tournament (i.e. tournament prize minus cost of effort) and based on that comparison he would decide to punish his opponent or not. However, if blame is a motivation for punishment, then when an opponent behaves in a nasty manner he should be punished whether or not his nasty behavior caused damage to his recipient. In the context of our tournament game, a strict theory of blame would indicate that if an opponent took a blameworthy action but lost the tournament, he is still a candidate for punishment.

To investigate which theory of justice is operating in our experiment we investigated the following probit regression using the data from the *high information* treatments where before the punishment stage subjects could see not only what their opponents' effort choices were but also what their payoffs were, and whether they won or lost the tournament in that round:

$$\Pr(D_{\text{p}} = 1) = \alpha + \beta_1 \Delta + \beta_2 \Delta_\pi + \beta_3 D_w + \beta_4 (D_w \times \Delta_\pi) + \beta_5 \Delta_{\text{past}}$$

where D_{p} is a dummy variable that takes value 1 if there is punishment by a subject.[12]

[12] When we distinguish advantaged and disadvantaged subjects we write $D_{\text{p\_adv}}$ and $D_{\text{p\_dis}}$ respectively.

Δ is our now familiar difference variable which indicates blame when it takes a negative value and no blame otherwise, Δ_π is the difference in payoffs between the players in the tournament stage, D_w is a dummy variable indicating whether the disadvantaged subject won the tournament or not, and Δ_{past} is the difference between one's advantaged opponents effort this period and the average of a disadvantaged subject's past effort choices when playing the advantaged role. Clearly, if blame is the main motivation for punishment we would expect that the coefficient associated with the Δ (or Δ_{past}) would be significant and all other coefficients to be insignificant. If inequality aversion was important then we should see a significant coefficient for Δ_π variable. (We also ran the regression for subjects in the advantaged role in our two-sided punishment treatment which we will discuss later).

Table 3 presents the results for the probit regression for both subjects in the disadvantaged and advantaged roles. After testing for differences in behavior across treatments we pool all observations.

Table 3: Probit regressions for punishment: two-sided treatments – disadvantaged and advantaged subjects.

	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$	$D_{\text{p\_adv}}$	$D_{\text{p\_adv}}$
Δ	$-.008$	-	$-.008$	-
	$(.002)^{***}$		$(.002)$	
Δ_{past}	-	$-.017$	-	$-.017$
		$(.005)^{***}$		$(.002)$
Δ_π	$-.006$	$-.006$	$-.002$	$-.002$
	$(.005)$	$(.005)$	$(.001)^{*}$	$(.011)$
D_w	1.583	1.674	.550	.584
	(1.195)	(1.120)	(.455)	(.460)
$D_w \times \Delta_\pi$	$-.002$	$-.002$	$-.006)$	$-.007$
	$(.002)$	$(.002)$	$(.005)$	$(.005)$
$D_w \times \Delta$	$-.001$	.002	$-.004$	-
	$(.003)$	$(.004)$	$(.002)^{**}$	
$D_w \times \Delta_{\text{past}}$	-	.002	-	$-.008$
		$(.004)^{**}$		$(.004)^{**}$
constant	-1.804	-1.918	$-.868$	$-.863$
	$(.701)^{***}$	$(.627)^{***}$	$(.332)^{***}$	$(.343)^{***}$
N	612	592	612	592

Robust standard errors in parenthesis, clustered by group.
* Significance at 10% level. ** Significance at 5% level. *** Significance at 1% level.

Looking first at the estimates for disadvantaged subjects (first two columns) this

regression offers strong support for our theory. It indicates that when subjects were offered information about payoffs differences, Δ_π, (an end-state distributional variable) and effort differences, Δ, (a process variable that indicates blame) they focus on the blame variable to the exclusion of the payoff difference. Such a result is not consistent with behavior under a theory of inequality aversion. In addition, note that punishment is not affected by whether a disadvantaged subject won the tournament or not. In other words, if your opponent chooses a blameworthy effort level but it happens to turn out that you win the tournament and hence no damaged was caused, that fact does not change your punishment behavior—it appears as if subjects punish the strategy and not its consequences.

It is slightly more difficult to separate our theory from those of Rabin [18], Dufwenberg and Kirchsteiger [5], and Falk and Fischbacher [10] (which we will refer to as Rabin et al.) but, given our data, the two theories do make different predictions for when the disadvantaged subjects should punish. While both theories predict punishment when one's opponent does something nasty (or their inferred intensions were nasty), they differ as to what nasty means. For Rabin et al., in the context of our tournament game, an advantaged opponent is nasty if he chooses a too large effort, i.e., one that decreases the probability of winning for his disadvantaged opponent unfairly. So nastiness is an increasing function of the difference between the effort levels of the advantaged and disadvantaged subjects since that translates directly into a difference in the probability of winning. On the contrary, in our theory an advantaged subject is nasty if he chooses a level of effort which is higher than what his disadvantaged opponent would do if he were in his shoes. To capture which type of nastiness is operational in our data we ran a probit regression where the left-hand variable was the usual dichotomous variable indicating punishment or not ($D_{\text{p\_dis}}$) while the right-hand variables were our difference variable Δ described above (which captures blame) and the difference in the probability of winning given the effort levels of the subjects (Δ_{prob})—which attempts to capture the Rabin et al. notion of nastiness. The results of this regression are presented in Table 4 below.

As can be seen, the only significant coefficient is Δ, which captures our notion of blame. In other words, it appears as if the motivation for punishment is best explained by a subject placing himself in the place of his opponent and performing a counterfactual thought experiment rather than simply noticing that one's opponent chose a high (and therefore nasty) effort level. In short, an advantaged opponent's action is only nasty if it was something that the disadvantaged subject would have not done if he were in that postilion. Absolutely high effort levels are not, per se, nasty nor are big differences in effort levels between advantaged and disadvantaged subjects.

Table 4: Probit regression for punishments in disadvantaged role: Rabin et al. vs. blame.

	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$
Δ	$-.005$	-
	$(.003)^{**}$	
Δ_{past}	-	$-.012$
		$(.003)^{***}$
Δ_{prob}	$-.130$	$-.454$
	$(.246)$	$(.275)^{*}$
constant	$-.707$	$-.609$
	$(.222)^{***}$	$(.241)^{***}$
N	2,312	2,244

Robust standard errors in parenthesis, clustered by group in the fixed matching rounds and by sessions in the random matching ones.
* Significance at 10% level. ** Significance at 5% level. *** Significance at 1% level

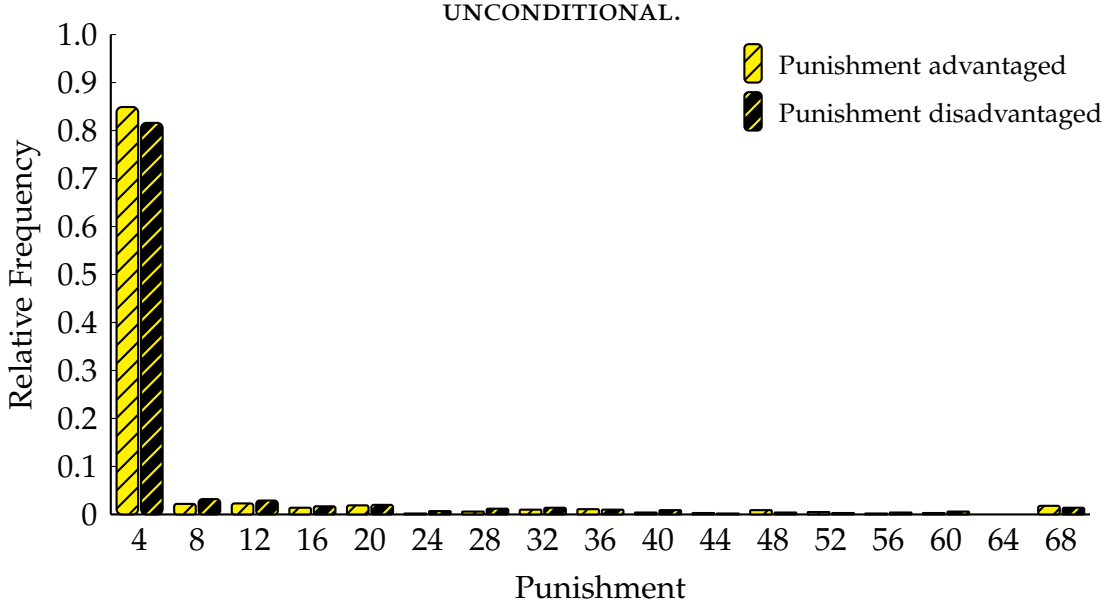
5.2.1 Punishment Behavior: Advantaged Subjects

Up until now all of our analysis of the blame-free hypothesis centered around the behavior of the disadvantaged subject. However, it is possible that the advantaged subject, despite the fact that he is advantaged, might still blame his disadvantaged subject for his actions and punish him. What we will attempt to establish below is that while the punishments of advantaged and disadvantaged subjects appear similar, advantaged subjects appear to punish for different reasons than disadvantaged subjects.

To get a first glimpse into the behavior of advantaged subjects, consider the histograms in Figure 7. This figure presents the frequency of different punishment levels by advantaged and disadvantaged subjects in our two-sided punishment treatment, the only sessions where the advantaged subjects could punish. As we see there does not seem to be any significant difference in the incidence or distribution of punishment between advantaged and disadvantaged subjects. More precisely, a Kolmogorov-Smirnov test indicates that these two distributions are not different ($p = 1.00$). If we restrict ourselves to the punishment behavior of advantaged and disadvantaged subjects when they punish only blame-worthy acts, then again we see that there is no difference between the punishment behavior of these two groups ($p = 1.00$). In other words, if punishment behavior is different between these two groups it is not due to different rates of punishment, but rather the circumstances under which they punish.

To investigate whether our blame-free theory explains the punishment behavior of advantaged subjects, we reran our probit regression that was previously run for disad-

Figure 7: Punishment by subjects in advantaged and disadvantaged roles: unconditional.



vantaged subjects, on advantaged subjects. This probit regression explains the probability of punishment as a function of our Difference variable. The results of this probit regression is presented in Table 5.

Table 5: Probit regressions for punishment in advantaged role.

	$D_{\text{p\_adv}}$	$D_{\text{p\_adv}}$	$D_{\text{p\_adv}}$
Δ	$-.001$	-	-
	$(.001)$		
Δ^+	-	$-.003$	-
		$(.003)$	
Δ_{past}	-	-	$-.001$
			$(.001)$
constant	$-.951$	$-.989$	$-.961$
	$(.328)^{***}$	$(.328)^{***}$	$(.339)^{***}$
N	1,224	1,224	1,118

Robust standard errors in parenthesis, clustered by group in the fixed matching rounds and by session in the random matching ones.
* Significance at 10% level. ** Significance at 5% level. *** Significance at 1% level.

Note that while in the same regression run for disadvantaged subjects (Table 2) all three variables were significant at least at the 10% level, the only significant variable here in any of the three regressions is the constant term. In other words, blame-freeness does not seem to do a good job of explaining behavior amongst our advantaged subjects.

There may be a number of reasons for this asymmetry. For example, advantaged subjects may have a very different blame function than disadvantaged subjects since they are in a favored position. Further, recall that in the experiment subjects were randomly assigned to their positions. Hence, an advantaged subject may not feel entitled to his advantaged position and may not blame his disadvantaged opponent for choosing a high effort and hence, may not wish to blame him for doing so. This same logic would have the opposite effect on disadvantaged subjects since any excessive effort by advantaged subjects may be viewed as unfair since disadvantaged subjects may feel that their advantaged opponents are not entitled to their advantage and therefore, not entitled to exploit it by choosing a high effort level. In other words, the technology of blaming changes when one is in an entitled role especially, in one which he is randomly placed.

Despite the fact that blame does not appear to explain punishment behavior well for advantaged subjects, it is still true that such subjects do punish their disadvantaged cohorts and so the question remains as to what exactly explains their behavior. Returning to Table 5, we see that while in the high information treatment, the only variable that was significant in explaining the punishment behavior of disadvantaged subjects was our Δ variable. When we look at the same regression run for advantaged subjects on the right hand side of Table 5, we see that these subjects seemed to care about the payoff differences Δ_π and the interaction term associated with winning (D_w) and Δ variable. The fact that the interaction variable has a negative coefficient associated with it indicates that advantaged subjects appear to punish according to blame-free principles only when they lose, which is interesting since losing seems to spark a blaming reaction. In summary, advantaged subjects seem to focus on different variables than disadvantaged subjects do to guide their punishment behavior, variables that are not entirely consistent with our blame-free hypothesis or variables that are consistent in a more indirect way.

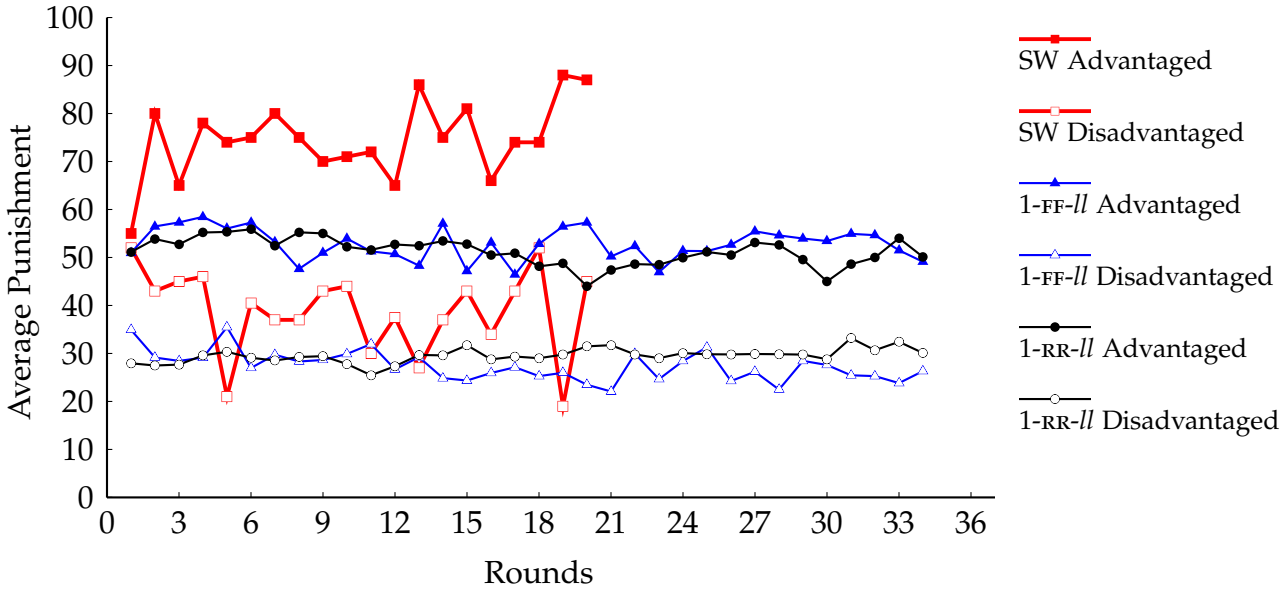
5.3 Question 3: Do Subjects Make Their Effort Decisions in a Manner Consistent with the Theory of Blame?

Given our Proposition 2 we would expect that, in comparison to a treatment where no punishments were allowed, the effort levels of the advantaged subjects would fall and those of the disadvantaged subjects would rise when one-sided punishment is introduced. This is so because an advantaged subject, fearing retaliation from a punishing (and blaming) disadvantaged subject, may decide to cut back on his effort in order to ward off costly punishments. In response, the disadvantaged subject then takes advantage of the reduced effort of his advantaged opponent and increases his effort. How large these changes are will depend on the distribution of caring parameters (the b_i's)

and, since these are not induced in the experiment, we can only make qualitative predictions here. Still we expect to see effort levels of the advantaged subjects fall and those of the disadvantaged subjects increase when one-sided punishments (and two-sided punishments as well) are introduced.

Data relating to this question can be seen in Figure 8 where we present the round-by-round mean effort choices of subjects in our one-sided punishment treatments as well as those same means for the SW experiment where no punishment was permitted.[13]

FIGURE 8: MEAN EFFORT LEVELS ONE-SIDED TREATMENT FIXED MATCHING, RANDOM MATCHING, SW.



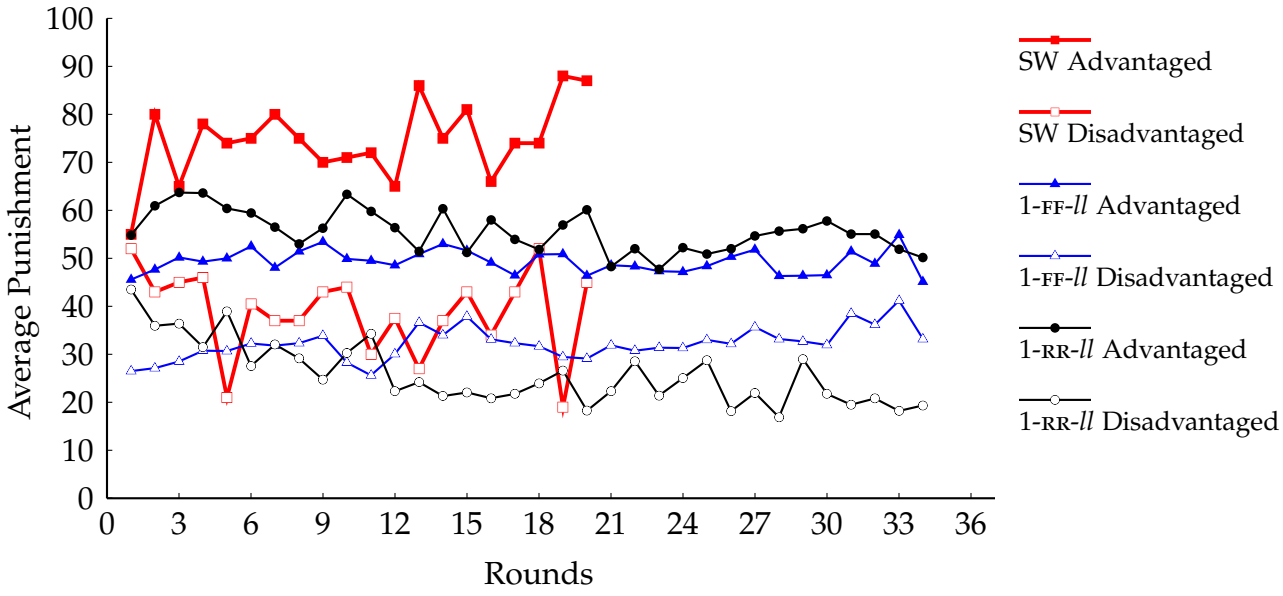
As we can see, the mean effort of advantaged subjects changes dramatically when punishments are made available and in the direction of the theory, but the same is not true for disadvantaged subjects. More precisely, while there is a dramatic drop in effort levels from the no-punishment treatment among the advantaged subjects, the effort levels of the subjects playing the disadvantaged role is actually lower. This latter result is true because in the no-punishment experiment of SW, disadvantaged subjects over exerted effort dramatically in an attempt to match the effort levels chosen by the advantaged subjects and, in essence, created a rat race that was eliminated when punishment was allowed.

In our two-sided punishment treatment (see Figure 9) we find the same results. Again

[13]Note that the SW experiments lasted only 20 rounds as compared to the 34 rounds in our experiments here. In addition, note that we have sorted our data by matching regime so that the—line represents the means of the subjects under the random matching regime while the—line presents those of the subjects under the fixed matching regime.

the efforts of the advantaged subjects decrease when the possibility of punishment exists and those of the disadvantaged subjects drop as well. This is true whether we have *effort* or *outcome* treatment or *fixed* or *random* matching, except in the case of *outcome* and fixed matching where for advantaged subjects we see a small over exertion of effort compared to the no punishment theoretical benchmark. Meanwhile for disadvantaged subjects there is no significant change.

FIGURE 9: MEAN EFFORT LEVELS ONE-SIDED TREATMENT FIXED MATCHING, RANDOM MATCHING, SW.



In summation, when comparing the behavior of subjects with and without punishment possibilities, it appears as if the qualitative predictions of our blame-free theory are born out. This is seen in a dramatic fashion when comparing the effort levels of the advantaged subjects in this experiment to those in SW. With respect to disadvantaged subjects, the results differ since when comparing effort levels with and without punishment the effort levels of the disadvantaged subjects actually decreases rather than increases, as expected. The lack of conformity in the behavior of our disadvantaged subjects is easily explained, since in this game without punishment, the subjects were involved in a high effort rat race where they drastically over exerted effort. When punishment was instituted this reduced the effort levels of the advantaged so much that there was no need for the disadvantaged subjects to increase their efforts and hence, in comparison to the no punishment case, effort levels dropped for the disadvantaged subjects.

5.4 Treatment Effects: Does Random Matching Have an Effect on Punishment Behavior?

As you will recall in our experimental design, we vary the matching protocol both within sessions and across sessions. While in one session subjects interacted in the same pairs for the first half of the session and were randomly re-matched to new opponents every round of the second half of it, in the other session the sequence was reversed. Mann-Whitney test for punishment behavior in these two sessions does not allow us to reject the hypothesis that both samples come from the same distribution. We can then conclude that the sequence of the matching protocol does not alter punishment behavior.

To substantiate this fact we ran a simple probit regression where the dependent variable was the dummy $D_{\text{p\_dis}}$, and the right-hand side variables were our difference variables, Δ, Δ^{+} (truncated difference), and Δ_{past} (mean of past differences), dummy variable D_{R}, indicating whether it is random matching or not, and an interaction term for these two variables. The results are presented in Table 6.

Table 6: The impact of random matching.

	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$	$D_{\text{p\_dis}}$
D_{R}	$-.010$	.031	.039
	(.345)	(.358)	(.367)
Δ	$-.005$	-	-
	(.002)		
Δ_{self}	-	$-.009$	-
		(.003)	
Δ^{+}	-	-	$-.007$
			(.004)
$\Delta \times D_{\text{R}}$	.003	-	-
	(.003)		
$\Delta_{\text{self}} \times D_{\text{R}}$	-	.004	-
		(.004)	
$\Delta^{+} \times D_{\text{R}}$	-	-	.004
			(.005)
constant	$-.738$	$-.736$	$-.827$
	(.195)	(.196)	(.184)

Robust standard errors in parenthesis, clustered by group in the fixed matching rounds and by session in the random matching ones.
* Significance at 10% level. ** Significance at 5% level. *** Significance at 1% level.

As we can see, the dummy variable D_{R} is never significant in any regression. This indicates that the motivation for blame (and punishment) is independent of whether the

subjects is matched repeatedly with the same subject. If blame is at work it is a type of disembodied blame that is not attached to a person but rather to the act taken even by an anonymous other.

6 Conclusions

This paper was motivated by the thought that if we are to understand reciprocity as a reward for kind behavior and punishment for unkind behavior then we will need an operational definition of what kindness means in strategic situations. This paper provides such a definition which we call blame freeness. This view of kindness requires a person to place himself in the position of his opponents and ask what he would do if he were in his strategic position. If he would have acted in a manner that would have increased his utility (i.e., if he would have been better off playing against himself rather than his opponent), then he has cause to blame his opponent for his actions and may punish him if the costs of doing so are not too high. We have then tested this notion in an experiment involving uneven tournaments where people can be blamed and punished for their behavior. We tested whether peoples' punishment behavior is blame-free consistent and whether they appear to punish when the theory asks then to a refrain otherwise.

By and large our data provides strong support for our blame-free concept with approximately 81% of subjects not punishing when they are not supposed to and 73% of them punishing when they should at least 50% of the time. These estimates are lower bounds to the adherence of our subjects to the theory since blame is only a necessary condition for punishment. Thus someone who does not punish when he blames his opponent, is not inconsistent with the theory. In addition, a set of probit regressions indicate that this theory may have stronger explanatory powers than theories that focus on inequality aversion and other intentions-based theories.

References

[1] Blount, S. (1995). "When Social Outcomes Aren't Fair: The Effect of Casual Attributions on Preferences." *Organizational Behavior and Human Decision Making*, 63(2): 131-144.

[2] Bolton, G. and A. Ockenfels, (2000). "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166-193.

[3] Carpenter, J. (2007). "The Demand for Punishment." *Journal of Economic Behavior & Organization*, 62: 522-542.

[4] Charness, G. and M. Rabin (2002). "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 817-869.

[5] Dufwenberg, M. and G. Kirchsteiger, (2004). "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47: 268-298.

[6] Fain, J. R. (2009). "Affirmative Action Can Increase Effort." *Journal of Labor Research*, 30: 168-75.

[7] Falk, A., E. Fehr, and U. Fischbacher, (2003). "On the Nature of Fair Division." *Economic Inquiry*, 41(1): 20-26.

[8] Falk, A., E. Fehr, and U. Fischbacher, (2008). "Testing Theories of Fairness—Intentions Matter." *Games and Economic Behavior*, 62: 287-303.

[9] Falk, A., E. Fehr, and C. Zehnder, (2006). "Fairness Perceptions and Reservation Wages - The Behavioral Effects of Minimum Wage Laws." *Quarterly Journal of Economics*, 121(4): 1347-1381.

[10] Falk, A. and U. Fischbacher, (2006). "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2): 293-315.

[11] Fehr, E. and S. Gaetcher, (2002). "Altruistic Punishment in Humans." *Nature*, 415: 137-140.

[12] Fehr, E. and K. M. Schmidt, (1999). "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics*, 114: 817-868.

[13] Fischbacher, U. (1999). "z-Tree - Zurich Toolbox for Ready-made Economic Experiments - Experimenter's Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University of Zurich.

[14] Gul, F. and W. Pesendorfer (2007). "The Canonical Space for Behavioral Types." mimeo.

[15] Kräkel, M. (2008). "Emotions and the Optimality of Uneven Tournaments." *Review of Managerial Science*, 2: 6–79.

[16] Levine, D. (1998). "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3): 593-622.

[17] Nikiforakis, N. and H. Normann, (forthcoming). "A Comparative Statics Analysis of Punishment in Public Good Experiments." *Experimental Economics.*

[18] Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83: 1281-1302.

[19] Schotter, A. (1990). "Free Market Economics: A Critical Appraisal." *Blackwell Publishers*, 2nd Edition, Cambridge, Mass.

[20] Schotter, A. and K. Weigelt, (1992). "Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results." *The Quarterly Journal of Economics*, 107: 511-539.

[21] Segal, U. and J. Sobel, (2007). "Tit for tat: Foundations of Preferences for Reciprocity in Strategic Settings." *Journal of Economic Theory*, 136(1): 197-216.

[22] Segal, U. and J. Sobel (2008). "A Characterization of Intrinsic Reciprocity." *International Journal of Game Theory* 36(3-4): 571-585.

[23] Sobel, J. (2005). "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 93: 392-436.