Introducing Social Norms in Game Theory*

Raúl López-Pérez†

December 2006

Abstract

This paper explicitly introduces norms in games, assuming that they shape (some) players' utility. People feel badly when they deviate from a binding norm, and the less other players deviate, the more badly they feel. Further, people anger at transgressors and get pleasure from punishing them. I then study how social norms and emotions affect competition, cooperation, and punishment in a variety of games. The model is consistent with abundant experimental evidence that alternative models of social preferences cannot explain.

Keywords: Competition, Cooperation, Emotions, Punishment, Reciprocity, Social Norms. JEL classification numbers: C72, D02, D62, D64, Z13.

---

* I am indebted to Ernst Fehr, Urs Fischbacher, Alexander Kritikos, Michael Kosfeld, Michael Näf, Christian Zehnder, participants at the March 2006 Greifensee seminar and the May 2006 Conference in Capua (Italy); and one anonymous referee for very helpful comments. Part of this research was conducted while visiting the Institute for Empirical Research in Economics at Zurich, and I would like to thank their members for their hospitality. I also gratefully acknowledge financial support from the European Union through the ENABLE Marie Curie Research Training Network.

† Universidad Autónoma de Madrid, Department of Economic Analysis, 28049 Madrid, Spain. E-mail: raul.lopez@uam.es

1. Introduction

A *norm* is a rule that prescribes behavior –that is, any statement of the form 'in situation *x*, you *ought to* do *y*'. For instance, all laws, codes of honor, moral principles, or religious commandments are norms according to this definition.[1]

Norms are a fundamental ingredient of human societies. Indeed, prominent social researchers like Emile Durkheim or Talcott Parsons have emphasized that *human behavior is shaped by norms* and that (some) norms foster cooperation and pro-social behavior, thus facilitating the attainment of social order –see also Arrow (1974), and Elster (1989).

In addition, social psychologists and other social researchers have pointed out that *norms affect behavior because they first affect motivation*. When someone internalizes a norm (Elster, 1989; Becker, 1996; Gintis, 2003), she becomes emotionally attached to it, that is, painful *emotions* get triggered when she deviates from it (shame, guilt), or when others deviate (anger, indignation). As a result, people tend to behave according to internalized norms in order to avoid (1) remorse (*internal punishment*) or (2) sanctions from an angry party (*external punishment*). [2]

This paper formalizes these ideas and investigates how norms affect behavior using the standard, well-known apparatus of preferences, rationality, games, and equilibrium concepts. To model the idea that (some) people care about norms, however, the model abandons the standard hypothesis that *all* players are selfish –i.e., *exclusively* motivated by their *own* consumption and leisure (material interest).

The results in this paper will be of particular interest for behavioral and experimental economists, who have gathered in the last 30 years an impressive amount of evidence contradicting the selfishness hypothesis. As a particular application of the model, I focus on a *norm of distributive justice* that exhibits a concern for both efficiency *and* maximin (the *EM-norm*), and show that if some agents have internalized *only* that norm, while remaining agents do not care about any norm at all, the model is then consistent with a large and varied array of well-replicated experimental results.[3]

---

[1] This definition is indeed very wide-ranging, and one may find more restrictive ones in the literature on norms -I survey this literature in López-Pérez (2005).

[2] The role played by the external punishment is particularly important when we consider *social norms* –i.e., norms that have been internalized by sufficiently many people *in a group* and are hence "partly sustained by their approval and disapproval" (Elster, 1989, p. 99). We must note two things in this regard: (1) People may internalize norms *even* if they are not social (see the discussion on private norms by Elster, 1989), and (2) there exist norms or morals which *nobody* cares about (at least if we circumscribe our attention to a particular group or society): Thus, the norm to wear black in a funeral was not a social norm in traditional China.

[3] In this paper, efficiency refers to the sum of players' material payoffs, and not to Pareto efficiency. Maximin or need refers to the worst-off player's income. López-Pérez (2004) study alternative norms (like egalitarian ones).

The model explains, for instance, why people cooperate conditionally (more generally, the model predicts that people respect binding norms in a *reciprocal* manner –i.e., they are more willing to comply if others comply as well), why *first movers* in a sequential social dilemma cooperate significantly more than players of a simultaneous dilemma, why punishment and cooperation depend on the menu of choices, why passive players are usually not punished, or why competitive markets induce principled people to behave as self-interested ones do.

The model is related to recent theories of social preferences and reciprocity, which also relax the selfishness hypothesis.[4] Rabin (1993) models reciprocity (that is, the idea that people are kind (unkind) to those who are kind (unkind) to them), and Dufwenberg and Kirchsteiger (2004) extend Rabin's ideas to multiple-player and dynamic games. Levine (1998) assumes type-based altruism and spitefulness, and both Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) propose models of inequity-averse players. Finally, Charness and Rabin (2002) and Falk and Fischbacher (2006) introduce both reciprocity and distributional concerns.[5]

Almost no one of these models explicitly introduces norms, the only exception being the reciprocity model of Charness and Rabin (2002). This model is very complex and presents a series of problems, though.[6] Indeed, Charness and Rabin do not see it 'as being primarily useful in its current form for calibrating experimental data, but rather as providing progress in conceptualizing what we observe in experiments' (p. 851). In line with this idea, one might view my model as a tractable version of theirs.

The model here has a number of advantages with respect to the other models. First, it explains better the experimental evidence in the range of games that I analyze here and in López-Pérez (2004). One crucial reason for this is that it assumes that agents care about *history*. More precisely, people's utility depends on whether others (or themselves) misbehaved –i.e., deviated from a binding norm- in the past. Hence, the model takes into account *procedural* justice. This is a key difference with outcome-based utility models like the inequity aversion ones.

Second, and contrary to Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) the model here is *not* based on the Psychological Game Theory of Geanakoplos et al. (1989), so that agents' utility does not depend on their beliefs. That makes the model much more parsimonious. Further, and contrary to Levine (1998), agents' utility does not depend on the co-players' types, thus significantly reducing

[4] Fehr and Schmidt (2006), Camerer (2003), and López-Pérez (2004) survey this literature.

[5] Except the seminal paper by Rabin (1993), all these models are quite general in that they apply to a large class of games. There exist some interesting theories which are more restrictive –thus, Cox, Friedman and Gjerstad (2006) only applies to sequential two-player games of perfect information. For expositional brevity, I do not consider them here –again, see Fehr and Schmidt (2006) on this.

[6] For a good discussion of this point, consult Fehr and Schmidt (2006, pp. 36-37).

the number of equilibria. In fact, my model predicts a unique *equilibrium outcome* in all the games that I analyze here, which is crucial to facilitate experimental testing.

Last, *but not least*, the model has a broader field of application because it explicitly introduces norms. Although I do not address such questions here, one might use it to explain why people tell the truth and punish cheaters contrary to their material interest, or why people follow rules of etiquette, or norms regulating sexual relations. The other models have troubles in explaining such behavior because they posit that utility depends on money allocations and/or on beliefs about such allocations -and it is unclear how, say, sexual intercourse may affect those things!

The rest of the paper is organized as follows. Section 2 describes the model and discusses its key assumptions. Section 3 studies how the EM-norm affects cooperation, competition, and punishment in different games. These predictions are summarized and briefly compared with those from other models in section 4. Section 5 concludes by mentioning some limitations of the theory, together with possible extensions.

2. A Model with Norms

Consider a n-player, extensive form game of perfect recall Γ. Let N = {1,…,n} denote the set of players, z denote a terminal node, $u_i(z)$ denote player i's utility payoff at z, and $x_i(z)$ denote player i's *monetary* payoff at z.[7] As $u_i(z)$ and $x_i(z)$ may differ for some players, it makes sense to distinguish between game Γ and its associated *lab game*. This is a mathematical object that contains the game form of Γ -i.e., all things that Γ comprises except utility payoffs $u_i(z)$ - *and* each monetary payoff $x_i(z)$.

2.1 Norms

Norms are *exogenous* rules that select actions in *lab games.* Let h denote an information set and $A(h)$ denote the set of available actions at h.

Definition 1: A norm is a nonempty correspondence $\Psi : h \rightarrow A(h)$ applying on any information set of any lab game, except on Nature's ones.

Throughout the paper, I will use indistinctively the following expressions: 'The norm selects action a at information set h', 'the norm commends to choose a at h', and 'according to the norm, (the relevant mover) should choose a at h.'

Given that norms select actions, a player is said to *respect* or *comply* with norm Ψ at h if (i) her choice at h is consistent with Ψ or if (ii) she is not the mover at h. Otherwise, she *deviates* from the norm. Suppose then that play reaches terminal node z. By considering all actions in the path of z, one may obtain the set of players who

[7] Apart of getting a monetary payment, players in some games might also consume goods and leisure in the history of z. In that case, $x_i(z)$ should represent the material utility that player i gets from consumption and money. I will not pursue this topic further here, though.

respected Ψ in the history of z, $R(\Psi, z)$, and its cardinality, $r(\Psi, z)$. If it is clear to which norm I refer, I will instead write $R(z)$ and $r(z)$.

2.2 Preferences

There are two types of players. S*elfish* players are standard money-maximizers who do not care about norms. To simplify, they are assumed to be risk-neutral, so that their utility function is

$u_i(z) = x_i(z)$.

In contrast, the utility of a *principled* player at z depends on the money earned $x_i(z)$ *and* the history of z. In other words, principled agents care about *what* they get and *how* they get it. Intuitively, different histories activate different emotions: If principled player A deviates from what an internalized norm commends then she feels ashamed or guilty, whereas if A complies but another player deviates then A feels angry at him.[8] More precisely, the utility function of a principled player i who has internalized norm Ψ equals

$$u_i(z) = \begin{cases} x_i(z) - \gamma \cdot r(z) & \text{if } i \notin R(z),\ (0 < \gamma) \\ x_i(z) - \alpha \cdot \max\limits_{j \notin R(z)} \{x_j(z)\} \cdot I(z) & \text{if } i \in R(z),\ (0 < \alpha), \end{cases}$$

where $I(z)$ is an indicator function that takes value 0 if nobody deviates –i.e., if $R(z) = N$ - and 1 otherwise.

2.3 The EM-Norm

One may think of infinite correspondences satisfying definition 1. As a particular example, consider a norm that selects any action pointing towards an efficient and maximin outcome, *conditional on others doing the same*. More formally, let $X(\Delta)$ denote the set of all *monetary* allocations of *lab game* Δ.

Definition 2: Allocation $x = \{x_1, ..., x_n\} \in X(\Delta)$ is efficient-cum-maximin (EM) if it maximizes function

$$F(x) = \sum_{i \in N} x_i + \delta \min_{i \in N} \{x_i\} \qquad (1)$$

over $X(\Delta)$, where $0 < \delta$. An EM-path of Δ is a path leading to an EM-allocation of Δ. An EM-action is an action that belongs to an EM-path.

Definition 3 (the EM-Norm): If h is on one EM-path, the EM-norm selects only the EM-actions in $A(h)$. Otherwise, the EM-norm selects the whole set $A(h)$.

In other words: As far as everybody respects the EM-norm, then one must strive to achieve an EM-allocation; but if it is known *for certain* that at least one player has deviated then any behavior is allowed –i.e., the norm is conditional in an extreme form. The reader

[8] Therefore, norms shape utility. This explains why norms are defined to apply on lab games and not on proper games: Circularity problems would appear if norms depended on utility payoffs (an ingredient of games) and at the same time affected utility.

can verify that this is truly a norm –i.e., a *nonempty* correspondence selecting at least one action at any information set of any lab game.[9]

Similar norms may be obtained by conveniently changing function (1). Thus, a crude egalitarian norm might correspond to the following function

$$F^{e}(x) = \min_{i \in N} \{x_i\} - \max_{i \in N} \{x_i\} . \quad (2)$$

The EM-norm is indeed extremely simple, and one might think of more sophisticated norms –for examples, consult López-Pérez (2005). For reasons that I mention later, however, it is posited that the EM-norm is the *only* norm that *all* principled players care about. Taking this into account, let ρ denote the fraction of principled players in the population –this parameter is common knowledge.

2.4 Information, Equilibrium Concept, and a Refinement

Unless otherwise noted, I assume for simplicity that each player's type is common knowledge. Taking into account this, and as I consider both simultaneous and sequential games, Subgame Perfect Equilibrium (SPE) is a natural solution concept to use.

We will concentrate our attention on pure strategy SPE. Now, some games may have multiple equilibria. In that case, and if at least one of the players is principled, I assume that the EM norm *shapes beliefs* by acting as a focal point (Schelling, 1960; Sugden, 1989). To formalize this, let s' denote a *pure strategy* SPE of a game, and $x(s')$ denote the associated vector of players' *monetary* payoffs.

Definition 4: Equilibrium s' is EM if $x(s')$ maximizes function (1) among *all* equilibria vectors $x(s)$ of the game.

To put it like that, an SPE is EM if it attains the 'fairest' equilibrium outcome (at least from the point of view of a principled player). Note that definition 4 can be easily extended to other norms of distributive justice by changing the corresponding function.

Assumption 1: A principled agent will play action a with some probability only if a is part of an EM *equilibrium* strategy –if any such equilibrium exists.

That is, *principled players* find obvious or prominent an EM equilibrium. Note well that the focal point acts in a heterogeneous way as it only affects expectations about *principled* players' behavior.

2.5 Discussion

Let me start with three remarks on principled types' utility function. First, parameter γ may be interpreted as an internalization index. Note that ceteris paribus the intensity of a deviator's bad feelings is assumed not to depend on the specific deviation she makes. That is, all deviations are equally 'bad'. Although this assumption is clearly

[9] If $X(\Delta)$ is not bounded then there might not be an EM-allocation, though. Consult López-Pérez (2005) on this point.

unrealistic, it greatly simplifies the model and suffices to explain many experimental facts. I come back to this issue in the conclusion.

Second, the more the people who respect the norm, the more badly a principled deviator feels. For simplicity, I have modeled this by means of a linear function, but any strictly increasing one would give the same qualitative results in the games I analyze. However, it is important for the results that no principled *deviator* feels badly at z if *all* the other players deviate as well. This implies that a principled player *never* complies with an internalized norm if (a) compliance is at odds with her material interest and if (b) she *expects* all other players to deviate, an implication that will be extensively used in the applications.[10]

Third, parameter α measures aggressiveness –more precisely, an angry player i is willing to spend α monetary units in order to reduce the best-off deviator's monetary payoff in one unit. Note that α is independent of the specific deviation that triggers the anger, a hypothesis that is again made for simplicity –in the conclusion I discuss this issue. For analogous reasons, I also assume that anger and the associated tendency impulse to retaliate focus on the best-off deviator if there are multiple deviators.

I pass now to consider a different question, that is, why should we assume that the EM-norm is the only norm that principled players care about? Note first that we must assume something about the specific norms that principled players have internalized in order to obtain precise behavioral predictions in games and test the model. Further, the number of norms should not be too high in order to keep the model tractable. Ideally, one simple norm should be able to explain a significant fraction of the experimental results.

This seems a difficult task because we know from sociologists and anthropologists' reports that human societies have myriads of norms, and it is not easy to discern which the key ones are. A prominent candidate, though, appear to be *norms of distributive justice* because concepts like fairness or justice are often employed to justify behavior.

The EM-norm, which views both efficiency *and* the welfare of the worst-off player(s) (and not, say, payoff equality) as the basic ingredients of distributive justice, is indeed an extremely rudimentary norm. In spite of this, we can explain a very good deal of the experimental evidence by assuming that principled people only care about it, as the results here and in López-Pérez (2004) attest.

In any case, more experiments are required to investigate what behavior people deem fair or just. For instance, it might be that nations or groups of people differ in what they view as fair. Thus, economists might be more concerned about efficiency than others - consult Fehr *et al.* (forthcoming) for evidence on this. Nevertheless, it must be pointed out that the model here is flexible enough to include such ideas. For instance, one could

[10] Taking into account some psychological evidence, I discuss the realism of this hypothesis (and its implications) in López-Pérez (2005). I also propose there alternative (and more complex) specifications.

introduce some heterogeneity by assuming that some principled people have internalized the EM-norm while others have internalized an egalitarian norm.

3. Applications

This section studies how the EM-norm affects cooperation, competition, and punishment in several games.

3.1 Cooperation

The Prisoner's Dilemma (PD) Lab Game

This *lab game*, represented at Figure 1, has received huge attention from experimentalists. The two players (John and Ana in the example) *simultaneously* decide whether they cooperate (action C) or defect (action D). Both earn c *monetary* units if they cooperate, and d if both defect. Further, a unilateral defector gets a 'temptation' payment of t while a unilateral cooperator gets a normalized payoff of zero. Payoffs satisfy $t > c > d > 0$ -i.e., defection strictly dominates cooperation in *monetary* terms- and $2c > t$ so that (c,c) is the only EM-allocation and cooperation is the only EM-action. In short, there exists a stark conflict between self-interest and compliance with the norm.

		John: C	John: D
Ana	C	c , c	0 , t
	D	t , 0	d , d

Figure 1: (Ana's, John's) Monetary Payoffs in the PD Lab Game

To illustrate players' utility payoffs, assume that Ana is selfish and John is principled (other cases can be analogously analyzed). Trivially, Ana's utility coincides with her own pecuniary payoff. On the other hand, John gets some disutility (shame) if he deviates *unilaterally* from the EM-norm or if Ana does so (anger), but he feels no disutility if both players defect. Figure 2 represents all this.

		John: C	John: D
Ana	C	c , c	0 , $t-\gamma$
	D	t , $-\alpha \cdot t$	d , d

Figure 2: Utility Payoffs if Ana is Selfish and John is Principled

Behavioral predictions are straightforward. First, mutual defection is the only Nash equilibrium if at least one player is selfish or if both players are principled and $\gamma < t - c$. Second, mutual cooperation is the unique *refined* equilibrium if both players are principled and $\gamma \geq t - c$ - although mutual defection is another Nash equilibrium, it can be ruled out because it is not an EM equilibrium (assumption 1).

To sum up, *principled players cooperate in a conditional manner in simultaneous dilemmas*: They cooperate only if the other player is expected to cooperate as well. Intuitively, this idea also extends to a setting where players' types are private information. More precisely, one can easily show that a principled player cooperates in the *simultaneous* PD if her prior is above threshold[11]

$$\rho^{sim} = \frac{d+\alpha\cdot t}{d+\alpha\cdot t-t+c+\gamma}. \qquad (3)$$

Consistent with the model, numerous experiments with one-shot prisoner's dilemmas –consult Rapoport and Chammah (1965), and Rabin (1993) for surveys; and Sally (1995) for a meta-analysis- find that a significant proportion of players cooperate, and that cooperation strongly depends on the expectation that the co-player will cooperate as well. Thus, in one of the treatments reported by Croson (2000), subjects played ten times a PD lab game against different co-players and had to guess at the start of each round her co-player's future choice. 83% of the participants that guessed their counterpart would cooperate cooperated themselves. On the contrary, when participants expected that their opponent would defect, only 32% of them cooperated.

To finish, inspection of threshold (3) indicates that ρ^{sim} depends negatively on c and positively on t and d. Interestingly, the same occurs with the expected price of cooperation $\rho\cdot(t-c)+(1-\rho)\cdot d$ –i.e., the *net*, expected material gain from defection. Taking into account that cooperation is hindered as the threshold ρ^{sim} grows, *a law of demand* follows: Cooperation decreases when its price increases. This prediction is again consistent with experimental evidence –see Rapoport and Chammah (1965, pp. 36-39), and Clark and Sefton (2001).

Fostering Cooperation: Sequential vs. Simultaneous Mechanisms

Assume now that the Prisoner's Dilemma is played in a sequential manner –e.g., Ana chooses after observing John's move. Apparently, this is a minor change. If the second player is principled, though, the sequential mechanism changes players' incentives to comply with the EM-norm, and fosters cooperation.

To understand this point, note first that the sequential PD has a unique EM-path. In it, both players cooperate one after the other, hence reaching the EM-allocation. As a result, the EM-norm commends the first mover to cooperate. Further, it also commends the second mover to cooperate if the first mover cooperates, but allows *any action* if the first mover defects (definition 3). Consequently, the first mover would be the *only* deviator from the EM-norm –i.e., the only person who 'misbehaves'- if both players chose defection. This is a subtle but key difference with the simultaneous PD, in which *both* players count as deviators if they mutually defect.

Given these norm prescriptions, the sequential PD has a unique Subgame Perfect Equilibrium for each parameter calibration. In this equilibrium (as one may easily prove), a

[11] Note that condition $1\geq\rho^{sim}$ requires $\gamma\geq t-c$.

selfish second mover always defects while a principled second mover reciprocates the first mover's choice if she is principled and $\gamma \geq t - c$ -that is, she cooperates if he cooperated and defects if he defected- whereas she always defects if $\gamma < t - c$.

Experimental evidence from Hayashi et al. (1999) and Clark and Sefton (2001) corroborates this. Second movers often cooperate conditional on the first mover's choice, while unconditional cooperation is negligible. In addition, Clark and Sefton (2001) show that reciprocation falls as its material cost rises, something that is also consistent with the model, as reciprocation only occurs if $\gamma \geq t - c$.

With regard to the first mover, it is fairly clear that his optimal strategy depends on his type and the second mover's. A selfish first mover cooperates only if the second mover is principled and $\gamma \geq t - c$ –this follows simply from $c > d$. In turn, a principled first mover cooperates if the other player is principled and $\gamma \geq \min\{\alpha \cdot t + d, t - c\}$, *or* if she is *selfish* and $\gamma \geq \alpha \cdot t + d$. This latter case is a bit paradoxical: The first mover cooperates even when he knows that his opponent will later defect! In that way, he avoids being the person who 'spoiled' cooperation, something that he finds particularly painful if $\gamma \geq \alpha \cdot t + d$.

The above mentioned results can be easily extended to an incomplete information setting. Since principled *second movers* reciprocate (if $\gamma \geq t - c$) and selfish ones always defect, a principled first mover cooperates in the *sequential* PD if $d - \gamma < (1 - \rho) \cdot (-\alpha \cdot t) + \rho \cdot c$, that is, if his prior is above threshold

$$\rho^{seq} = \frac{d + \alpha \cdot t - \gamma}{\gamma \cdot t + c}. \qquad (4)$$

Comparison between equations (3) and (4) indicates that $\rho^{sim} > \rho^{seq}$ if $\gamma \geq t - c$. That is, a principled mover in the simultaneous PD requires a larger prior to cooperate that a principled first mover in the sequential PD. This occurs because any deviation from the EM-norm (or from any conditional norm of cooperation) in the sequential PD is *unilateral*. As a result, a transgression is psychologically more disturbing (in expected terms) than in the simultaneous PD, in which it is possible that both players deviate *simultaneously*.

When analyzing the *simultaneous* PD, we also proved that selfish players never cooperate. On the contrary, they cooperate in the sequential PD if they move first and –this is again easy to prove- their prior belief is large enough. They find profitable to comply with the EM-norm because they understand that they can 'emotionally force' a principled second mover to comply as well.[12]

To sum up, the last two paragraphs imply that first movers' rate of cooperation in the sequential PD is significantly larger than the average cooperation rate in the simultaneous PD. This is consistent with the lab evidence from Hayashi et al. (1999) and Clark and Sefton (2001).

[12] See Rabin (1993, p. 1296) on this regard.

On Positive Reciprocity

In some models of reciprocity -Rabin (1993), Levine (1998), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006)- one may distinguish between positive reciprocity (being kind with those who are kind) and negative one (being unkind with those who are unkind). Positive reciprocity implies that people are more kind with an active *and kind* player than with a passive player who makes no choice in the game.

To illustrate this, consider again the *sequential* PD lab game but assume now that the first mover –that is, John- has only action C available –i.e., he is a passive player. The only active player is Ana, who must choose therefore between (Ana's, John's) allocations (c, c) and $(t,0)$. Clearly, the above mentioned reciprocity models predict that Ana will choose $(t,0)$ significantly *more* if John is passive (call this the passive *cooperation* case) than if John actually chose 'kind' action C (active *cooperation* case).

However, the available experimental evidence does not seem to support this prediction. Thus, Camerer (2003, pp.89-90) survey some results in this regard and concludes that the effect of positive reciprocity is insignificant or small. Consistently with such experimental evidence, my model predicts *invariance,* or *no positive reciprocity*. Indeed, Ana makes the same move in both cases *whatever* her type: She defects and attains allocation $(t,0)$ if she is selfish, and cooperates (the EM-action) if she is principled and $\gamma \geq t - c$.

The intuition behind the invariance result is twofold. On one hand, selfish types only care about available outcomes, and not about previous history, so that invariance makes no surprise. On the other hand, and in case everybody previously complied with the norm, it makes no difference for a principled player whether compliance happened because everybody was active and compliant or because everybody was passive –passive players, recall, respect the norm by definition. In other words, *principled players treat equally well both passive players and active compliant players*.

What explains Invariance?

The previous comparison has pointed out one key difference between my model and other models of reciprocity. However, models like Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and the model of quasi-maximin preferences of Charness and Rabin (2002) also predict invariance. This occurs because these models assume that players only care about the distribution of income –i.e., players have consequentialistic utility functions. Is it possible to discriminate between this explanation and the one this paper offers?

Although I will investigate this issue in more detail and give some evidence when studying punishment, it may be worth to consider again the sequential PD. In this case, however, consider Ana's behavior in the following two situations: (i) John is active and has chosen action D (active *defection* case), and (ii) John is passive and has only action D available (passive *defection* case).

Since Ana faces (Ana's, John's) allocations (d, d) and $(0,t)$ in both cases, a consequentialistic model predicts invariance –i.e., Ana always chooses the same allocation. My model, on the contrary, predicts some variance if Ana is principled. On one hand, she chooses (d, d) in the active defection case because then she feels angry at John. On the other hand, Ana does not feel any anger at a passive John and moreover the EM-allocation is $(0,t)$ if $(2+\delta)\cdot d < t$. Hence, she chooses $(0,t)$ if t and γ are large enough.

To sum up, *while a principled player may treat kindly a passive player, she will never do that with a deviator*. This idea is absent in a consequentialistic model, but it is important to appreciate why institutions making tough decisions that affect others have incentives to signal that they had no other choice or were forced by external forces to do that. In such a way, other agents cannot blame institutions for violating prevailing norms and hence do not get angry at them. For instance, many European governments and politicians who advocate for reforms in their Welfare States often argue that Globalization leaves them no way out. Though some of them may sincerely believe that, such type of arguments might be also part of a strategy designed to prevent voters' indignation.

Efficiency and maximin versus equality

In the passive *defection* case of the previous example, Ana must choose between (Ana's, John's) allocations (d, d) and $(0,t)$. The former allocation is completely egalitarian while the second one is not. As the EM-norm commends to care about efficiency and maximin, and not about equality, it follows that a principled Ana chooses the latter allocation. Hence, this example shows the importance of the assumption that principled people care about the EM-norm and not, say, about an egalitarian norm. In general, this hypothesis is particularly well supported by the evidence coming from individual decision lab problems with externalities.[13] As I summarized much of this evidence in López-Pérez (2004), I will only provide here two implications of this hypothesis.

First, people are willing to spend money for the sake of efficiency and maximin. Consider a situation in which agent B has no say whereas player A must choose between (A, B) pecuniary allocations (4, 4) and (4-ε, 10). If ε and δ are small enough (more precisely, $\varepsilon\cdot(1+\delta)<6$), the only EM-allocation is (4-ε, 10) and hence the EM-norm commends to choose it. My model predicts that behavior if player A is principled and if her internalization parameter γ is larger than ε -incidentally, she would clearly opt for (4, 4) if she were selfish. In contrast, she would unequivocally choose (4, 4) if she had internalized an egalitarian norm like that of function (**¡Error! No se encuentra el origen de la referencia.**.

[13] The reader may consult Frohlich and Oppenheimer (1992), Charness and Rabin (2002), Konow (2003), and Engelmann and Strobel (2004). See also Fehr *et al.* (forthcoming) for some evidence to the contrary.

Second, people are not willing to spend money just to promote equality. To see this, suppose now that A must select either (3, 3) or (4, 6). According to my model, a principled or a selfish A always opts for the latter allocation. In contrast, she would choose (3,3) if she had internalized an egalitarian norm and her γ was large enough.

Lab Games with n Players: Public Goods

In a simple Voluntary Contribution Mechanism (VCM) public good *lab game*, $n \geq 2$ subjects, each one with an endowment of e monetary units, choose *simultaneously* whether to contribute e to a public good or to keep the endowment for them.[14] Subject i's monetary payoff at terminal node z is given by $m \cdot e \cdot c(z)$ if she contributes and by $e + m \cdot e \cdot c(z)$ if she does not contribute, where m denotes the monetary payoff per unit of public good and is such that $m < 1 < n \cdot m$, and $c(z)$ stands for the number of players that contributed to the public good in the history of z. Since $m < 1$, the dominant strategy in monetary terms is not to contribute. Nevertheless, many experiments report aggregate contribution levels around 40-60% -for a survey, consult Ledyard (1995).

To get behavioral predictions, note first that the EM-norm commends every player to contribute because $1 < n \cdot m$. Let then n_p ($0 \leq n_p \leq n$) denote the number of principled players in the group (recall that players' types are assumed to be common knowledge). For any n_p and γ, α, the VCM lab game has a unique *refined* equilibrium:

- If $\gamma < \alpha \cdot m \cdot e$, no player contributes.
- If $\gamma \geq \alpha \cdot m \cdot e$, no selfish player contributes while a principled player contributes only if $n_p = n$ or if $n_p < n$ and

$$e \cdot m \cdot n_p - \alpha \cdot e \cdot (1 + m \cdot n_p) \geq e + (m \cdot e - \gamma) \cdot (n_p - 1) \Leftrightarrow n_p \geq \frac{e \cdot (1 - m + \alpha) + \gamma}{\gamma - \alpha \cdot m \cdot e} = n_p^*(m, \alpha, \gamma). \quad (5)$$

In other words: Principled players respect the EM-norm if sufficiently many others do it as well. Note that there exist other equilibria if $n_p \geq n_p^*$, but they are not EM because at least one principled player does not contribute in them (assumption 1).

Observe also that n_p^*, the minimal number of principled agents necessary to sustain positive contributions (the *critical mass*), does not depend on the total number of players n. Consequently, the probability that a group of n agents independently drawn from the population contains n_p^* or more principled players grows with n, so that

[14] In more complex VCM games, players are allowed to contribute a fraction of the endowment, and not only the whole one. This is unsubstantial for my model –I come back to this in the conclusion. Note also that results do not change substantially if players have heterogeneous endowments e_i, although I assume it for expositional simplicity.

cooperation should get facilitated as n increases, a result supported by experimental evidence from Isaac and Walker (1994).

In case player's types are private knowledge, it is fairly easy to show that principled types contribute if ρ is large enough. In other words: There exist a positive correlation between the expectations of a principled agent about aggregate contribution levels and her decision to contribute. Abundant experimental evidence bears this point –consult Orbell and Dawes (1991), and Sonnemans et al. (1999).[15]

In addition, experimental evidence from Isaac and Walker (1988) –see also Ledyard (1995) for a survey- shows that contribution levels raise if m increases. In this regard, inspection of equation (5) points out that n_p^* depends negatively on m only if γ and α are large and small enough, respectively, so that an increase in m will foster contributions only in those cases. The intuition is that, since contributing has the side-effect of increasing *deviators*' earnings, the emotional cost of anger must be offset by the emotional cost of transgressing the norm in order to find contribution optimal.

3.2 Competition: Market Lab Games

Experimental evidence from a broad class of market lab games supports the standard prediction that prices *converge* to the competitive equilibrium –see, for instance, the survey in Fehr and Schmidt (1999, p. 829). Is the model here consistent with that? To study this point, consider a market game with *proposer* competition: $n-1$ sellers (proposers) make simultaneous price offers p_1, p_2, ..., and p_{n-1} to sell one unit of a good to a single buyer (responder) who demands only one unit of the good. The buyer can accept the offer *she prefers* or reject all of them.

Assume that the responder values one unit of the good in V monetary units. Hence, the responder's monetary payoff if she accepts price offer p_i ($i \in \{1,2,..,n-1\}$) is $V - p_i$, whereas seller i's income is p_i -unsuccessful sellers get zero money. Finally, all players get no money if the responder accepts no offer.

Before applying the model to this game, it is convenient to consider first the prediction when all players are selfish. For any $n \geq 3$, the game has then a basically unique SPE: The responder always accepts the minimum price offer and at least two

[15] In experiments with finitely repeated public goods games, aggregate contributions fall over time, getting very close to the zero level. I will not address this point in detail here, but the model suggests that such phenomenon might be due to learning about the number of principled players. According to this, (some) principled subjects might arrive at the lab with upwardly biased priors that they revise when they observe actual contribution levels. This revision downwards might explain the decrease in contributions. Of course, we should abandon the assumption that priors are common in order to model such process.

proposers offer a price equal to zero.[16] The intuition why this equilibrium is unique is similar to that behind the Bertrand Duopoly equilibrium, and the reader is directed to a Microeconomics textbook for a proof. Finally, note that the standard equilibrium result is radically different if $n=2$ because then the proposer reaps the whole surplus V -this is the so-called *ultimatum game*; I briefly study it in section 3.3.

Consider now the prediction of my model if $n\geq 3$. The key point here is that *all* allocations in this game are EM -except those that are attained when the responder rejects. In effect, all these allocations are efficient and moreover the worst pecuniary payoff is zero in all of them –if $n\geq 3$ there is always at least one unsuccessful seller who gets nothing. Therefore, offering any price and accepting it are EM-actions, whereas rejecting it is not. This implies in turn that the utility payoffs of any type of player coincide with monetary ones unless the responder rejects –in that case, she suffers a utility cost if she is principled whereas principled sellers anger at her. It is then easy to show that the game has a (basically) unique SPE that coincides with the standard one. Clearly, this result does not depend on players' types being common knowledge.

Consider now a market lab game with *responder* competition. Opposite to the game with proposer competition, this game has just one seller (proposer) and $n-1$ buyers (responders). The proposer moves first by choosing a selling price p and then each responder decides, unaware of other responders' choices, whether she accepts or rejects p. All players receive a monetary payoff of zero if *all* responders reject p. In turn, the proposer gets p and the buyer $V-p$ if at least one responder accepts - a random draw selects with equal probability one of the accepting responders in case more than one accepts-, and all other responders receive zero.

Note first that there exists a unique SPE if all players are selfish: Responders accept any selling price while the proposer makes a price offer of $p=V$, thus reaping the whole surplus. In fact, one may prove that the game has this unique SPE whatever the players' types if $n\geq 3$ -the game is the ultimatum game if $n=2$; subsection 3.3 studies it. The reasons are now familiar: All *Pareto-efficient* allocations in this game are EM-ones so that accepting any price offer is consistent with the EM-norm. Further, as rejection is never *pecuniary* profitable for principled or selfish responders, it follows that responders always accept in equilibrium, and a seller consequently asks for the whole surplus. Experimental evidence roughly supports this equilibrium prediction –see Fehr and Schmidt (1999, p. 832) for references.

3.3 Punishment

In a two-player game, player A punishes B when she imposes a cost on B without getting any immediate material reward as a result. According to the model, A punishes B

[16] Many strategy profiles satisfy this, but they only differ in the distribution of offers of the remaining $n-3$ sellers, which is inconsequential for the final result. Hence, the equilibrium outcome is unique.

only if B has transgressed a norm that A cares about and which A herself has not violated. Intuitively, B's deviation triggers an aggressive emotion in A that goes associated with an impulse to retaliate.

As an illustration, consider the decision tree at Figure 3, where only *monetary* payoffs are depicted. The first mover can offer either (player 1's, player 2's) allocation (8, 2) or (5, 5), and then the second mover can accept (A) or reject (R) the offer. Both players get zero money if she rejects. Otherwise, the offer is implemented. This lab game is a simplified version of an Ultimatum Game with stakes equal to 10 monetary units –the difference is that the range of offers in the ultimatum game consists of *all* possible divisions of the stakes. I stick to this simple version because it is sufficient to show how punishment works -for a detailed analysis of the model's predictions in the Ultimatum lab game, consult López-Pérez 2004.

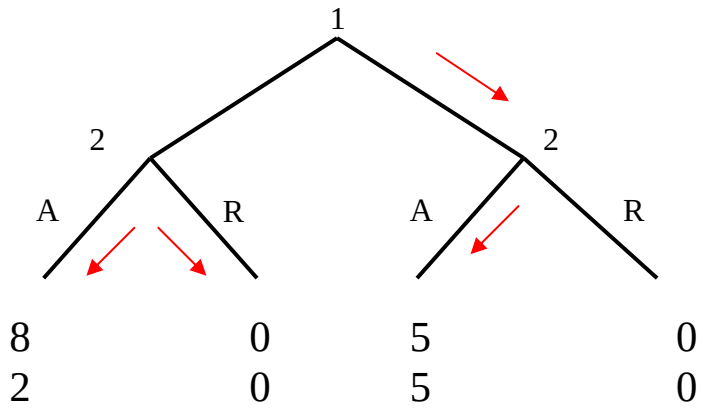

Figure 3: A Mini-Ultimatum Lab Game

As (5, 5) is the unique EM-allocation of this game, the EM-norm clearly commends player 1 to offer (5, 5) and player 2 to accept it. On the other hand, if player 1 deviates from the norm and offers (8, 2), the EM-norm allows player 2 to choose *any* move. Arrows in Figure 3 indicate that the associated action is selected by the EM-norm.

The game has a unique *refined* SPE. In it, a selfish second mover accepts any offer. Further, a principled second mover accepts offer (5, 5), rejects (8, 2) if $0 > 2 - 8 \cdot \alpha$ and accepts it if $0 < 2 - 8 \cdot \alpha$. In the marginal case $\alpha = 0.25$, a principled second mover is indifferent between accepting or rejecting (8, 2) so that there are two SPE. However, only the one in which the second mover accepts is EM (assumption 1).

In turn, the first mover's offer depends on both players' types, as Table 1 indicates. The first column in this matrix shows player 1's type, while the first row shows player 2's type. For instance, player 1 abides by the EM-norm and offers (5, 5) independently of the co-player's type if she is principled and $\gamma \geq 3$.[17]

[17] There are two SPE if $\gamma = 3$ and the second mover accepts (8, 2). However, the equilibrium in which player 1 offers (8, 2) is not EM and can thus be ruled out (assumption 1).

Player 2's type is… / Player 1's type is…	… selfish or principled with $\alpha < 0.25$.	… principled With $\alpha > 0.25$.
… selfish or principled with $\gamma < 3$.	(8, 2)	(5, 5)
… principled and $\gamma \geq 3$.	(5, 5)	(5, 5)

Table 1: Player 1's SPE offer depending on her type and the second mover's.

Note that this result can be easily extended to an incomplete information setting. Obviously, a selfish first mover or a principled one with $\gamma < 3$ should condition their choice on their prior about the second mover's type -the reader may easily compute the minimal prior that makes offer (8, 2) optimal.

Experimental data from ultimatum games –see Camerer (2003, pp. 48-55) for an informative survey- confirms that the 50-50 offer is almost always accepted, whereas low offers face a high probability of rejection. Studies also show that "very large changes in stakes have only a modest effect on rejections",[18] something that is barely consistent with my model –if a principled second mover rejects (8, 2) in the lab game of Figure 3 then she also rejects offer (8·k, 2·k) when stakes are k>0 times bigger.

The analysis shows that punishment depends on parameter α -which, incidentally, could be estimated from experimental data. In fact, if one extended the model by assuming that principled players are heterogeneous regarding their aggressiveness –i.e., parameter α -, *a law of demand* would follow: The more costly punishment is the less of it there is. To see this, consider a slightly modified version of the lab game at Figure 3 in which allocation (6, 4) replaces allocation (8, 2). Since (5, 5) is still the only EM-allocation, a principled second mover will anger if she is offered (6, 4). Nevertheless, punishing (i.e., rejecting) such offer is more costly than rejecting offer (8, 2) and hence only optimal if α is relatively large –more precisely, if $\alpha > 2/3$ holds. To sum up, principled agents use relatively costly punishment technologies only if they are aggressive enough.

Another interesting issue concerns *responsibility* (or 'intentions', to use the usual terminology). As an illustration, assume that player 1 has no say in the lab game of Figure 3 and that his offer is decided by a random device. As player 1 cannot be blamed for anything that happens in the game, a principled player 2 will not anger at him and hence will not reject any offer. Therefore, and in comparison with the intentional treatment, the model predicts a *smaller* rate of rejection in the random treatment, something that is consistent with the experimental results reported by Blount (1995). To sum up, the model

[18] Camerer (2003, pp. 61).

indicates that *responsibility* is crucial to understand *who is punished* because it predicts that only wrongdoers get punished.

The word 'intentions' also refers sometimes to the *influence of non-chosen alternatives*. To illustrate this point, consider a slight variation of the lab game of Figure 3, in which allocation (10, 0) replaces allocation (5, 5), and compare the rejection rate of offer (8, 2) in this new game and in the former game. Does the model predict a difference? Yes. As offer (8, 2) constitutes a deviation from the EM-norm when the alternative is (5, 5), but not when the alternative is (10, 0), the model clearly predicts a larger rejection rate in the former case –in fact, it predicts that nobody rejects (8, 2) if the alternative is (10, 0). More generally, whether an action constitutes a norm transgression depends on the available alternatives, and that explains why an act with the same material consequences may be punished in one game but not in another. This prediction is highly consistent with the experimental evidence –see Camerer (2003, p. 81-82).

4. Comparison with Other Utility Models

It can be illustrative to compare the behavioral predictions of the model with those from other models.[19] With regard first to cooperation and punishment, the model has been shown to be consistent with seven well-replicated experimental phenomena:

(1) A significant number of people cooperate in the *simultaneous* PD lab game, or contribute in a one-shot public good lab game.

(2) Subjects also contribute in the *sequential* PD, and the rate at which first movers cooperate is larger than average cooperation in the simultaneous PD.

(3) Subjects often give money to passive players (dummies).

(4) Subjects tend to treat equally kindly both dummies and *kind* active players (absence of positive reciprocity).

(5) Many subjects sacrifice equality of payments in order to increase efficiency and/or the worst-off player's payoff.

(6) Punishment depends on the menu of alternatives that the *punished* person had available.

(7) Dummies are rarely punished (responsibility).

Table 2 indicates whether other utility theories are consistent with facts (1) to (7). Entry YES indicates that the corresponding theory is consistent with the fact, whereas entry NO indicates the opposite. For brevity, I consider just four models, each one representing a different research line in the existing literature. Models of inequity aversion like Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) represent pure consequentialistic models in which people only have distributional concerns –other examples include the model of quasi-maximin preferences of Charness and Rabin (2002). Rabin (1993) is a pure reciprocity model with no distributional concerns, as Dufwenberg

[19] Consult López-Pérez, R. (2004) for a more lengthy discussion.

and Kirchsteiger (2004). Falk and Fischbacher (2006) introduce both reciprocal and distributional concerns. Finally, Levine (1998) is a model of type-based reciprocity.

Facts / Theories	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rabin (1993)	YES	NO	NO	NO	NO	YES	YES
Levine (1998)	YES	YES	YES	NO	YES	YES	NO
Inequity aversion	YES	YES	YES	YES	NO	NO	NO
F&F (2006)	NO	YES	YES	NO	NO	YES	NO

Table 2: Predictions by Other Utility Models

The interested reader is directed to the relevant papers for a detailed explanation of these predictions. Finally, all models but Rabin (1993) can explain the experimental results from market lab games, although some remarks could be made. To mention one, models of inequity aversion predict the standard solution in the market game with proposer competition if the responder is restricted to accept or reject the *highest price* offer but not if she is given the opportunity to choose *any* offer, in which case a very inequity-averse responder would rather accept an egalitarian sharing of the surplus. Note that this distinction is immaterial in my model.

5. Conclusion and Extensions

This paper shows that a large set of experimental evidence can be explained if one assumes that (some) people care about a particular norm of distributive justice. The model appears to be empirically more accurate than other models of non-selfish preferences. Moreover, it is much simpler and precise than other models of reciprocity.

There are some possible ways to extend the model. A natural one is considering other norms that the EM-norm. For instance, one could assume that some people have internalized a norm of honesty, and study how it affects communication. One could also think of more realistic norms of distributive justice. The EM-norm is too strict in that it allows any behavior after a deviation occurs. Less draconian norms would take into account the welfare of those players who *have hitherto respected* the norm –López-Pérez (2005) gives particular examples. Further, the EM-norm is probably too austere in that it only allows EM-actions. However, people seem to have a more flexible view of what is correct: 'Small' deviations from the ideal moral behavior –e.g. the EM-path in the model- are usually considered valid as well, and they do not trigger anger.

Some of the motivational hypothesis of the model could be also relaxed. First, the model assumes that the intensity of remorse does not depend on the specific deviation one makes from an internalized norm. But it seems realistic to assume that remorse is higher depending on the material consequences of the deviation[20] –e.g., killing someone should generate stronger remorse than just hurting him. This hypothesis and an additional one of

[20] I have investigated this point in López-Pérez (2005).

decreasing marginal utility of money could explain, for instance, why participants in public good games often contribute something between zero and their endowment. Second, but closely related, it might be more realistic to assume that the intensity of anger at a deviator depends on the specific misbehavior and its consequences.

As a final remark, the model here should motivate further experimental research on social norms, emotions, and reciprocity. Further, it might be used to study how norms based on political ideologies (or religious beliefs) and aggressive emotions like anger shape political violence, terrorism, and revolutions; or how a sense of duty and the associated emotions of guilt and shame affect voters' turnout, to give some examples.

Bibliography

- Arrow, Kenneth J. (1974). *The Limits of Organization*. Norton & Company.
- Becker, G. (1996). *Accounting for Tastes*, Harvard University Press.
- Blount, S. (1995). "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences", Organizational Behavior & Human Decision Processes, 63(2), 131–144.
- Bolton, G. E., and A. Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition", American Economic Review, 90(1), pp. 166-93.
- Camerer, C. (2003). *Behavioral Game Theory-Experiments in Strategic Interaction*, Princeton University Press.
- Charness, G., and M. Rabin (2002). "Understanding Social Preferences with Simple Tests", Quarterly Journal of Economics, 117, 817-869.
- Clark, K., S. Kay, and M. Sefton (2001). "When are Nash Equilibria Self-Enforcing? An Experimental Analysis", International Journal of Game Theory 29, 495-515.
- Cox, J., D. Friedman, and S. Gjerstad (2006). "A Tractable Model of Reciprocity and Fairness", mimeo, University of Arizona.
- Croson, R. T. A. (2000). "Thinking like a Game Theorist: Factors affecting the Frequency of Equilibrium Play." Journal of Economic Behavior and Organization, 41, 299-314.
- Dufwenberg, M., and G. Kirchsteiger (2004). "A Theory of Sequential Reciprocity", Games and Economic Behavior, 47, 268-98.
- Elster, J. (1989). "Social Norms and Economic Theory", Journal of Economic Perspectives, 3(4), 99-117.
- Engelmann, D., and M. Strobel (2004). "Inequality Aversion, Efficiency and Maximin Preferences in Simple Distribution Experiments", American Economic Review, 94(4), 857-869.
- Falk, A., and U. Fischbacher (2006). "A Theory of Reciprocity", Games and Economic Behavior 54, 293-315.

- Fehr, E. and K. Schmidt (1999). “A Theory of Fairness, Competition and Cooperation”, Quarterly Journal of Economics, 114(3), 817-68.
- Fehr, E., and K. Schmidt (2006). “The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories”, in S. C. Kolm, and J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, Volume 1, Elsevier B. V.
- Fehr, E., M. Näf, and K. Schmidt. “The Role of Equality, Efficiency, and Rawlsian Motives in Social Preferences: A Reply to Engelmann and Strobel.” Forthcoming in American Economic Review.
- Frohlich, N., and J. A. Oppenheimer (1992). *Choosing Justice: An Experimental Approach to Ethical Theory*. Berkeley and LA: University of California Press.
- Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). “Psychological Games and Sequential Rationality.” Games and Economic Behavior 1, 60-79.
- Gintis, H. (2003). “The Hitchhiker’s Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms.” Journal of Theoretical Biology 220(4), 407-418.
- Hayashi, N., E. Ostrom, J. Walker, and T. Yamagishi (1999). “Reciprocity, Trust and the Sense of Control: A Cross-Societal Study.” Rationality and Society, 11, 27-46.
- Isaac, R. M., and J. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism." Quarterly Journal of Economics, 103, pp. 179-99.
- Isaac, R. M., J. Walker, and A. Williams (1994). "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing very Large Groups." Journal of Public Economics, 54, pp. 1-36.
- Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). "Fairness and the Assumptions of Economics." Journal of Business, 59 (4, 2), 285-300.
- Konow, J. (2003). “Which Is the Fairest One of All? A Positive Analysis of Justice Theories.” *Journal of Economic Literature*, 41 (4), pp. 1186-1237.
- Ledyard, J. (1995). “Public Goods: A Survey of Experimental Research”, in J. Kagel and A. E. Roth (Eds.), *Handbook of Experimental Economics*, Princeton Univ. Press.
- Levine, D. K. (1998). “Modeling Altruism and Spitefulness in Experiments.” Review of Economic Dynamics, 1, 593-622.
- López-Pérez, R. (2004). “Emotions Enforce Fairness Norms”, mimeo.
- López-Pérez, R. (2005). “Guilt and Shame in Games”, mimeo.
- Orbell, J., and R. Dawes (1991). "A Cognitive Miser' Theory of Cooperators' Advantage." American Political Science Review, 85, 515-28.
- Parsons, T. (1967). *Sociological Theory and Modern Society*, New York: Free Press.
- Rabin, M. (1993). “Incorporating Fairness into Game Theory and Economics”, American Economic Review 83, 1281-1302.
- Rapoport, A. and A. M. Chammah (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, MI: University of Michigan Press.

- Schelling, T. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.
- Sonnemans, J., A. Schram, and T. Offerman, 1999. "Strategic Behavior in Public Good Games: When Partners Drift Apart." Economics Letters 62, 35-41.
- Sugden, R. (1989). “Spontaneous Order”, Journal of Economic Perspectives, 3(4), 85-97.