Equality of opportunity and the acceptability of outcome inequality

Robert Sugden[1] and Mengjie Wang[2]

21 October 2020

Abstract: In many real-world situations, unfairness of outcomes is not directly related to fairness-related properties of individual decisions; it is an unintended consequence of procedures in which individuals interact. Attitudes to such unfairness may be revealed in emotions of anger and resentment rather than in preferences over alternative decision outcomes. We conjecture that inequality is viewed with relatively little disfavour when it results from procedures that allow individuals equal strategic opportunities. We define a concept of procedural fairness which formalises intuitions about equality of opportunity. We report a Vendetta Game experiment in which negative attitudes to inequality can be expressed in costly and counter-productive 'taking' of co-players' assets. A given degree of material inequality induces more taking if the procedure that has generated it is unfair rather than fair. Surprisingly, there is excess taking by players whom procedural unfairness has benefited as well as by those it has harmed.

Keywords: equality of opportunity; procedural fairness; inequality; Vendetta Game;

JEL classifications: C92 Laboratory, group behavior; D63 Equity, justice, inequality, and other normative criteria and measurement; C72 Noncooperative games; D91 Role and effects of psychological, emotional, social, and cognitive factors on decision making.

Acknowledgements: Financial support for the experiment was provided by the Centre for Behavioural and Experimental Social Science (CBESS) and the University of East Anglia. The authors' work received funding from the Economic and Social Research Council of the UK (award no. ES/P008976/1). We thank Bertil Tungodden, Theodore Turocy, Daniel Zizzo, an Editor and three anonymous referees for advice and comments.

Declarations of interest: None.

[1] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia; r.sugden@uea.ac.uk; corresponding author.

[2] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia; m.wang3@uea.ac.uk.

1 Introduction

Economists' interest in *distributional* or *outcome* fairness – the equality or inequality with which economic outcomes are distributed between individuals – has a long history. Recently, there has also been growing interest in *procedural* fairness – the fairness or unfairness of the procedures by which those outcomes are generated. It is now widely recognised that a given degree of outcome inequality may be judged more or less socially or morally acceptable, depending on the procedure that led to it. However, the premise of our paper is that this literature focuses on an unduly narrow range of forms of procedural fairness.

There is ample evidence (a small part of which is cited later in our paper) that individuals judge outcome inequality to be unfair if it has been deliberately and arbitrarily chosen, either by someone who is favoured by it or by a third party. It is also well established that outcome inequalities are viewed less negatively if they are construed as rewarding differences in effort or merit, or if they result from purely random and unbiased mechanisms – as in Machina's (1989) famous example in which Mom tosses a coin to decide which of her two children will receive an indivisible treat. Two types of experimental design have been particularly widely used to elicit attitudes to inequality. One of these uses variants of the Ultimatum Game. A proposal for a distribution of payoffs is generated by some combination of decision-making by one player (the Proposer) and the realisation of some random process, and then the other player (the Responder) chooses whether to accept the proposal. In a second type of experiment, two or more players (the Producers) engage in independent production activities that contribute to a total 'social surplus'. One player (the Dictator) chooses how to distribute this surplus between the Producers.

These approaches seem inadequate for addressing one of the most fundamental normative questions in economics – are *markets* fair? In a competitive market, each participant's payoff depends on other participants' actions, but each participant, acting individually, has only limited power to affect the overall distribution of payoffs. When a competitive market generates outcome inequality, that inequality has not been chosen by any Proposer; but nor is it the result of a random mechanism. The total product of an ongoing market economy is not a fixed surplus that a Dictator can redistribute at will: the volume of that product depends on the procedures by which the market operates. A market is an example of what we will call a *Hayekian procedure* – a rule-governed procedure in which interactions between individuals can produce consequences that were not deliberately chosen

by any individual participant, but are not merely the result of chance. As some great economists and philosophers have recognised, people's attitudes to outcome inequalities induced by such procedures are of fundamental importance for the stability of market economies.

Defending the market system, Hayek (1976) acknowledges that markets can produce outcomes which, had they been consciously chosen by any identifiable person, would be judged to be unfair. He poses the rhetorical question:

> Are we not all constantly disquieted by watching how unjustly life treats different people and by seeing the deserving suffer and the undeserving prosper? And do we not all have a sense of fitness, and watch it with satisfaction, when we recognize a reward to be appropriate to effort or sacrifice? (p. 68)

Hayek accepts that the market can treat people unjustly in the same sense that life can: '[I]n the cosmos of the market we all constantly receive benefits which we have not deserved in any moral sense' (p. 94). Such perceptions of deservingness and undeservingness are psychologically natural, but they are not relevant for an evaluation of the market:

> In a spontaneous order the position of each individual is the resultant of the actions of many other individuals, and nobody has the responsibility or the power to assure that these separate actions of many will produce a particular result for a certain person... There can, in a spontaneous order, be no rules which will determine what anyone's position ought to be. (p. 33)

Buchanan (1964, p. 219) voices a similar thought:

> The 'market' or market organization is not a *means* toward the accomplishment of anything. It is, instead, the institutional embodiment of the voluntary exchange processes that are entered into by individuals in their several capacities. That is all there is to it.

For Hayek and Buchanan, a general willingness to follow the rules of the market is a valuable form of social capital – capital that is liable to be eroded if people evaluate market outcomes as if they were the results of someone's deliberate choice. Rawls (1971, pp. 16, 177, 453–462) makes a similar point about rules in general when he argues that a social institution is more likely to sustain itself over time if it is *psychologically stable* – that is, if the operation of its rules tends to reproduce a general belief that those rules are fair.

Hayek's and Buchanan's lines of thought raise the question of whether it is possible, even in principle, to assess the fairness of a Hayekian procedure. It might seem that even the criterion of equality of opportunity is inapplicable to such procedures. On one interpretation, advocated for example by Roemer (1998, p. 15), equality of opportunity requires that

'individuals who try equally hard should end up with equal outcomes' – that is, that all individuals are able to transform effort into reward on the same terms. Related principles of 'liberal egalitarianism' (Cappelen et al., 2007) and 'choice egalitarianism' (Cappelen et al., 2013) have been discussed in the experimental literature. If Hayek is right, this kind of ex post patterning of rewards cannot be guaranteed in a spontaneous order. However, we will argue that equality of opportunity can be defined in a different way, which *can* apply to Hayekian procedures. The essential idea is that there is equality of opportunity if social interaction is governed by 'rules of the game' that treat all individuals symmetrically. We will develop formal criteria of *equal opportunity* and *procedural bias* (i.e., one person having 'more' or 'better' opportunities than another) that are applicable to procedures in which individuals interact strategically. From now on, we will use the term 'equality of opportunity' in this procedural sense.

No real-world economy can provide complete equality of opportunity, but one might reasonably treat equality of opportunity as an ideal to which a social market economy – an economy in which markets are coupled with redistributive taxes and transfers and with publicly-funded social services – should approximate. But even in such an economy, individuals who enjoy (approximately) equal opportunities can experience significant inequalities of outcome that cannot be construed as rewards for differences of effort or sacrifice, or as the results of purely random mechanisms. Adapting an example from Sugden (2004), consider two school-leavers, Joe and Jane, with similar tastes, skills and qualifications who choose to train for different careers. At the time, the two careers seem to require roughly similar amounts of training and to offer roughly similar rewards, but each school-leaver feels more optimistic about the prospects of his or her chosen career than about the prospects of the other one. Over time, as a result of developments in technology and changes in consumers' tastes, Jane's optimism turns out to be justified, and Joe's does not: Jane's efforts are better rewarded than Joe's. Such inequalities may be unfortunate, but they are not unfair in the sense that can be said of outcome inequalities that result from procedural biases, such as systematic discrimination by ethnicity, gender or social class.

If economists are to understand the attitudes that people take to inequalities generated through the workings of Hayekian procedures, they need to take account of the fairness or unfairness of the rules by which the relevant procedures operate. Our paper investigates a conjecture made by Isoni et al. (2014) when discussing evidence from experimental bargaining games – the conjecture that inequality is viewed with relatively little disfavour

when it is the result of self-interested behaviour in an interaction in which all participants had the same strategic opportunities.

This statement of our research question hides a significant imprecision: What is an attitude to inequality? In economics, it is a standard practice to represent attitudes as preferences, and to assume that these are revealed in individuals' choices. But what is the appropriate choice problem for revealing attitudes to outcomes that no one has chosen, and that no one has unilateral power to change?

Our starting point is the idea that, for individuals who have participated in interactive procedures, perceptions of procedural unfairness can reveal themselves in emotions of inchoate dissatisfaction, resentment or anger rather than in purposeful decision-making. Such emotions may be expressed in actions and reactions that have negative material consequences for participants on both sides – advantaged and disadvantaged – of the perceived unfairness. Adapting the Vendetta Game of Bolle et al. (2014), we develop an experimental design that allows such material effects to be observed and measured, and that can be appended to any interactive procedure that results in a distribution of outcomes between participants. We treat the size of these effects as an inverse measure of the acceptability of the relevant procedure. Crucially, and we think usefully, our construct of 'acceptability' does not depend on any specific assumptions about the motivations that lie behind the apparently dysfunctional behaviour observed in Vendetta Games. Of course, the nature of those motivations is an important research topic in its own right; but it is not the topic of the present paper.

Using the Vendetta Game design, we investigate whether a given degree of inequality between individuals' material outcomes is more acceptable if it is generated by a procedure that gives equal strategic opportunities than if it is generated by one that is biased. To set a benchmark for our findings, we also compare the acceptability of an inequality generated by a strategically fair procedure with the acceptability of the same inequality when generated by a procedure that rewards differential effort.

The paper proceeds as follows. In Section 2, we briefly review the existing theoretical and experimental literature on fairness. We argue that there is a fundamental distinction between procedural and non-procedural principles of fairness, and that theories that treat randomness as the epitome of procedural fairness fail to capture significant aspects of equality of opportunity. In Section 3, we present our formal concepts of equality of

opportunity and of procedural bias. In Section 4, we explain why the Vendetta Game is particularly suitable for eliciting attitudes to these concepts. We then describe our experimental design (in Section 5) and the hypotheses it is intended to test (in Section 6). We report our results in Section 7 and discuss their implications in the final Section 8.

2 Procedural and non-procedural principles of fairness

In the theoretical and experimental literature about social preferences, principles of fairness are most commonly understood as properties of an individual's preferences over alternative allocations of material payoff between two or more individuals (possibly but not necessarily including the individual who holds the preferences). Many of the principles considered in this literature make no reference to the process by which the available surplus came into existence. Such principles include inequality aversion (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), the maximisation of social welfare or efficiency (e.g., Andreoni and Miller, 2002; Charness and Rabin, 2002), and kindness in the sense of being willing to forgo one's own payoff to achieve an 'equitable' allocation (e.g., Rabin, 1993). The social preference literature also studies second-order principles of reciprocity. In most applications, such principles are represented as preferences, held by individuals, for rewarding other people for acting on first-order principles of fairness or kindness and for punishing other people for acting contrary to such principles (e.g., Rabin, 1993; Charness and Rabin, 2002; McCabe et al., 2003; Falk and Fischbacher, 2006).

Another branch of literature investigates the degree to which individuals endorse various principles of distributive justice that have been proposed by economic and social philosophers. A typical experiment has two stages. In the first stage, two or more individuals engage in independent activities, each of which can produce variable quantities of some good. In the second stage, individuals' production is pooled into a 'social surplus'. Taking the role of the active player in a Dictator Game, each experimental respondent (who may or may not have been involved in the production stage) reports how she wants to distribute the surplus between the producers. Among the distributive principles that have been studied are ones that reward individuals according to effort exerted, skill shown, contribution made or luck experienced in the creation of the surplus (e.g. Ruffle, 1998; Konow, 2000; Capellen et al., 2007; Capellen et al., 2013; Mollestrom et al., 2015). An experimental set-up of this kind elicits the preferences of participants who act as social planners, choosing how to reward individuals for actions they have already taken, or for

characteristics they have already exhibited, when acting independently of one another. By treating the creation of the surplus as independent of the procedure by which it will be distributed, this type of design abstracts from the problem identified by Hayek – that in a spontaneous order, there can be no rules that determine what each individual's outcome ought to be, independent of the workings of the order itself.

There is also a substantial literature concerned with the kind of procedural fairness exhibited by Machina's Mom – *fairness as randomness*. For example, Blount (1995) reports an experiment that uses a variant of the Ultimatum Game in which the proposed payoff allocation is generated by an unbiased random device. She finds that second movers are more willing to accept disadvantageous inequality if the proposal is generated in this way than if it is deliberately chosen by a co-player. Bolton et al. (2005) investigate another variant of the Ultimatum Game in which the first mover chooses between three options – a proposal that favours him, a proposal that favours his co-player, and using an unbiased random device to select one of those two proposals. Second movers are more willing to accept disadvantageous proposals that come about through the random device than ones that have been chosen directly. There is an obvious sense in which the random devices in these experiments are procedurally fair, and in which Blount's and Bolton et al.'s results are evidence of the relative acceptability of inequalities that are generated in this way. The concept of fairness as randomness is generalised by Trautmann (2009), Krawczyk (2011) and Saito (2013) in theories of (respectively) 'process fairness', 'procedural fairness' and 'equality of opportunity' (which Saito [p. 3084] treats as synonymous with equality of ex ante expectations). In different ways, these theories develop the idea that, in strategic interactions, equality should be defined in terms of individuals' ex ante expectations of payoff rather than in terms of ex post payoffs. Notice that equality of expected payoffs is a property of the *outcomes* of a game, even though the viewpoint is ex ante; it is not a property of the *rules* of the game.

Building on the work of Rabin (1993) and Dufwenberg and Kirchsteiger (2004), Sebald (2010) proposes a theory of sequential reciprocity which expresses a 'Machina's Mom' intuition in a generalised form. If Mom chooses to give the treat to Mark rather than Mary, she has been unkind to Mary; Mary may then derive utility from an action that punishes Mom. But if Mom tosses a coin and *chance* favours Mark, the resulting inequality is not attributable to Mom and so neither child will want to punish or reward her. Aldashev et al. (2015) use Sebald's theory to explain the phenomenon of 'resentful demoralisation' in

randomised controlled trials (RCTs). In their model, the RCT is a trial of some policy intervention that is beneficial to the clients who receive it but whose effectiveness depends partly on the clients' own efforts. The policy is applied to a *treatment* group but not to a *control* group. If both groups know about this assignment, the experimenter has the role of Mom. If the assignment is not based on 'an explicit and credible randomisation mechanism' (p. 875), clients in the treatment (respectively: control) group may reward (punish) the experimenter by exerting more (less) effort than normal; in consequence, the RCT is liable to overestimate the effectiveness of the intervention.

However, the 'fairness as randomness' approach cannot easily represent attitudes to the fairness of procedures in which individuals interact strategically. As an illustration, consider a zero-sum strategic interaction in which two players compete to receive some material benefit. Suppose that each player chooses the strategy that maximises his probability of winning, given his belief about his co-player's strategy, and that these beliefs are in fact correct. Ex post, one player receives the benefit and the other does not. To whom is this inequality attributable? If we use the theory of reciprocity developed by Rabin, Dufwenberg, Kirschsteiger and Sebald, we must conclude that *both* players' intentions were maximally unkind (and thereby calling for punishment): each player faced a set of feasible ex ante probabilities of winning, and chose the one that was best for him and worst for his co-player. Notice that this conclusion holds *whatever the rules of the game*. The theory does not provide a way of saying that if the rules of the game treated both players equally, the players judge one another's self-interested intentions as morally permissible and the ex post inequality as fair.

It must be said that such judgements are compatible with the theory of reciprocity proposed by Falk and Fischbacher (2006), in which a player shows kindness (unkindness) by choosing to be at the unfavoured (favoured) side of outcome inequality. If the players of a procedurally fair zero-sum game act according to self-interest, their expected payoffs are equal, and so their intentions are neither kind nor unkind. But notice how such a theory would represent players' attitudes to a zero-sum game with procedurally *unfair* rules. A procedurally advantaged player who acted with the objective of winning such a game would be judged to be acting on unkind intentions and to be deserving of punishment. We suggest that social-preference theories of reciprocity are not recognising the distinction between kind (unkind) intentions and fair (unfair) procedures.

Intuitively, our understanding of the concept of procedural fairness within strategic interactions is of a set of rules that ensure that individuals interact on equal terms. There is a sense in which this understanding has an intermediate position in conceptual space between the idea of fairness as randomness and the idea of giving due reward to merit and effort. That this position creates a conceptual difficulty may help to explain why equality of opportunity has received relatively little attention in the social preference literature. Formally, any Hayekian procedure can be represented as a game between individuals, the outcome of which depends wholly or (if the game includes 'moves of nature') partly on the players' strategy choices. If those choices are not made at random, it must in principle be possible to define a concept of decision-making skill that the game tends to reward. That skill presumably depends on some combination of talent and cognitive effort that might conceivably be thought to deserve reward.

Schurter and Wilson (2009) raise this issue in relation to an experiment reported by Hoffman et al. (1994), in which subjects compete for the right to be the first mover in an Ultimatum Game. The competition is a general knowledge quiz. This experiment finds evidence that first movers in the Ultimatum Game offer less, and second movers are willing to accept less, when the role of first mover has been chosen by competition than when it has been selected at random. Schurter and Wilson argue that this design confounds the effects of (procedural) 'fairness' and (merit-based) 'justice', and that fairness is better implemented experimentally by using explicit random procedures.

Although the problem of confounds is real, it is important to recognise that fair competition can be valued *in itself*, irrespective of any belief about whether competition rewards merit. This is an attitude to the rules of procedures in which individuals interact; it cannot be elicited by investigating random processes. Recall the example of the two school-leavers (Section 1). Conceivably, one might judge the outcome inequality between Joe and Jane to be 'just' because it rewards Jane's merit as an economic forecaster. But, following Hayek, one might instead say that Jane does not *deserve* her favourable position, but the inequality is nonetheless fair because it is the outcome of a rule-governed market procedure in which Joe and Jane had the same opportunities.

Our paper is premised on the belief that it is important to develop concepts of procedural fairness and unfairness that apply to Hayekian procedures, and to investigate people's attitudes to inequalities that are generated by fair and unfair rules. There are some similarities between our theoretical approach and that of Chlaß et al. (2019), who propose a

concept of equality of 'decision rights' in games. However, their proposal measures the extent of an individual's decision rights by the cardinality of the set of non-dominated options from which she can choose, without reference to what those options are. As will emerge later, our concept of equality of opportunity is very different. Our methodological strategy is complementary with that of Hoffman and Spitzer (1985) and Hoffman et al. (1994). Instead of comparing procedurally fair interaction with randomness, as those papers do, we compare it with procedurally unfair interaction. The games that we use to make these comparisons were chosen with the aim of making the dimension of procedural fairness as salient as possible relative to the dimensions of talent and effort. Further, these games are zero-sum, so as to avoid prompting the thought that one person's talent or effort creates benefits for others, and is thereby worthy of reward.

3 Formalising the concept of equality of opportunity

The intuitive idea of equality of opportunity can be represented by a game (defined in terms of material consequences rather than utilities) in which all players have exactly the same strategic opportunities. To keep things simple, we focus on the case of a finite two-player game in strategic form. Formally, player 1 has a set of m pure strategies $S_1 = \{s_{11}, \ldots, s_{1m}\}$; player 2 has a corresponding set of n pure strategies $S_2 = \{s_{21}, \ldots, s_{2n}\}$. For each player $i \in \{1, 2\}$ there is a *payoff function* π_i which assigns a material payoff $\pi_i(k, l)$ to each profile (s_{1k}, s_{2l}) of pure strategies. We will normally interpret payoffs as quantities of some good – for example, money – that both players value positively. When we consider games that involve random processes or 'moves of nature', a payoff will be interpreted as the mathematical expectation of the quantity of that good.

Initially, we restrict attention to games in which $m = n$ and in which all strategies are distinct.[3] A sufficient condition for an $n \times n$ game to give the same opportunities to both players is that, for all $k, l \in \{1, \ldots, n\}$, $\pi_1(k, l) = \pi_2(l, k)$: player 1's payoff if he chooses his strategy k and player 2 chooses her strategy l is the same as player 2's payoff if she chooses her strategy k and player 1 chooses his strategy l. Notice that this condition uses the labelling convention that if strategies for the two players are indexed by the same number, they are 'the same'. To avoid this arbitrariness, we define an $n \times n$ game as giving *equality of opportunity* if

[3] Two strategies $s_{1k'}$, $s_{1k''}$ for (say) player 1 are *distinct* if there is some strategy s_{2l} for player 2 such that *either* $\pi_1(k', l) \neq \pi_1(k'', l)$ *or* $\pi_2(k', l) \neq \pi_2(k'', l)$. Throughout the paper, we assume that strategies are distinct.

the two players' positions are isomorphic in the following sense: it is possible to renumber strategies (separately for each player) in such a way that, after renumbering, $\pi_1(k, l) = \pi_2(l, k)$ holds for all k and l. In such a case, we will say that the new numbering of strategies is 'isomorphic'.

For example, consider the following Battle of the Sexes game, in which Arthur and Betty need to coordinate on where to go for an evening outing:

Game 1:		Betty	
		boxing	*opera*
Arthur	*boxing*	3, 2	0, 0
	opera	0, 0	2, 3

If we denote Arthur as player 1 and Betty as player 2, and if for each player we denote *boxing* by strategy 1 and *opera* by strategy 2, we find $\pi_1(1, 1) \neq \pi_2(1, 1)$ and $\pi_1(2, 2) \neq \pi_2(2, 2)$, contrary to the sufficient condition proposed above. But the same game can be re-described as one in which each player chooses between *better* and *worse* entertainments. If we keep the previous numbering of Arthur's strategies, but renumber Betty's so that *opera* becomes her strategy 1 and *boxing* becomes her strategy 2, $\pi_1(k, l) = \pi_2(k, l)$ holds for all k and l: the game gives equality of opportunity (and the new numbering is isomorphic).

The concept of *isomorphism* (or *symmetry*) between players and strategies, as used in our definition of equality of opportunity, is standard in game theory (e.g., Harsanyi and Selten, 1988: 73, 76).[4] However, we are using that concept in a different way. In game theory, isomorphism is usually invoked as a formal restriction on solution concepts. In that context, it expresses the principle that a theory of rational play should not discriminate between players or strategies that differ only in how they are labelled. In Game 1, for example, the probability with which a rational Arthur plays *boxing* must be equal to the probability with which a rational Betty plays *opera*. Notice the implication that, under the assumption of label-independent rational play, players who are isomorphic to one another have equal ex ante expectations of payoff. In contrast, we propose to interpret 'equality of opportunity' as a property of symmetry between the players' *pure* opportunities – that is, opportunities defined without any reference either to rationality or to predictions about

[4] Because we are using material payoffs, we do not follow Harsanyi and Selten in treating payoffs as utility numbers that are unique only up to positive affine transformations.

players' actual behaviour.[5] There is no implication that when ordinary human beings play an equal-opportunity game, their ex ante expectations are equal. Conversely, a demonstration that rational players of some game would have equal ex ante expectations does not imply that there is equality of opportunity.

Our definition of equality of opportunity has the following significant implication. Consider any equal-opportunity game with isomorphically numbered strategies, and any profile of strategies (s_{1k}, s_{2l}) in that game. Suppose this profile induces an outcome inequality that favours player 1, i.e., $\pi_1(k, l) > \pi_2(k, l)$. Suppose player 2 (she) objects that this inequality treats her unfairly. Player 1 (he) can point out that player 2 could have made the same choice (from the same opportunity set) as he did, and that he could have made the same choice as she did. Had that been the case, there would have been exactly the same outcome inequality, but in player 2's favour. In this sense, the actual inequality that results from (s_{1k}, s_{2l}) is attributable to the players' own decisions, and not to any unfairness in the rules of the game. We will say that this inequality is *counterbalanced* by the symmetrical inequality that would result from (s_{1l}, s_{2k}). In an equal-opportunity game, all possible inequalities are counterbalanced.[6]

We now propose a concept of procedural bias for $n \times n$ games (with distinct strategies). We do not claim that this concept encompasses every feature of such a game that could plausibly be considered a bias. Our objective is more modest: to characterise certain properties of games which, when viewed in the same perspective of pure opportunity as we have used to define equality of opportunity, can be considered to be unambiguous biases.

We define an $n \times n$ game to be *procedurally biased towards player1* if there is some renumbering of strategies such that $\pi_1(k, l) \geq \pi_2(l, k)$ holds for all k and l, and $\pi_1(k, l) > \pi_2(l,$

[5] Like standard game-theoretic analyses of isomorphism, our definition of 'pure opportunity' takes no account of how players or strategies are labelled. We use this property only as a simplifying assumption. For example, if a formally symmetrical Battle of the Sexes game is labelled in a way that makes one of the pure-strategy equilibria uniquely salient, the player who is favoured in that equilibrium might reasonably be deemed to have a strategic advantage (Schelling, 1960; Isoni et al., 2013).

[6] That all outcome inequalities are counterbalanced is a necessary but not a sufficient condition for equality of opportunity. For example, consider any 2×2 game in which the entries in the cells of the payoff matrix are (0, 0), (1, 1), (2, 2) and (3, 3). However these entries are positioned in the matrix, the game is not isomorphic between players but there can be no outcome inequality.

k) holds for some *k* and *l*.[7] In such a case, we will say that the new numbering is 'isomorphic'.

This definition is immediately intuitive. It also has an implication that is related to the concept of counterbalanced inequalities discussed above. Consider any $n \times n$ game (with isomorphically numbered strategies) that *either* gives equality of opportunity *or* is procedurally biased towards one of the players. Consider any profile of strategies (s_{1k}, s_{2l}). We will say that this profile imposes an *unbalanced inequality* on player 2 if two conditions hold. First, $\pi_1(k, l) > \pi_2(k, l)$, i.e., the profile induces an outcome inequality that favours player 1. Second, $\pi_1(l, k) \geq \pi_2(k, l)$ and $\pi_2(l, k) \leq \pi_1(k, l)$, and at least one of these inequalities is strict. In other words, if the two players' strategy choices were transposed, there would not be a corresponding transposition of payoffs (as there would be in an equal-opportunity game); relative to a symmetric transposition, player 1 would be favoured. The following result can be proved (proofs are given in Appendix 1):

> *Proposition 1 (Unbalanced Inequality)*: In any $n \times n$ game (with distinct, isomorphically numbered strategies) that is procedurally biased towards player1: (i) there is no strategy profile that imposes an unbalanced inequality on player 1, and (ii) there is at least one strategy profile that imposes an unbalanced inequality on player 2.

As an illustration, consider the following game:

Game 2:

		Betty	
		better	*worse*
Arthur	*better*	0, 0	4, 2
	worse	2, 3	0, 0

Clearly, this game is procedurally biased towards Arthur. If Arthur chooses *better* and Betty chooses *worse*, there is outcome inequality in favour of Arthur. Betty may object that, had both players made opposite choices, Arthur would have had the same payoff that she actually has, but she would have had a smaller payoff than Arthur actually has.

[7] In this and subsequent definitions and results, it should be taken as read that the player numbers '1' and '2' may be transposed.

Now consider any game in which there are m distinct pure strategies for player 1 and n for player 2, with $m > n$. Since the positions of the two players are clearly not isomorphic, we make it a matter of definition that such a game does *not* give equality of opportunity. We define the game to be *procedurally biased towards player 1* if two conditions are met. The first is that there is some renumbering of strategies such that $\pi_1(k, l) \geq \pi_2(l, k)$ holds for all $k \leq n$ and for all l. Equivalently, by deleting some subset of player 1's pure strategies, we can arrive at an $n \times n$ game that *either* gives equality of opportunity *or* is procedurally biased towards player 1. The second condition applies only if the $n \times n$ game gives equality of opportunity; it is that (given the renumbering of strategies) there is some $k > n$ and some l such that $\pi_1(k, l) > \pi_2(k, l)$, i.e., an outcome inequality favouring player 1 can occur as a result of player 1 choosing a strategy that is not available to player 2. Notice that, in defining procedural bias in $m \times n$ games, we do not consider whether the $m - n$ additional strategies available to player 1 are ones that might be chosen by a rational player. To do that would be inconsistent with our aim of analysing pure opportunity.[8]

As an illustration of our definitions, consider the following game:

Game 3:		Betty	
		better	*worse*
Arthur	*better*	0, 0	3, 2
	worse	2, 3	0, 0
	extra	1, 2	2, 1

This game is procedurally biased towards Arthur because it can be reduced to an equal-opportunity game by deleting his strategy *extra*, and because the strategy profile (*extra*, *worse*) leads to outcome inequality that favours Arthur. Suppose that this profile is played. Betty can object that the resulting inequality is unfair because the game did not allow her a strategy symmetrical with Arthur's *extra*. In contrast, if the strategy profile (*extra*, *better*) is played, Arthur's experience of unfavourable inequality is not unfair, because it was the result of his using an opportunity that was not available to Betty.

[8] Chlaß et al.'s (2019) concept of 'equality of decision rights' is not a criterion of pure opportunity in this sense, because it excludes dominated strategies when counting the number of strategies available to each player.

Our definitions of equal opportunity and bias are not exhaustive: many games that do not give equality of opportunity cannot be classified as biased in favour of either player. However:

> *Proposition 2 (Mutually Exclusive Classification)*: The classifications 'equality of opportunity', 'procedurally biased towards player 1' and 'procedurally biased towards player 2' are mutually exclusive.

In general, the fact that a game is procedurally biased towards one of the players does not imply that, given Nash equilibrium play, that player's payoff will be at least as great as her co-player's: having more or 'better' opportunities in a game can be a strategic disadvantage.[9] However, in the special case of zero-sum games, there is a significant relationship between our concepts of procedural fairness and bias and theories of rational play:

> *Proposition 3 (Strategy-matching):* For every two-player game G with players i, $j \in \{1, 2\}$: if G gives equality of opportunity [respectively: is procedurally biased towards player i], then for every probability mix σ_j of strategies that is available to player j, there is some response σ_i by i such that, if the mixed strategy pair (σ_i, σ_j) is played, i's expected payoff is equal to [no less than] j's. *Corollary*: if in addition G is zero-sum, then i's expected payoff is equal to [no less than] j's in every Nash equilibrium.

If 'strategic advantage' is interpreted in terms of Nash equilibrium, Proposition 3 shows that if a zero-sum game has equality of opportunity, neither player has a strategic advantage; if such a game is procedurally biased towards one player, that player may have a strategic advantage, and definitely does not have a strategic disadvantage.

In thinking about our definitions of equality of opportunity and procedural bias, it is important not to conflate a game-theoretic model of a real-world interaction with the interaction itself. It is often possible to recognise the presence or absence of equality of opportunity in a real-world situation by noticing obvious symmetries or asymmetries that are inherent in the structure of the relevant interaction, without explicitly describing the strategic

[9] This is true even if Nash equilibrium is unique. For example, consider the game formed by substituting 4 for 0 as the payoff to Arthur from (*opera, boxing*) in Game 1. The revised game is procedurally biased towards Arthur, but its unique Nash equilibrium is (*opera*, *opera*).

form of a game that models it. As we explain in Section 5, our experimental design uses interactions whose symmetry or asymmetry properties have this kind of obviousness.

4. Eliciting attitudes to procedural unfairness

As we explained in Section 1, our experimental objective is to investigate individuals' attitudes to inequalities that they experience as participants in interactive procedures governed by rules that they themselves have not chosen.

In social preference research, the most common method of eliciting attitudes to fairness is to ask individual respondents to choose between alternative payoff distributions – for example, by taking the active role in a Dictator Game, or the role of Respondent in an Ultimatum Game. But it is not obvious that attitudes to procedural fairness can be represented as stable preferences over payoff distributions. In many real-world situations in which issues of procedural fairness are salient, ordinary individuals do not have opportunities to make unilateral, uncontested changes to the distribution of payoffs. Commonly, if one individual has an opportunity to impose losses on another, that other person has some opportunity to retaliate. The implications of this fact are now widely recognised in relation to acts of punishment. Fehr and Gächter (2000) show that the existence of one-sided punishment opportunities increases contribution levels in public good games. However, Nikiforakis (2008) shows that when both punishment and counter-punishment are allowed in a Public Good game, cooperators' willingness to punish decreases, leading to the breakdown of cooperation. In the presence of counter-punishment opportunities, people also reveal strong desires to reciprocate punishment (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007). These findings cast doubt on the external validity of experimental designs that provide one-sided opportunities for punishment or redistribution. Further, the concept of punishment seems out of place when the outcome of an interactive procedure has not been chosen by any particular person. When a Hayekian procedure is unfair, the blame attaches to its rules, and not to the behaviour of particular participants. We believe that many of the real-world situations in which people are able to express opposition to procedural unfairness are better modelled by Vendetta Games, and that the attitudes that are expressed in these situations often correspond with 'hot' emotions of resentment and anger rather than stable distributional preferences.

In Bolle et al.'s (2014) Vendetta Game, there are two players. There is a single monetary prize which will be received either by one of the players, or by neither of them.

States of the game are described by the players' respective probabilities of winning the prize. Some pair of probabilities is set as the initial state of the game. The players then take turns to decide whether to stick with the current state of the game or to take probability from one another in fixed-size 'blocks'. Whenever a block is taken, a fixed proportion of its probability is lost; the remainder passes to the taker. The game ends if both players choose not to make further takes, or if no further takes are possible.

In Bolle et al.'s experiment, this game induced substantial frequencies of repeated taking moves by both players, even when those moves were contrary to received theories of rational play. In many cases, players continued taking until no further taking was possible. Vendettas (i.e., sequences of taking and counter-taking) were common even when the initial state was equal, but more frequent when it was unequal. Repetition of the game had only a weak tendency to reduce the amount of taking, with the implication that players' behaviour was not driven by false beliefs about what their co-players would do.

Participants' self-reported emotional states were elicited repeatedly during games, and also retrospectively (and in more detail) at the end of the experiment. There was a strong positive correlation between taking behaviour and self-reported within-game anger. In the post-experiment data, recollections of anger were strongly and positively correlated with recollections of envy, irritation and jealousy (emotions that were not elicited during games) (pp. 115–117).

In the light of this evidence, we believe that the Vendetta Game is best understood as an experimental model of individually and collectively dysfunctional behaviour, associated with mutually-reinforcing emotions of resentment and anger – as in real-world vendettas. One might reasonably claim that, once a vendetta has been initiated, resentment and anger are emotions of negative reciprocity, understood as psychologically primitive desires to return hurt for hurt. However, players' unwillingness to terminate vendettas, despite repeatedly incurring equal losses, suggests myopia rather than the kind of rationality described by theories of social preference. (A social-preference explanation would require the assumption that each player was willing to sacrifice at least one unit of payoff for each unit of punishment inflicted on her co-player.) The *initiation* of vendettas would be particularly hard to explain by such theories. Prior to the first taking move in Bolle et al.'s experiment, neither player had been kind or unkind to the other. Since there was ample scope for a player to retaliate *after* a first taking move by her co-player, initiating a vendetta in anticipation of possible unkindness would incur an unnecessary risk.

Precisely because the Vendetta Game is so effective in inducing negative emotions and dysfunctional behaviour, it is a useful experimental tool for measuring the psychological stability of different procedures – that is, the degree to which the outcomes generated by those procedures are perceived as acceptable. As far as we know, our experiment is the first to use Vendetta Games in this way. We must emphasise that our experiment was not intended to investigate alternative explanations of taking behaviour in Vendetta Games: that would have required a very different experiment. Nor does our design depend on any particular hypotheses about the explanation of this behaviour. What matters is that the Vendetta Game is a set-up in which, as in many real-world environments, individuals can engage in socially wasteful taking from others who have opportunities to retaliate. In our design, experimental control comes from the fact that the same Vendetta Game is used in conjunction with different procedures, all of which generate the same payoff inequality.

5 Experimental design

Our experimental design elicited individuals' attitudes to inequalities that had been generated by different procedures. Part 1 of the experiment generated the inequality; Part 2 elicited attitudes towards it.

Our main comparison is between two treatments. In Part 1 of each of these treatments, participants were paired to compete in a series of identical games. In the Fair Rule treatment, the games were zero-sum and procedurally fair according to our equal-opportunity condition; in the Unfair Rule treatment, they were zero-sum but procedurally biased towards one of the players. To allow some calibration of our results in relation to the existing experimental literature, we also used a Real Effort treatment in which inequality was generated as the result of a procedure in which paired participants competed in a real effort task, the rules of which treated the two participants symmetrically. In all treatments, the relevant competition ended with a winner and a loser; the winner received nine tickets for a specified lottery and the loser received three**.** Thus, all three treatments induced the same inequality of outcome between the two participants.

In principle, the Real Effort competition might be modelled as a non-zero-sum game in which each player chooses an effort level and the player who chooses the higher level is the winner. If the relationship between effort and cost is assumed to be the same for both players, this is an equal-opportunity game. Real-effort competitions are common components of experimental designs, but are usually interpreted by experimenters as – and

are expected to be understood by participants as – procedures that give rewards according to a principle of desert or merit. The Fair Rule competition is much more obviously a strategic interaction. There is therefore some interest in comparing the acceptability of inequalities resulting from these two types of competition. In our design, however, controlled comparisons between fair and biased procedures are possible only between the Fair and Unfair Rule treatments.

In Part 2 of each treatment, the paired participants played a Vendetta Game. Participants were given opportunities to take lottery tickets from their co-players, in alternating turns, to increase their holdings of tickets by a fraction of what they took. By using or not using these opportunities to take tickets, participants were able to reveal their attitudes to the inequality generated in Part 1. At the end of the experiment, for each pair of participants, one of the twelve lottery tickets was drawn at random. If that ticket had not been wasted in the Vendetta Game, its holder won a money prize.

We now describe the components of the experiment in more detail. The full instructions to participants can be found in the Supplementary Material.

5.1 Assignment of roles

Each experimental session was assigned to one of the three treatments. At the beginning of each session, participants were admitted to the lab one by one and were asked to sit at any vacant computer terminal. At this stage, they were given no indication that the choice of where to sit might be significant. When the experiment began, participants were told that those sitting at odd-numbered seats would be 'participant As' and those at even-numbered seats would be 'participant Bs'; each participant A would be randomly and anonymously matched with a participant B for the duration of the experiment. By allocating seats before the start of the experiment and in this unstructured way, we intended that the assignment of A and B roles would be separated in participants' minds from the rules of the competition they faced in the experiment itself.

Considered in relation to the theory developed by Aldashev et al. (2015) (discussed in Section 2 above), this assignment of roles was made by implicit rather than explicit randomisation, but (since subjects had chosen their own seats) it was clear that there was no intent by the experimenters to favour particular participants over others. In the Unfair Rule treatment, the assignment affected participants' expected payoffs, and so some versions of reciprocity theory might predict that players assigned to the disadvantaged role would want to

harm *the experimenter*.[10] But the Vendetta Game is a set-up in which players harm *one another*. If resentment against the perceived source of unfair rules were to be expressed in costly 'punishment' of co-players, that would be evidence of the kind of non-instrumental attitudes that our experiment was designed to investigate.

5.2 Part 1: the fair and unfair Search Competition Games

In the Fair Rule and Unfair Rule treatments, each pair of participants played a series of *Search Competition Games*, simulated on their computer screens. At the start of each game, each player was 'dealt' a card. Each card had a number of 'points', which could be any of the whole numbers from 1 to 100. Each of these numbers was equally likely at each deal (as if a new deck of 100 cards was used every time a card was dealt). Each player was offered a pre-specified number of opportunities to choose whether to stick with the card currently held or to return it and replace it with a newly-dealt card. The number of opportunities available to each player was common knowledge. In the Fair Rule treatment, participants A and B were both allowed to make up to three replacements in each game. In the Unfair Rule treatment, A was allowed to make no more than one replacement while B was allowed to make up to three. We will refer to A and B in this treatment as the *disadvantaged* and *advantaged* players respectively. (These terms were not used in the experiment itself.) During this stage of the game, neither player could see what cards her co-player was being dealt, or whether the co-player was using his replacement opportunities. After both players had chosen to stick with a card they had been dealt or had used up all their replacement opportunities, each saw the other's final card and was shown how many replacement opportunities that player had used. The player holding the higher-numbered card was the winner of the game. If both cards had the same number, the game was a draw. The first player to win four games was the overall winner of the series of games: draws were not counted. (In fact, only 1.2 per cent of games were drawn.) A screen shot of the Search Competition Game is shown in the Supplementary Material.

Given our aim of investigating attitudes to procedural fairness and unfairness, these games have a number of desirable features. Their rules are very easy to understand. The procedural fairness of the Fair Rule game, and the procedural bias of the Unfair Rule game,

[10] The only obvious way in which players could harm the experimenter was by increasing the cost of the experiment to the experimenter's budget. Since each act of taking reduces this cost, the effect of this kind of behaviour would be to make taking *less* frequent when the initial inequality was unfair – contrary to the hypotheses tested in experiment.

are transparent and salient. (It is clear from a simple statement of the rules of the Fair Rule game that the two players are treated exactly equally. In the Unfair Rule game, anything that A can do with his one replacement opportunity can also be done by B with her three; but B also has opportunities that are not available to A and that can help her to win.[11]) The bias in the unfair rules is experienced directly by A whenever he is dealt a low-valued second card, and by B whenever she replaces a second or third card. It is also obvious that each player's chance of winning depends partly on how she uses her replacement opportunities, and that in broad-brush terms, what a player needs to do is to replace relatively low-valued cards and to stick at relatively high-valued ones. Thus, the game induces purposeful engagement by the players. When one player learns that her co-player is using replacement opportunities, she can reasonably infer that he is trying to increase the probability that that he wins and she loses.[12] Nevertheless, the outcome of any single game, even under the unfair rules, is predominantly a matter of luck: in the Nash equilibrium of the unfair game, the probability of a win for the advantaged player is approximately 0.63, implying that this player's probability of winning the series is 0.77.[13] Since, for each player, the series of games has only two possible outcomes ('win' or 'lose'), attitudes to risk are irrelevant, at least for rational players.

Although it would be very difficult for a participant to assess the exact balance of skill and luck in the game, the fact that luck plays an important part (and that luck is more important in the Fair Rule treatment than in the Real Effort treatment) is easy to recognise. For the reasons explained in Section 2, our games must provide *some* opportunity for skill or effort if they are to be useful in an investigation of procedural fairness. We cannot rule out the possibility that some participants believed that Search Competition Game winners deserved their higher payoffs as a reward for their merit and/or cognitive effort in strategic

---

[11] That the Unfair Rule game is procedurally biased towards player B can be shown formally by considering any given strategy s_A for A. There is an exactly equivalent strategy for B: use s_A at the first replacement opportunity, and then use no more opportunities. But B also has many strategies that are not available to A and which, if played in response to some of A's strategies, give B a greater than 0.5 probability of winning. To show that this is the case it is not necessary to enumerate the players' enormous strategy sets.

[12] Only three of the 104 participants in the Fair Rule treatment and only four of the 150 participants in the Unfair Rule treatment failed to use any replacement opportunities.

[13] These numbers are approximations because they are calculated for a game in which card numbers are uniformly distributed over the real interval [0, 100]. One of our reasons for using a series of games rather than a single one was that we wanted the game itself to be simple, while giving the advantaged player a high probability of winning overall.

reasoning. However, our prior expectation was that negative attitudes to inequality would be revealed primarily by the losers in each treatment. Given what is known about the psychology of *locus of control* (Rotter, 1966), it seems unlikely that *losers* would believe that winning was evidence of effort or merit.

A more serious potential concern is that, if cognitive ability *in fact* had a significant effect on a participant's probability of winning Search Competition Games, there would be some selection bias in the assignment of participants to roles in the Vendetta Game. (For example, cognitive ability might be positively correlated with rational play in Vendetta Games.) Our design was premised on the assumption that, for laboratory subjects with no previous experience of this type of game, there would be no major differences of cognitive ability between winners and losers. In Section 7.1, we report evidence that supports that assumption.

5.3 Part 1: the real effort task

Real effort tasks, in which experimental subjects are rewarded for their performance in burdensome low-skill tasks, are commonly used to investigate desert-based concepts of fairness (e.g. Burrows and Loomes, 1994; Fahr and Irlenbusch, 2000; Konow, 2000). Part 1 of our Real Effort treatment was designed to create inequalities that could be construed as rewarding differences in effort.

We used the *encryption task*, as developed by Erkal et al. (2011). For each pair of matched participants, each member of that pair was given an encryption table which assigned a number to each letter of the alphabet in a random order. She was then presented with words in a predetermined sequence and was asked to encrypt them by substituting the letters with numbers using the encryption table. All participants were given the same words to encode in the same order. After a participant encoded a word, the computer would tell her whether the word had been encoded correctly. If it had been encoded wrongly, she would be asked to check her codes and correct them. Each time she encoded a word correctly, she was given another word to encode. This process continued for six minutes. The participant who had encoded more words in this period was declared the winner. If there was a tie, the winner was the participant who had encoded the words in a shorter time. A screen shot of the real effort task is shown in the Supplementary Material.

5.4 Part 2: the Vendetta Game

Part 2 took the same form in all three treatments. Participants remained in the same pairs as in Part 1, but now were referred to as 'the winner in Part 1' and 'the loser in Part 1'. The game involved twelve lottery tickets, numbered from 1 to 12. Initially, nine of these tickets, selected at random, were assigned to the Part 1 winner; the remaining three were assigned to the loser. The players moved in turn, the loser moving first. When it was a player's turn to move, she was asked to choose whether she wanted to take lottery tickets from her co-player, and if so, how many tickets to take. Tickets had to be taken in blocks of three (so the number of tickets taken could be three, six or nine, up to as many as the co-participant held at the time). For each block of three tickets that the player took, one of those tickets was transferred to her and the other two were wasted (i.e., lost to both players). Players always had the option of not taking any tickets. The game ended if one of two cases applied. The first case occurred if one or both players were still able to take tickets, but the player(s) who was (were) able to do this had chosen not to do so on two consecutive moves.[14] The second case occurred if both players held less than three tickets, and therefore no positive multiple of three tickets could be taken from either of them. Because we wanted to be able to pick up the effects of hot emotions, we used a design in which Vendetta Games were played in real time, rather than using the strategy method to elicit players' conditional decisions at all nodes in the game tree.

Figure 1 shows all the distributions of tickets that can be reached in the game. The starting point is (3, 9), i.e., the loser holds three tickets and the winner holds nine. For each possible distribution of tickets, the solid arrows show the possible moves by the loser from that point in the game; any continuous sequence of these arrows is a possible move. For example, if the current distribution is (3, 9) and the loser has an opportunity to take, she can move to (4, 6), (5, 3) or (6, 0). Similarly, the broken arrows show possible moves by the winner. The only distribution at which no further taking moves are possible is (0, 2).

[Figure 1 near here]

This game has the same general structure as the Vendetta Games studied by Bolle et al. (2014), but it was framed in more concrete terms and displayed in a more intuitive visual form. In Bolle et al.'s experiment, states of the game were described as numerical

[14] The 'two consecutive moves' rule was a feature of the game used by Bolle et al. (2014). By allowing players to reconsider 'stick' decisions but not 'take' decisions, this rule may contribute to the power of the Vendetta Game to induce dysfunctional behaviour. If so, that is not a problem for our cross-treatment comparisons.

probabilities of winning, which could be changed according to algebraic rules. In our design, these states were represented as distributions of physical objects between three 'baskets' – one for each player, and a 'bin' for wasted tickets. Players' decisions transferred specific objects between these baskets. Figure 2 shows a screen shot of the beginning of the game.

[Figure 2 near here]

If players are rational and self-interested and if this is common knowledge, the distribution (4, 6) is the unique backward-induction solution. Other outcomes are possible for rational players who act on social preferences. For example, if both players are mildly altruistic (specifically, if each player treats one unit of the other's payoff as equivalent to more than one third of a unit of her own), the game will stay at (3, 9). But it is important to recognise that our design does not depend on any particular assumptions about players' reasoning. Indeed, as explained in Section 4, the Vendetta Game is a useful experimental tool precisely because of its power to induce irrational behaviour.

5.5 Final earnings

At the end of each session of the experiment, twelve tickets numbered from 1 to 12 were put into a bag and a participant was asked to draw one at random. This draw determined the winning ticket number for every Vendetta Game in that session. In each pair, if the winning ticket was in the basket of one of the participants, that participant received a prize of £24; the other participant earned nothing from the game. If the winning ticket was in the bin, both participants earned nothing from the game. In addition, every participant was paid a participation fee of £3.

5.6 Implementation

The experiment was conducted between November 2015 and January 2016 at the CBESS Experimental Laboratory at the University of East Anglia. Participants were recruited from the general student population via the CBESS online recruitment system (Bock *et al.*, 2012). The experiment was programmed and conducted with the experimental software z-Tree (Fischbacher, 2007). We ran 18 sessions in total: six for the Fair Rule treatment (with 104 participants), eight for the Unfair Rule treatment (150 participants), and four for the Real Effort treatment (72 participants). We needed more participants in the Unfair Rule treatment than the Fair Rule treatment because our main aim was to compare the Vendetta Game behaviour of Fair Rule participants with that of Unfair Rule participants from series of card

games that had been won by the advantaged player; as explained in Section 5.2, we expected around 20 to 25 per cent of these series to be won by the disadvantaged player. We recruited fewer participants for the Real Effort treatment, as this was intended only to provide a point of comparison with the existing literature. Participants' ages ranged from 18 to 63; approximately 60 per cent were female.[15] The experiment lasted about 50 minutes. Every participant earned either £3 (the participation fee) or £27 (that fee plus the £24 lottery prize). Average earnings were £10.67 per participant.

6 Hypotheses

Informally stated, our fundamental hypothesis is that there is greater willingness to accept inequality if it results from an equal-opportunity game than if it results from a game that was procedurally biased in favour of the player who ended up as the winner. Thus, we are primarily concerned with behaviour in the Vendetta Game played in Part 2 of the experiment, interpreted as revealing participants' attitudes to the inequality created in Part 1. That inequality could be created in four different ways – through a Fair Rule game (FR), through an Unfair Rule game that was won by the advantaged player (URA), through an Unfair Rule game that was won by the disadvantaged player (URD), or through a Real Effort task (RE). It is convenient to refer to these as different 'treatments' applied to a single game.[16] Our fundamental hypothesis requires comparisons between the FR and URA treatments.

This hypothesis can be firmed up in various ways, depending on whether 'acceptance' is interpreted as an individual or collective phenomenon and on whether it is measured by players' propensities to stick at the initial token allocation or (inversely) by their propensities to take from one another in the Vendetta Game as a whole. We test the following two hypotheses about collective behaviour:

> *Hypothesis 1 (collective sticking at the initial allocation):* The Vendetta Game is more likely to end at the (3, 9) allocation in the FR treatment than in the URA treatment.

[15] The gender question allowed 'prefer not to say' as an answer. Of the 326 participants, 123 identified as male and 196 as female; 7 preferred not to say.

[16] This is a slight misnomer, because the allocation of participants between URA and URD 'treatments' was partly the result of decisions made in Part 1. But, as explained in Section 4.2, our working assumption was that this allocation was effectively random.

Hypothesis 2 (collective taking): In total, more tokens are wasted in the URA treatment than in the FR treatment.

Intuitively, one might expect non-acceptance of inequality to take different forms for winners and losers. A loser can perceive at least some acts of taking both as protests against the initial inequality and as partial rectifications of it. Thus, if inequality is perceived as less acceptable in the URA treatment than in the FR treatment, we should expect losers' willingness to take to be greater in URA games. Our prior expectation was that there would be an opposite effect for winners, that is, a tendency for them to view modest amounts of taking by losers as more legitimate, and so less deserving of retaliation, in URA games. Even if there is some tendency for winners to believe that, by winning, they have shown skills that deserve reward, one would expect that belief to be less strong when the rules had been biased in their favour. To the extent that winners engage in the apparently myopic act of taking at (3, 9) – adding to the initial inequality by moving the game to a position where their opponents have nothing to lose by counter-taking – one might expect that behaviour to be less likely, the less acceptable the initial inequality. These expectations are encapsulated in the following hypotheses about individual behaviour:

Hypothesis 3 (individual sticking at the initial allocation): (a) For losers, sticking at the (3, 9) allocation is less likely in the URA treatment than in the FR treatment; and (b) for winners, sticking at that allocation is more likely in the URA treatment than in the FR treatment (unless there is 100 per cent sticking in both).

Hypothesis 4 (individual taking): (a) For losers, the overall propensity to take is greater in the URA treatment than in the FR treatment; and (b) for winners, the overall propensity to take is lower in the URA treatment than in the FR treatment.

The concept of 'overall propensity to take' in Hypothesis 4 needs some explanation. Because of the interactive nature of Vendetta Games, and because we can observe players' behaviour only at the nodes in the game they actually reach, disentangling winners' and losers' underlying attitudes to taking is not straightforward. The fundamental problem is that which nodes are reached by one player depends in part on the behaviour of the other. If we are to identify determinants of behaviour separately for winners and losers, we need to control for this effect. In principle, it is possible to test for overall cross-treatment differences

in taking propensities by carrying out a separate test at each decision node.[17] Consider any given decision node *N* for the winner (for example, the winner's second opportunity to take at (4, 6), reached after the loser took three tokens at her first opportunity at (3, 9)). Suppose we observe behaviour at this node in both FR and URA games, and we want to test whether this observation is consistent with the null hypothesis that the distribution of winners' behaviour strategies (i.e. strategies that specify a player's behaviour at *every* node in the game) is the same for FR and URA players. A winner can reach *N* only if a particular combination of previous moves is played. But, given the null hypothesis and provided that (within FR and URA games considered separately) there is no correlation between the strategies of matched players, the distribution of winners' strategies conditional on reaching *N* is the same for FR and URA games. Thus, if behaviour at *N* is significantly different in the two games, that is evidence against the null hypothesis. Intuitively, the implication is that Hypothesis 4 should be read as referring to overall patterns in the results of the node-specific tests we have described. In Section 7.3 we will explain how we used this idea to design practicable tests of that hypothesis.

Our experiment allows tests, analogous with those of Hypotheses 1 to 4, of differences between the FR and RE treatments. Because comparisons between the two types of fair competition is orthogonal to our main research question, we do not propose any formal hypotheses about differences between these treatments. In principle, our experiment also allows similar tests of differences between behaviour in URA and URD games. Intuitively, one might expect the initial inequality in the distribution of tokens to be more acceptable in URD games, since in these games the winner has succeeded despite a handicap imposed by unfair rules. But, as explained in Section 6.6, the URD games exist only as a by-product of the procedure by which the URA games were created; the number of these games (17) is too small for useful analysis.

7 Results

For completeness, we report data for all four treatments. Our primary concern is with Hypotheses 1 to 4, which involve comparisons between the FR and URA treatments, but we also report tests of differences between the FR and RE treatments.

[17] Following the standard practice in game theory, we define nodes in relation to game trees. Since the Vendetta Game is a game of perfect information, each decision node for a specified player has a unique history of previous moves by all players.

7.1 Behaviour in the Search Competition Games

As noted in Section 5.2, our design assumes that, in each of the two types of search competition (i.e., fair and unfair), the players who end up as Part 1 winners are a random sample of the whole set of players. There would be selection bias if, and to the extent that, the probability of winning was correlated with players' cognitive ability.[18] Clearly, the probability that a given participant wins a competition depends on the strategy she follows in the games, and on luck (i.e., the cards she is dealt). To the extent that it depends on luck, the selection of winners is random. Thus, selection bias can occur *only if* there is a systematic difference between the strategies used by winners and losers. Excluding the effects of sampling error, it is tautological that if there is any such difference, winners' strategies are 'better' than losers' according to the *empirical best response* criterion (i.e., better as responses to the actual distribution of strategies in the population). Thus, one would need to be cautious in interpreting such a difference as evidence of a transferable form of cognitive ability. But the *absence* of a difference would definitely imply the absence of selection bias.

We test for such a difference by examining players' replacement decisions following the first card deal in the first four games of the Fair Rule competition. We use the Fair Rule competition because the symmetry between the two players gives us a particularly large sample size, and because one would expect the correlation between ability and success to be strongest when the rules are fair. We use the first four games because (under the 'first to win four games' rule) everyone played at least four games. We use the first deal because everyone faced a first deal in every game.[19]

Figure 3 shows Lowess-smoothed graphs of the probability that the first card was accepted, as a function of the value of that card, separately for Part 1 winners and Part 1 losers. If we assume that, in any given game, each player's strategy for accepting or rejecting the first card can be described by some *cut-off* value (i.e., the minimum value at which the first card is accepted), these graphs can be interpreted as distributions of cut-off values. As

[18] The analysis reported in this subsection was carried out in response to concerns expressed by readers of an earlier version of this paper. Since the experimental design did not elicit independent measures of cognitive ability, our analysis is based on observations of players' behaviour in the Search Competition games.

[19] We have run the same tests (separately) for advantaged and disadvantaged players in the Unfair Rule competition and found similar results (i.e., no significant differences between the strategies of winners and losers). We have also tested whether the average number of replacement opportunities used in the first four Fair Rule games was different for winners and losers. The difference was not significant; the averages were almost the same (1.024 per game for winners, 1.067 for losers).

benchmarks: the Nash equilibrium cut-off is 83; the empirical best response is a cut-off of 76.[20] Three features of these distributions are immediately obvious. First, there is a high degree of heterogeneity in players' cut-offs. Second, most players' cut-offs are well below the rational-choice benchmarks (the median cut-offs are 61 for winners and 59 for losers). Third, and most important, there is almost no difference between the distributions of cut-offs for winners and losers.[21] The regression analysis in Table 1 shows that the difference is not statistically significant. We find no evidence that the experiment was affected by selection bias.

[Figure 3 near here]

[Table 1 near here]

It is also interesting to ask whether there was any tendency for advantaged players in the Unfair Rule competition to offset their advantage by deliberately not using some or all of their 'unfair' additional replacement opportunities. Recall from Section 2 that, according to some theories of reciprocity, an advantaged player who did not exercise this kind of restraint would be acting on unkind intentions. In answering this question, we need to take account of a possible confound: the Nash equilibrium cut-offs of a player with a given number of replacement opportunities differ according to the number of opportunities available to her opponent. Advantaged players in the Unfair Rule competition are '3 vs 1' players (i.e., have three opportunities against an opponent with one); both players in the Fair Rule competition are '3 vs 3' players. Under the assumption of Nash equilibrium play, the expected number of cards replaced is higher for 3 vs 3 players (1.88) than for 3 vs 1 players (1.58). Thus, were we find fewer replacements by 3 vs 1 players, that would not necessarily indicate intentional restraint by players who were unfairly advantaged. In fact, however, the average number of cards replaced in the first four games was slightly *higher* for 3 vs 1 players (1.20) than for 3 vs 3 players (1.05), although not significantly so ($p = 0.146$ in a two-tailed Mann-Whitney test). We conjecture that players did not recognise the strategic significance of the number of opportunities available to their opponents; although they were aware of the procedural

[20] Explanations of the calculation of the Nash equilibrium and the estimation of the empirical best response are available from the authors.

[21] Given the heterogeneity of cut-offs, it may seem surprising that there is so little difference between the overall behaviour of winners and losers. A possible explanation is that this much of this heterogeneity resulted from variation in the behaviour of each player across card deals, rather than variation between 'types' of player with differing degrees of rationality.

unfairness of the Unfair Rule competition, they did not think of this as imposing moral constraints on the behaviour of advantaged players within the given rules.

7.2 Vendetta game outcomes

Table 2 summarizes the outcomes of the Vendetta Game in the four treatments. To get a broad-brush picture of how the game was played, we begin by looking at the aggregate data in the final column of the table.

[Table 2 near here]

Recall that, if players act on rational self-interest and if this is common knowledge, the outcome of the game's unique backward-induction solution is the (4, 6) allocation. In fact, this was the end point of only 12.9 per cent of games. In 32.5 per cent of games, the game ended at the initial (3, 9) allocation. In these games, losers chose not to take, in preference to the move (taking three tokens) that is specified by the backward-induction solution. This behaviour can be rationalized in various ways – for example, in terms of altruism on the part of the loser, respect for entitlements ('taking' might be viewed as a kind of stealing), a perception of not taking as the default option, or a belief that the winner might retaliate against a loser who took three tokens.[22] But notice that 47.9 per cent of games ended with allocations that were strictly worse for both players than both (3, 9) and (4, 6). In these cases, the behaviour induced by the Vendetta Game seems dysfunctional, both for each player individually and for the two players collectively.

Although the overall amount of taking was rather less than in the experiment reported by Bolle et al. (2014), the patterns of behaviour in the two experiments are qualitatively very similar. In the light of the evidence (discussed in Section 4) about the emotional states of players in Bolle et al.'s experiment, we conjecture that the mutually damaging behaviour we observed was driven by hot emotions of resentment and anger (and perhaps, in the case of winners who took at (3, 9), greed), and by each player's failure to anticipate the full effects of the other player's experiences of those emotions.[23]

---

[22] In the context of the experiment, a belief of this kind would in fact have been justified. Of the 49 winners who reached (4, 6), 14 chose to take at the first opportunity; of the 22 winners who, having declined the first opportunity, were given a second, 5 chose to take.

[23] The relatively large proportion of games ending at (2, 4) may be evidence of some minimally sophisticated reasoning by players who had experienced the effects of taking and counter-taking. If a loser takes at (2, 4), the winner has nothing to lose by retaliation, and the loser has no way of counter-retaliating.

We now consider cross-treatment comparisons. In the FR treatment, 44.2 per cent of pairs stuck at the initial (3, 9) allocation, compared with only 22.4 per cent in the URA treatment; this difference is significant in a two-tailed chi-squared test (χ^2= 5.928, p = 0.015).[24] Hence:

> *Result 1*: Consistently with Hypothesis 1, the Vendetta Game was more likely to end at the (3, 9) allocation in the FR treatment than in the URA treatment.

In the RE treatment, 38.9 per cent of pairs stuck at (3, 9), which is not significantly different from the FR treatment (χ^2= 0.249, p = 0.618).

Table 2 shows the average number of tickets held at the end of the game by each player in each treatment. For both winners and losers, these numbers are markedly greater in the FR treatment (5.87 and 2.40 respectively) than in the URA treatment (4.67 and 1.47); the RE treatment (5.58 and 2.36) is similar to FR. Figure 4 shows the cumulative distributions of final total holdings (i.e. the sum of the holdings of the two players) in the three treatments. Final total holdings in the URA and FR treatments are significantly different from one another (Mann-Whitney p = 0.046; the probability that a randomly selected URA subject has smaller final holdings than a randomly selected FR subject is 0.607). Hence:

> *Result 2*: Consistently with Hypothesis 2, more tokens were wasted in the URA treatment than in the FR treatment.

There is no significant difference in total final holdings between the RE and FR treatments (Mann-Whitney p = 0.679; the probability that a randomly selected RE subject has smaller final holdings than a randomly selected FR subject is 0.525).

[Figure 4 near here]

7.3 Sticking at the initial allocation by winners and losers

Table 3 gives a breakdown of winners' and losers' decisions at the initial allocation (3, 9). A player's decision is classified as $take_1$ if she took (any positive number of) tokens at her first opportunity. It is $pass_1$ if she passed (i.e., took no tokens) at her first opportunity, and the other player then immediately took tokens.[25] It is $take_2$ if she passed at her first opportunity, the other player then passed, and she then took at her second opportunity. It is $pass_2$ if she

[24] Throughout the paper, significance levels are reported for two-tailed tests.

[25] If the other player was the winner, this was that player's first opportunity to take; if he was the loser, it was his second.

passed at her first opportunity, the other player then passed, and she then passed at her second opportunity. If a loser takes at her first opportunity, the winner faces no decision problem and so is not given any classification. Notice that if a player's decision is classified as $pass_1$, we do not know how she would have used a second opportunity to take. But a $take_1$ or $take_2$ player has unambiguously chosen to take, and a $pass_2$ player has unambiguously chosen to pass. We therefore define the *sticking rate* for a given role (loser or winner) as the number of $pass_2$ decisions as a proportion of the total of $take_1$, $take_2$ and $pass_2$; n is the total number of decisions of all four types (i.e. the number of players in the relevant role who had at least one opportunity to take).

[Table 3 near here]

For losers, the sticking rate is higher in the FR treatment (47.1 per cent) than in the URA treatment (30.4 per cent); this difference is significant at the 10 per cent level (χ^2= 3.150, p = 0.076).[26] But contrary to our prior expectation, we find a significant difference in the same direction for winners.[27] The sticking rate is 92.0 per cent for FR winners and 68.4 per cent for URA winners (χ^2= 4.035, p = 0.045).[28] Hence:

> *Result 3*: Consistently with Hypothesis 3, we find weak evidence that losers were more likely to stick at (3, 9) in the FR treatment than in the URA treatment. But, contrary to that hypothesis, winners were more likely to stick at (3, 9) in the FR treatment.

Comparisons of sticking rates between FR and RE treatments show no significant difference either for losers (χ^2= 0.573, p = 0.450) or for winners (χ^2= 1.181, p = 0.277).

7.4 Winners' and losers' taking propensities

As explained in Section 6, an ideal test for cross-treatment differences in taking propensities would be based on separate comparisons at each decision node. However, since the Vendetta Game has 82 decision nodes for winners and 46 for losers, disaggregating behaviour to the decision-node level would produce numbers that are much too small for useful analysis. Our

[26] Given our sample size, with a type 1 error rate α = 0.05 and power = 0.80, this test can pick up an effect size of approximately 0.27. The observed effect size is 0.34.

[27] Sticking rates for winners should be interpreted with some caution, because of the small numbers of winners who chose to take at (3, 9). Intuitively, a winner who takes at (3, 9) is making a move that is both unprovoked and rash: it leaves the loser with no tokens, and with opportunities to retaliate.

[28] Given our sample size, with a type 1 error rate α = 0.05 and power = 0.80, this test can pick up an effect size of approximately 0.42. The observed effect size is 0.88.

analysis is therefore based on a coarser disaggregation of the data. It represents our best attempt to disentangle the behaviour of losers and winners, but should be viewed as exploratory.

We define a *taking opportunity* for a given player as a set of decision nodes for that player, specified by the current allocation of tokens and (if the allocation is one at which it is possible for both loser and winner to take tokens) whether the relevant player is facing her first or second opportunity to take. (At an allocation such as (0, 6), at which only one player is able to take, the distinction between first and second opportunities has no strategic significance.) Using L and W subscripts to refer to loser and winner, and 1 and 2 subscripts to refer to first and second opportunities, the set of taking opportunities for the loser is O_L = $\{(3, 9)_{L1}, (3, 9)_{L2}, (4, 6)_{L1}, (4, 6)_{L2}, (5, 3)_{L1}, (5, 3)_{L2}, (0, 10)_L, (1, 7)_L, (2, 4)_L\}$; the corresponding set for the winner is O_W = $\{(3, 9)_{W1}, (3, 9)_{W2}, (4, 6)_{W1}, (4, 6)_{W2}, (5, 3)_{W1}, (5, 3)_{W2}, (6, 0)_W, (3, 1)_W\}$. In defining taking opportunities in this way, we ignore the history of how a given allocation was reached. If (as is the case for (1, 7), (2, 4) and (3, 1)) a given allocation can be reached by an immediately preceding taking move by either player, we implicitly ignore whether, at that allocation, the loser or the winner has the first opportunity to take. A full breakdown of behaviour at each taking opportunity is reported in Tables A1–A4 in Appendix 2.

Using data from these tables we construct an index of 'excess taking' for each player in the FR and URA treatments. Intuitively, this index measures each player's taking propensity in the game as a whole, relative to the average taking propensity of players in the same role (loser or winner) in those treatments, taken together. We now explain how this index is defined for losers; the definition for winners follows the same principles.

Consider any taking opportunity k in O_L, and any individual i who played as a loser in the FR or URA treatment. Let a_k be the observed average number of tokens taken by players who faced that taking opportunity in the FR and URA treatments, taken together. We define a variable t_{ik} as follows. If i faced opportunity k, t_{ik} is the number of tokens taken by i at k; otherwise, $t_{ik} = a_k$. The null for a test of part (a) of Hypothesis 4 is that the behaviour strategies of losers in the two treatments are drawn from the same distribution. Under this null hypothesis, the expected value of $t_{ik} - a_k$ for any given k is zero in both treatments, irrespective of whether winners' behaviour differs between those treatments. Suppose that, for each taking opportunity in O_L, we fix any strictly positive and finite weight q_k, and then define $X_i = \sum_{k \in O_L}(t_{ik} - a_k)q_k$. Whatever weights are used, $\Sigma_i(t_{ik} - a_k) \equiv 0$ for all k, and

hence X_i necessarily has a mean of zero for the two treatments taken together. Under the null hypothesis, the expected value of X_i is zero in each treatment taken separately.

As a convention, we define the *index of excess taking* for each loser *i* as the value of X_i if, for each taking opportunity *k*, q_k is the proportion of FR and URA games in which that opportunity *k* was faced. Intuitively, this index weights each taking opportunity by the probability that it is reached in an 'average' game. The value of this index for any given player *i* can then be thought of as an estimate of the difference between *i*'s taking propensity and the taking propensity of an 'average' player, based only on observations of *i*'s behaviour. We interpret part (a) of Hypothesis 4 as predicting that indices of excess taking for losers are higher in the URA treatment than in the FR treatment. Part (b) is interpreted as predicting that indices of excess taking for winners are lower in the URA treatment.

Our tests of these hypotheses are reported in the first column of Table 4. The entries in the table report, separately for losers and winners, the difference between the mean value of the index of excess taking in the URA and FR treatments; positive values indicate a greater taking propensity in the URA treatment. The test statistic is the *p*-value for a *t*-test with bootstrap. Recall that, summing over the behaviour of winners and losers, total taking in the URA treatment (5.11 tokens per game) was significantly greater than in the FR treatment (3.73 tokens per game). Roughly speaking, a comparison between the 'loser' and 'winner' entries in the first column of Table 4 is informative about the relative contributions of losers and winners to this overall difference in taking behaviour. For losers, as one would expect, the propensity to take is greater in the URA treatment than in the FR treatment, but the difference is not statistically significant. More surprisingly, the propensity to take *by winners* is also greater in the URA treatment, and this difference is significant at the 10 per cent level. Hence:

> *Result 4:* Our index-based test finds no significant support for Hypothesis 4(a), i.e. the prediction that losers' taking propensities are greater in the URA treatment than in the FR treatment. We find weak evidence that, contrary to Hypothesis 4(b), winners' taking propensities are greater in the URA treatment.

The second column of Table 4 reports similar comparisons between propensities to take in the RE and FR treatments. There are no significant differences between treatments, either for losers or winners.

[Table 4 near here]

8 Discussion

The main objective of our experiment was to investigate individuals' attitudes to inequalities of outcome that have not been directly chosen by anyone, but instead result from rule-governed interactions in which those individuals have participated and in which each of them has acted on self-interest. The experiment was designed to elicit responses mediated by 'hot' emotions of resentment and anger about inequality rather than by 'cool' social preferences or distributional judgements. Our main finding (expressed in Results 1 and 2) is that inequalities are perceived as more acceptable if the interaction that generated them gave the two participants equal opportunities than if it was biased towards the individual who received the higher payoff from it. None of our measures of acceptability revealed any significant difference between the FR and RE treatments. In other words, inequalities generated by self-interested behaviour in a procedurally fair competition whose outcomes were largely determined by luck were perceived in much the same way as inequalities that could be interpreted as rewards for differences in effort. We interpret this as evidence that people are relatively tolerant of inequalities that result from interactions that are governed by procedurally fair rules, even if those inequalities are not perceived as rewarding effort or talent.

Our initial expectation was that, in the URA treatment, the driving force for wealth-destruction in the Vendetta Game would be resentment and anger *felt by the losers of the Search Competition Game* – that is, by players who had been disadvantaged by the unfair rules under which that game had been played. We expected an opposite effect for winners in that treatment. Our intuition was that winners would realise that they had already benefited from unfair rules, and so would be less likely to initiate taking than in the FR treatment, and less likely to retaliate against taking moves by their co-players. Surprisingly, however, we found no evidence that advantaged winners were inhibited about taking. We recognise that our findings about the separate behaviour of losers and winners are less firm than those about their combined behaviour, because of the smaller sample sizes involved and (in the case of the exploratory analysis leading to Result 4) the difficulty of disentangling the behaviour of players who interact in a sequential game. But Results 3 and 4 both suggest that winners were *more* likely to take in the URA treatment than in the FR treatment.

As this effect was unexpected, we can offer only post hoc conjectures about how it might be explained. Why would the fact that a person has already benefitted from unfairness make him or her more willing to act aggressively against the victim of that unfairness? There

may be some analogy with psychology experiments which have famously shown that ordinary people can act with (what they have reason to believe to be) cruelty towards others when that cruelty has been licensed by an apparently legitimate authority (Milgram, 1963; Zimbardo, 2007). In a less extreme form, our Unfair Rule treatments may have been perceived as licensing subjects to override their inhibitions against taking advantage of their co-participants. Or, putting this more charitably, interacting with others in procedures that are governed by fair rules may tend to prime social norms that impose fairness constraints on individual behaviour. By failing to activate such norms, the Unfair Rule treatments may have made it psychologically easier for both losers and winners to engage in behaviour that was collectively dysfunctional.

To repeat, we do not want to place too much weight on our findings about winners and losers considered separately. We believe that the main contribution of this paper is to highlight the importance of procedural fairness in the sense of equality of opportunity, as opposed to mere randomness. Individuals have negative attitudes to procedural unfairness, independently of any preferences for social welfare, equality of outcomes, or rewards for effort. When interactions take place under unfair rules, those attitudes may induce emotions of anger, resentment and greed that are revealed in behaviour that is socially costly.

Table 1: Acceptance of first card by winners and losers

	FR	
	(1)	
	β	ME
Card 1	0.134***	0.028***
	(0.019)	(0.004)
Winner	-0.301	-0.062
	(0.456)	(0.093)
Constant	-7.817***	
	(1.121)	
Observations	416	
Pseudo R^2	0.564	
LR chi-square	81.283	
Prob > chi-square	0.000	
Baseline predicted probability	0.420	

Notes: ∗ 10% level, ∗∗ 5% level, ∗∗∗ 1% level. The dependent variable in the model is a dummy equal to 1 if the subject accepted and 0 if the subject rejected the first card in the search competition game. *Card 1* takes the value of the first card dealt to the subject in the first four games. *Winner* is equal to 1 if the subject is the overall winner of the series of search competition games, and 0 if the subject is not the overall winner. The left column reports coefficients, and the right column reports marginal effects. Results for the model are based on random effect logit estimations in which subject-specific random effects are controlled.

Table 2: Outcomes of Vendetta Games

	Fair Rule (FR)	Unfair Rule with advantaged winner (URA)	Unfair Rule with disadvantaged winner (URD)	Real Effort (RE)	All
Allocation of tickets:					
(3, 9)	23 (44.2%)	13 (22.4%)	3 (17.7%)	14 (38.9%)	53 (32.5%)
(4, 6)	4 (7.7%)	10 (17.2%)	3 (17.7%)	4 (11.1%)	21 (12.9%)
(5, 3)	4 (7.7%)	1 (1.7%)	0 (0%)	3 (8.3%)	8 (4.9%)
(6, 0)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
(0, 10)	0 (0%)	0 (0%)	1 (5.9%)	0 (0%)	1 (0.6%)
(1, 7)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
(2, 4)	10 (19.2%)	16 (27.6%)	2 (11.8%)	6 (16.7%)	34 (20.9%)
(3, 1)	0 (0%)	2 (3.5%)	0 (0%)	0 (0%)	2 (1.2%)
(0, 2)	11 (21.2%)	16 (27.6%)	8 (47.6%)	9 (25.0%)	44 (27.0%)
Average number of tickets held by the winner	5.87	4.79	4.67	5.58	5.30
Average number of tickets held by the loser	2.40	2.10	1.47	2.36	2.19
Number of pairs	52	58	17	36	163

Table 3: Sticking at the initial allocation

Loser's decision at (3, 9)

	FR	URA	URD	RE
n	52	58	17	36
$take_1$	24	35	12	17
$pass_1$	1	2	1	0
$take_2$	3	4	1	5
$pass_2$	24	17	3	14
sticking rate (%)	47.1	30.4	18.8	38.9

Winner's decision at (3, 9)

	FR	URA	URD	RE
n	28	23	5	19
$take_1$	1	2	1	0
$pass_1$	3	4	1	5
$take_2$	1	4	0	0
$pass_2$	23	13	3	14
sticking rate (%)	92.0	68.4	75.0	100.0

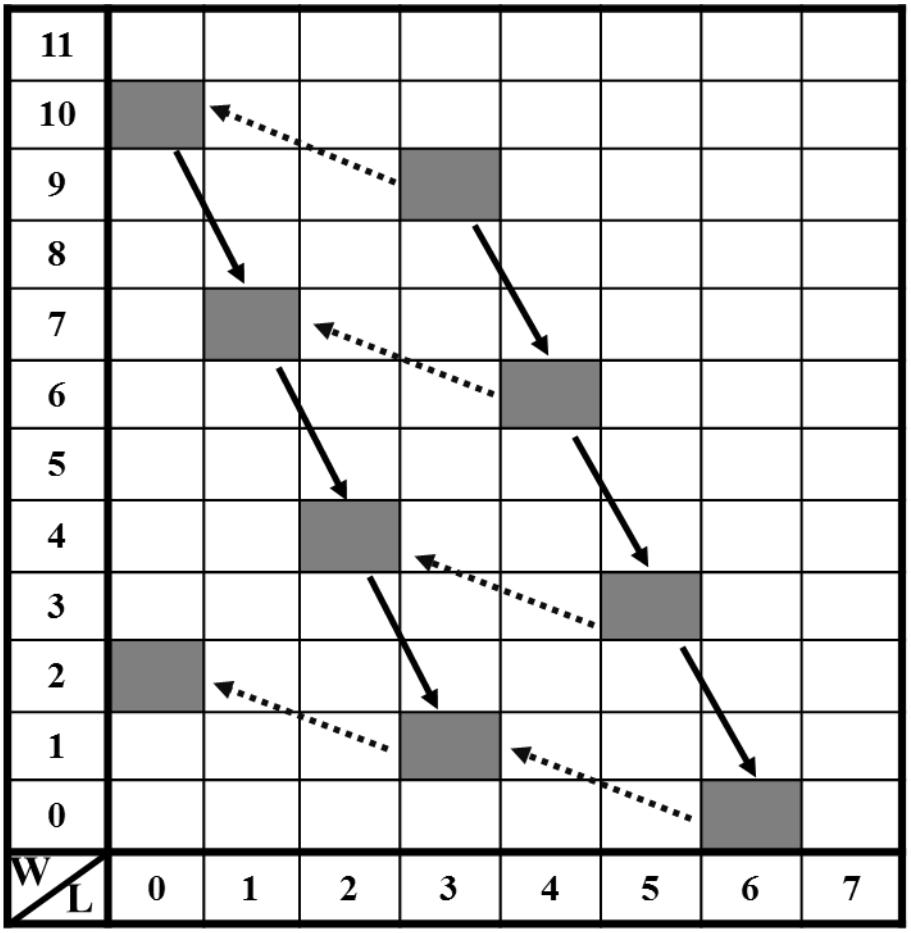
Notes: FR = Fair Rule treatment; URA = Unfair Rule treatment with advantaged winner; URD = Unfair Rule treatment with disadvantaged winner; RE = Real Effort Treatment. Sticking rate = $pass_2$ as percentage of ($take_1$ + $take_2$ + $pass_2$).

Table 4: Excess taking relative to Fair Rule treatment

	URA	RE
Loser	0.553 ($p = 0.354$)	–0.291 ($p = 0.671$)
Winner	0.147 ($p = 0.083$)	0.105 ($p = 0.457$)

Notes: URA = Unfair Rule treatment with advantaged winners; RE = Real Effort Treatment. This table shows the net excess of the mean index of excess taking for the URA (respectively RE) treatment over the mean index for the Fair Rule treatment; p values are from a *t*-test with bootstrap.

Figure 1: Schematic representation of Vendetta Game



Notes: The number of tickets held by the loser (winner) is shown on the horizontal (vertical) axis. The shaded cells are ticket distributions that can be reached in the game. Continuous sequences of solid (broken) arrows correspond with possible moves by the loser (winner).

Figure 2: Screen shot of Vendetta Game

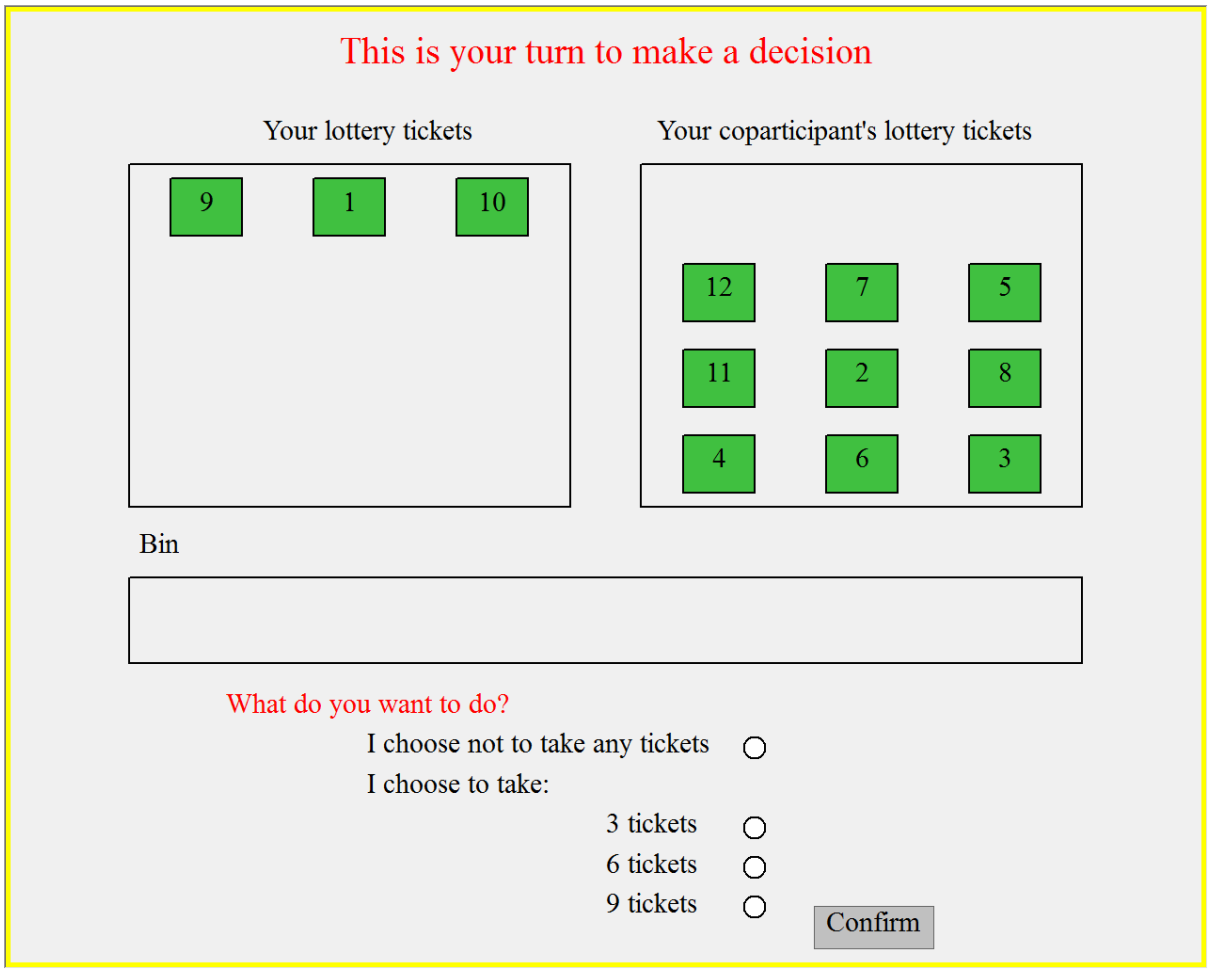


Figure 3: Probabilities of accepting first card in Fair Rule competition

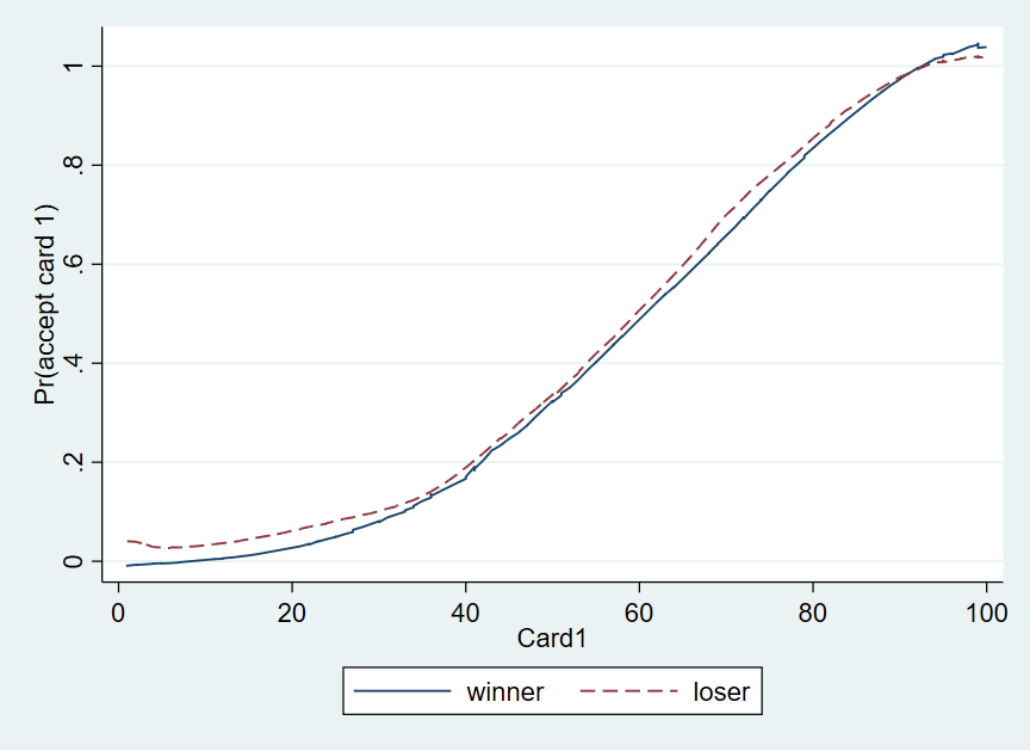
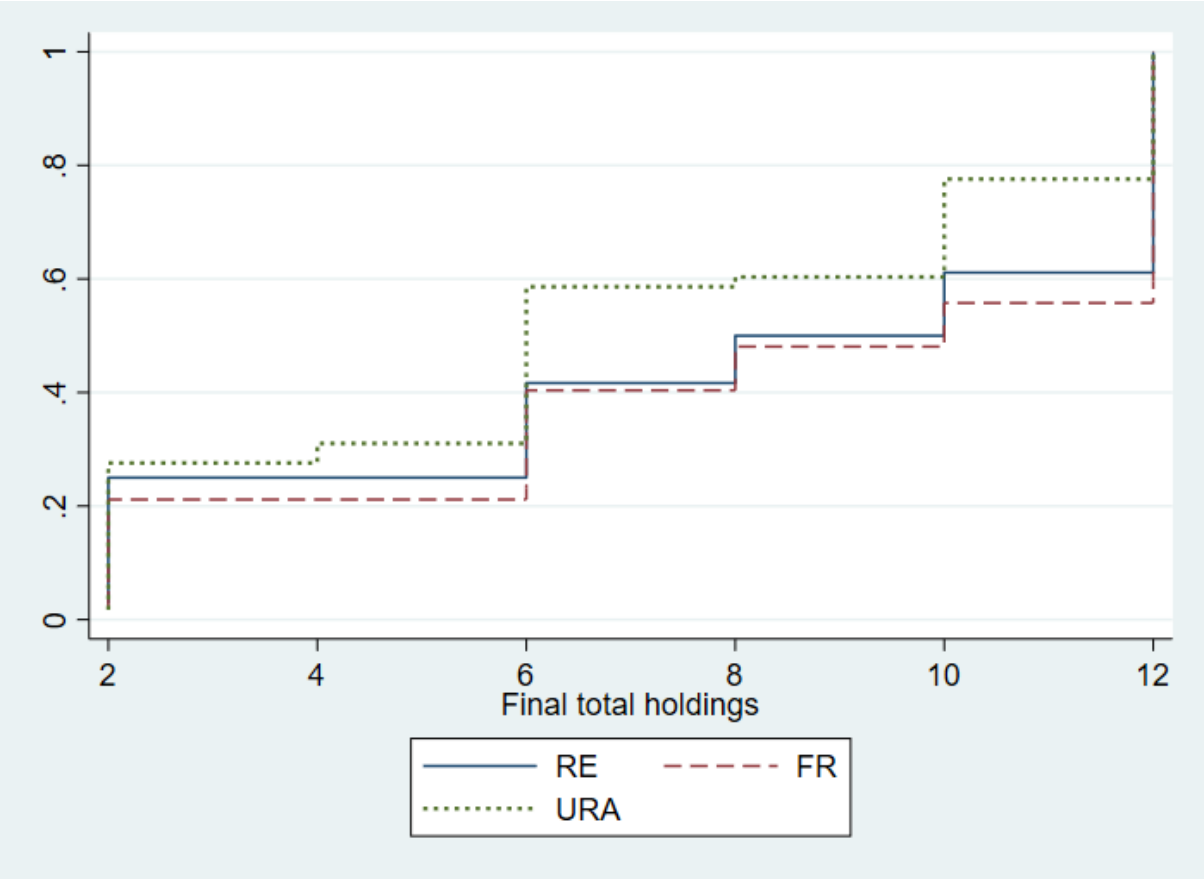


Figure 4: Cumulative distributions of final total holdings



Appendix 1: Proofs of theoretical results

Proof of Proposition 1

Consider any $n \times n$ game (with isomorphically numbered strategies) that is procedurally biased towards player1. Let $T(k, l)$ denote the set $\{(s_{1k}, s_{2l}), (s_{1l}, s_{2k})\}$. Notice that $T(k, l) \equiv T(l, k)$. We will say that $T(k, l)$ is *transpositionally biased* towards player 1 if $\pi_1(l, k) \geq \pi_2(k, l)$ and $\pi_2(l, k) \leq \pi_1(k, l)$, and at least one of these inequalities is strict. It follows immediately from the definition of procedural bias that there is no (k, l) such that $T(k, l)$ is transpositionally biased towards player 2, which proves part (i) of result 1, and that there is some (k', l') such that $T(k', l')$ is transpositionally biased towards player 1. Suppose $\pi_1(k', l') \leq \pi_2(k', l')$, i.e., $(s_{1k'}, s_{2l'})$ does not induce an outcome equality that favours player 1. Chaining this inequality with the inequalities implied by transpositional bias gives $\pi_1(l', k') > \pi_2(l', k')$, i.e., $(s_{1l'}, s_{2k'})$ induces an outcome inequality that favours player 1. So at least one of $(s_{1k'}, s_{2l'})$ and $(s_{1l'}, s_{2k'})$ induces an outcome inequality that favours player 1 and (since $T(k', l')$ is transpositionally biased towards player 1) imposes an unbalanced inequality on player 2.□

Proof of Proposition 2

Consider any game G in which there are m strategies for player 1 and n strategies for player 2, with $m \geq n$. If $m > n$, our definitions imply that G cannot have equal opportunity or be biased towards player 2. Now suppose $m = n$. Define *payoff advantage* to player 1 as the sum of all possible payoffs to player 1 minus the sum of all possible payoffs to player 2. If G has equal opportunity, payoff advantage to player 1 is necessarily zero. If it is procedurally biased towards player 1, payoff advantage to player 1 is strictly positive (and therefore payoff advantage to player 2 is strictly negative). So G cannot fall into more than one of our classifications. □

Proof of Proposition 3

Consider any game G that gives equality of opportunity [respectively: is procedurally biased towards player 1]. Let the number of pure strategies for players 1 and 2 be m and n. By definition, $m = n$ [$m \geq n$], and there is some renumbering of pure strategies such that $\pi_1(s_{1k}, s_{2l}) = \pi_{2j}(s_{1l}, s_{2k})$ [$\pi_1(s_{1k}, s_{2l}) \geq \pi_{2j}(s_{1l}, s_{2k})$] holds for all k [for all $k \leq n$] and for all l. Consider any strategy mix σ_2 for player 2, defined as a probability distribution over the renumbered strategies. Let $\sigma_1' = \sigma_2$. By simple algebra, player 1's expected payoff from (σ_1', σ_2) is equal

to [no less than] player 2's. Thus, in the zero-sum special case, player 1's payoff from (σ_1', σ_2) is zero [non-negative]. So player 1's best response to σ_2 must give a zero [non-negative] payoff to player 1, and a zero [non-positive] payoff to player 2. Since this holds for all σ_2, it follows that player 1's payoff is equal to [no less than] player 2's in every Nash equilibrium.

□

Appendix 2: Behaviour at each taking opportunity

These tables show, for each taking opportunity in each treatment, the number of players who faced that opportunity (*n*), and the distribution of these players' decisions between 'take 0', 'take 3', 'take 6' and 'take 9'.

Table A1: Fair Rule Treatment (FR)

Loser					
Taking opportunity	*n*	take 0	take 3	take 6	take 9
$(3,9)_{L1}$	52	28	12	7	5
$(3,9)_{L2}$	27	24	1	2	0
$(4,6)_{L1}$	10	8	2	0	
$(4,6)_{L2}$	6	4	2	0	
$(5,3)_{L1}$	5	5	0		
$(5,3)_{L2}$	4	4	0		
$(0,10)_{L}$	2	0	0	0	2
$(1,7)_{L}$	5	0	4	1	
$(2,4)_{L}$	13	0	3		
Winner					
Taking opportunity	*n*	take 0	take3	take 6	
$(3,9)_{W1}$	28	27	1		
$(3,9)_{W2}$	24	23	1		
$(4,6)_{W1}$	13	10	3		
$(4,6)_{W2}$	8	6	2		
$(5,3)_{W1}$	13	5	8		
$(5,3)_{W2}$	5	4	1		
$(6,0)_{W}$	5	0	1	4	
$(3,1)_{W}$	7	0	7		

Table A2: Unfair Rule treatment with advantaged winner (URA)

Loser					
Taking opportunity	*n*	take 0	take 3	take 6	take 9
$(3,9)_{L1}$	58	23	18	12	5
$(3,9)_{L2}$	21	17	3	1	0
$(4,6)_{L1}$	15	12	3	0	
$(4,6)_{L2}$	10	10	0	0	
$(5,3)_{L1}$	2	2	0		
$(5,3)_{L2}$	2	1	1		
$(0,10)_{L}$	6	0	0	2	4
$(1,7)_{L}$	8	0	5	3	
$(2,4)_{L}$	21	16	5		
Winner					
Taking opportunity	*n*	take 0	take 3	take 6	
$(3,9)_{W1}$	23	21	2		
$(3,9)_{W2}$	17	13	4		
$(4,6)_{W1}$	21	15	6		
$(4,6)_{W2}$	12	10	2		
$(5,3)_{W1}$	16	2	14		
$(5,3)_{W2}$	2	2	0		
$(6,0)_{W}$	6	0	1	5	
$(3,1)_{W}$	13	2	11		

Table A3: Unfair Rule treatment with disadvantaged winner (URD)

Loser					
Taking opportunity	*n*	take 0	take 3	take 6	take 9
$(3,9)_{L1}$	17	5	5	3	4
$(3,9)_{L2}$	4	3	0	1	0
$(4,6)_{L1}$	3	3	0	0	
$(4,6)_{L2}$	3	3	0	0	
$(5,3)_{L1}$	0	0	0		
$(5,3)_{L2}$	0	0	0		
$(0,10)_{L}$	1	1	0	0	0
$(1,7)_{L}$	2	0	0	2	
$(2,4)_{L}$	4	2	2		
Winner					
Taking opportunity	*n*	take 0	take 3	take 6	
$(3,9)_{W1}$	5	4	1		
$(3,9)_{W2}$	3	3	0		
$(4,6)_{W1}$	5	3	2		
$(4,6)_{W2}$	3	3	0		
$(5,3)_{W1}$	4	0	4		
$(5,3)_{W2}$	0	0	0		
$(6,0)_{W}$	4	0	0	4	
$(3,1)_{W}$	4	0	4		

Table A4: Real Effort Treatment (RE)

Loser					
Taking opportunity	*n*	take 0	take 3	take 6	take 9
$(3,9)_{L1}$	36	19	6	9	2
$(3,9)_{L2}$	19	14	4	1	0
$(4,6)_{L1}$	7	5	2	0	
$(4,6)_{L2}$	4	4	0	0	
$(5,3)_{L1}$	5	3	2		
$(5,3)_{L2}$	3	3	0		
$(0,10)_{L}$	0	0	0	0	0
$(1,7)_{L}$	4	0	3	1	
$(2,4)_{L}$	10	6	4		
Winner					
Taking opportunity	*n*	take 0	take 3	take 6	
$(3,9)_{W1}$	19	19	0		
$(3,9)_{W2}$	14	14	0		
$(4,6)_{W1}$	10	7	3		
$(4,6)_{W2}$	5	4	1		
$(5,3)_{W1}$	12	5	7		
$(5,3)_{W2}$	3	3	0		
$(6,0)_{W}$	4	0	1	3	
$(3,1)_{W}$	6	0	6		

References

Aldashev, Gani, Georg Kirchsteiger and Alexander Sebald (2015). Assignment procedure biases in randomised policy experiments. *Economic Journal* 127: 873–895.

Andreoni, James and John Miller (2002). Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* 70: 737–753.

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

Bolle, Friedel, Jonathan Tan and Daniel Zizzo (2014). Vendettas. *American Economic Journal: Microeconomics*, 6: 93–130.

Bolton, Gary, Jordi Brandts and Axel Ockenfels (2005). Fair procedures: evidence from games involving lotteries. *Economic Journal*, 115 (506), 1054–1076.

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

Buchanan, James (1964). What should economists do? *Southern Economic Journal* 30: 213–222.

Burrows, Paul and Graham Loomes (1994). The impact of fairness on bargaining behaviour. In John Hey (ed.), *Experimental Economics*, Heidelberg: Physica-Verlag HD, 21–41.

Cappelen, Alexander, Astri Hole, Erik Sørensen, and Bertil Tungodden (2007). The pluralism of fairness ideas: An experimental approach. *American Economic Review*, 97 (3), 818–827.

Cappelen, Alexander W., James Konow, Erik Ø. Sorensen and Bertil Tungodden (2013). Just luck: an experimental study of risk-taking and fairness. *American Economic Review* 103(4): 1398–1413

Charness, Gary and Matthew Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117: 817–869.

Chlaß, Nadine, Werner Güth and Topi Miettinen (2019). Purely procedural preferences: beyond procedural equity and reciprocity. *European Journal of Political Economy* 59: 108–128.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman (2006). Can second-order

punishment deter perverse punishment? *Experimental Economics* 9: 265–279.

Denant-Boemont, Laurent, David Masclet and Charles Noussair (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33: 145–167.

Dufwenberg, Martin and Georg Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior* 47, 268–98.

Erkal, Nisran, Lata Gangadharan and Nikos Nikiforakis (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review* 101: 3330–3348.

Fahr, René and Bernd Irlenbusch (2000). Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment. *Economics Letters* 66: 275–282.

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

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

Fehr, Ernst and Simon Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–94.

Harsanyi, John and Reinhard Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: The MIT Press.

Hayek, Friedrich (1976). *Law, Legislation and Liberty. Vol 2: The Mirage of Social Justice.* Chicago: University of Chicago Press.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat and Vernon L. Smith (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7 (3), 346–380.

Hoffman, Elizabeth and Matthew L. Spitzer (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies*, 14 (2), 259–297.

Isoni, Andrea, Anders Poulsen, Robert Sugden and Kei Tsutsui (2013). Focal points in tacit bargaining problems: experimental evidence. *European Economic Review* 59(April): 167–188.

Isoni, Andrea, Anders Poulsen, Robert Sugden and Kei Tsutsui (2014). Efficiency, equality and labelling: an experimental investigation of focal points in explicit bargaining. *American Economic Review* 104: 3256–3287.

Konow, James (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90 (4), 1072–1092.

Krawczyk, Michal (2011). A model of procedural and distributive fairness. *Theory and Decision* 70: 111–128.

Levine, David (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1: 593–622.

McCabe, Kevin, Mary Rigdon and Vernon Smith (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization* 52: 267–275.

Machina, Mark (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27: 1622–1668.

Milgram, Stanley (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67 (4), 371–378.

Mollerstrom, Johanna, Bjørn-Atle Reme and Erik Ø. Sørensen (2015). Luck, choice and responsibility: an experimental study of fairness views. *Journal of Public Economics* 131: 33–40.

Nikiforakis, Nikos (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* 92: 91–112.

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

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review* 81: 1068–1095.

Rotter, Julian B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* 80: 1–28.

Rawls, John (1971). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Roemer, John E. (1998). *Equality of Opportunity*. Cambridge, Mass.: Harvard University Press.

Ruffle, Bradley J. (1988). More is better, but fair is fair: tipping in dictator and ultimatum games. *Games and Economic Behavior*, 23 (2), 247–265.

Saito, Koto (2013). Social preferences under risk: equality of opportunity versus equality of outcome. *American Economic Review* 103(7): 3084–3101.

Sebald, Alexander (2010). Attribution and reciprocity. *Games and Economic Behavior* 68: 339–352.

Schelling, Thomas (1960). *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Schurter, Karl and Bart Wilson (2009). Justice and fairness in the Dictator Game. *Southern Economic Journal* 76: 130-145.

Sugden, Robert (2004). Living with unfairness: the limits of equality of opportunity in a market economy. *Social Choice and Welfare* 22: 211–236.

Trautmann, Stefan (2009). A tractable model of process fairness under risk. *Journal of Economic Psychology* 30: 803–813.

Zimbardo, Philip (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. New York: Random House.