IESE Business School
University of Navarra

SP-SP

Working Paper
WP no 645
September, 2006

PEER PRESSURE AND INEQUITY AVERSION IN THE JAPANESE FIRM

Gianandrea Staffiero

IESE Business School – University of Navarra
Avda. Pearson, 21 – 08034 Barcelona, Spain. Tel.: (+34) 93 253 42 00 Fax: (+34) 93 253 43 43
Camino del Cerro del Águila, 3 (Ctra. de Castilla, km 5,180) – 28023 Madrid, Spain. Tel.: (+34) 91 357 08 09 Fax: (+34) 91 357 29 13

The Public-Private Center is a Research Center based at IESE Business School. Its mission is to develop research that analyses the relationships between the private and public sectors primarily the following areas: regulation and competition, innovation, regional economy and industrial politics and health economics.

Research results are disseminated through publications, conferences and colloquia. These activities are aimed to foster cooperation between the private sector and public administrations, as well as the exchange of ideas and initiatives.

The sponsors of the SP-SP Center are the following:

- Accenture
- Ajuntament de Barcelona
- Official Chamber of Commerce, Industry and Navigation of Barcelona
- BBVA
- Diputació de Barcelona
- Garrigues, Abogados y Asesores Tributarios
- Catalan Government (Generalitat de Catalunya)
- Sanofi-Aventis
- Telefonica
- T-Systems
- VidaCaixa

The content of this publication reflects the conclusions and findings of the individual authors, and not the opinions of the Center's sponsors.

PEER PRESSURE AND INEQUITY AVERSION IN THE JAPANESE FIRM

Gianandrea Staffiero*

Abstract

The herein study presents an explanation of the high frequency of team production and high level of peer monitoring found in Japanese firms, in terms of a simple and empirically grounded variation in individual utility functions. We argue that Japanese agents are generally characterized by a higher degree, with respect to their Western counterparts, of aversion to unfavourable inequality, a feature which explains seemingly puzzling experimental evidence. In combination with long term employment and various organizational practices, this creates the conditions for obtaining willingness to exert mutual monitoring and peer pressure which facilitates the convergence towards cooperative equilibrium in dilemma type situations.

*Post-Doctoral Research Fellow, IESE

JEL Classification: C7, C91, C92, D63, H41, L23

Keywords: Team production, fairness, cooperation, punishment, reciprocity.

PEER PRESSURE AND INEQUITY AVERSION IN THE JAPANESE FIRM*

1. Introduction

The free-riding problem in team production is one of the most well-known issues within the literature in economics and in social science in general. Since the classical work by Alchian and Demsetz (1972), it has become one of the typical examples of "market failure" connected to the "public good nature" of, in this case, the total output produced by a group, whose members receive a share which does not depend solely on their personal effort. As the cost of effort is born entirely by the individual, while the benefits are reaped by the whole group, we find that individual payoff maximization brings about an inefficiently low level of team production.

This result rests on the standard assumption of a rather narrow definition of individual prederences, based only on a function of one's own result in term of a combination of compensation and effort. However, everyday evidence suggests the importance of social aspects related to working in a group. In particular, members who do provide a substantial effort typically feel resentful towards free-riders. On the other hand, the latter may feel guilty towards high contributors. These aspects are related to the concepts of shame and guilt, proposed as the two components of peer pressure proposed by Kandel and Lazear (1992). They show how peer pressure can overcome the free-riding problem in a context of profit sharing among partners in a small group. Their analysis is extended in Barron and Paulson Gjerde (1997), where a principal takes into account mutual monitoring and pressure among workers when designing incentive schemes.

There is nowadays widespread experimental evidence that social aspects affect people's decisions in such a way that, in many contexts, own payoff maximization and perfect rationality perform quite poorly as sources of predictions for human behavior. That is the case even quite simple games, in which we can safely assume that the lack of perfect rationality is not the main issue. Among recent works, Fehr and Gaechter (2000), Andreoni et al. (2002) and Sefton et al. (2002) provide evidence of willingness to spend in order to punish free-riders even

* A previous version of this paper was part of my Ph.D. thesis at the I.D.E.A. program at the Universitat Autònoma de Barcelona. I would like to thank my supervisor, Jordi Brandts, for his continuous support, and Reinhard Selten, Carlo Filippini, Joan Maria Esteban, Pierluigi Sacco, Inés Macho-Stadler, Steffen Huck and Rosella Nicolini for useful suggestions. The usual disclaimer applies. Financial support from the Generalitat de Catalunya duriong my Ph.D. studies (Beca FI 2000 487) and from the SP-SP Research Center at IESE Business School-University of Navarra is gratefully acknowledged.

in the last stage of a repeated game and in "stranger" conditions, where two opponents are not going to interact more than once and, therefore, strategic considerations are not behind this behavior. It is noteworthy how free-riders learn to cooperate, when given the chance to play the same game, even though with a different match. So we find that willingness to punish, one of the conditions for the effectiveness of the external part of peer pressure ("shame"), is often fulfilled.

As for the internal pressure, guilt, we find that a number of agents adapt their contribution choices to their peers'. This is defined as the "conditional cooperative" behavior in Fischbacher et al. (2001) and Keser and Van Winden (2000). However, while it is clear that agents who cooperate when uninformed about others' behavior quickly withdraw their contributions in case of free-riding by the other members, less conclusive is the evidence about free-riders' desire to increase theirs in order to make it reach the group average (see, e.g., Brandts and Fatás, 2001).

Overall experimental evidence on dilemma and public good games (see Ledyard, 1995) spurred new theories of human preferences which try to explain findings apparently at odds with "conventional" assumptions. Among the "social" or "other regarding" preferences, the ones which include inequity and inequality aversion proposed by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) are compatible both with evidence which suggests equity or reciprocity as leading forces of behavior and with data which fits with "standard" kinds of preferences. In fact, in some settings even the inequality averse behaves in the same way a selfish would; more important for our purposes, in others it is the selfish who finds it worthwhile to behave "nicely". Especially in Fehr and Schmidt's analysis we find examples of this kind of outcomes as a result of interaction between the selfish and the fairness minded, for given (and experimentally tested) proportions in the population.

A natural question arising in this respect is how different cultural contexts can bring about different outcomes in various aspects of social life. For instance some of our results are derived from modelling the cultural context as a provider of initial conditions in an evolutionary setting. The initial conditions are thought precisely as proportions of fairness minded agents in the population. We choose inequity aversion as the representation of "fairness orientation" for the strong experimental support it received and for the simplicity it provides in deriving predictions.[1]

Our analysis investigates the reasons for the emergence of peer pressure underlined by several scholars as a relevant and distinctive aspect of life within Japanese firms when compared to their "Western" counterpart. Common wisdom refers to the communitarian type of culture the Japanese society is embedded in as the driving force behind team cooperation and mutual monitoring as means to achieve it, while the more individualistic orientation typical of Western societies drives towards a system of incentives and control based on hierarchy and individual performance evaluation. We are going to show how these aspects can be studied and explained applying standard concepts of Nash equilibrium and subgame perfection (Nash, 1950; Selten, 1975) based on simple and empirically grounded variations in the individual utility functions, while traditionally they are studied in more "holistic" terms.

---

[1] Other models propose to incorporate the reward and punishment of good and bad intentions, following Rabin (1993) seminal contribution. The main problem of these models (e.g. Dufwenberg and Kirchsteiger, 2004) lies in their high complexity. We argue that the inequality aversion models are preferable as the loss of accuracy as predictors is very limited with respect to the advantage of their simplicity when compared to their intentions based counterparts.

More specifically, our results are based on critical parameters included in the utility functions, representing the aversion to inequity as in Fehr and Schmidt (1999). We are going to use assumptions on the different frequency, in Japan and in the "West", of certain critical levels of these parameters. These assumptions are based not just on what seems intuitive reasonable in the light of "group orientation" culture frequently mentioned to be stronger in Japan (we later discuss this point), but especially on empirical evidence arising from cross-cultural experiments. We find these experiments particularly useful for our purpose, as they permit to isolate aspects of behavior that are normally hard to disentangle with the multitude of elements which can potentially affect prominent features of field data. For example, evidence of high productivity in teams could be due to intrinsic desire to contribute, or to altruism, or fear to be punished by peers or by bosses; also, different situations in the specific workplace of a given country could play an important role. In the experimental lab it is possible isolate the effect of a punishment device available to group members, by comparing results between treatments where the presence or absence of such device is the only difference. Moreover, cross cultural experiments allow to make sets of individuals, whose main difference is the country where they live (university students are the usual subject pool), interact in similar environment under the same rules. In the next section we discuss about how the Japanese "orientation" towards group members needs to be carefully defined if one wants to be consistent with experimental evidence, and propose aversion to "being behind" as a plausible explanation. In section 3 we introduce the game we use in order to represent the interaction among team members. Section 4 presents the results derived when players interact in couples and are averse only to unfavorable inequality. Section 5 deals with a more general case, where group size varies and aversion to favorable inequality also exists. In section 6 we discuss how results allow to reconcile experimental evidence with existing organizational practices in the Japanese firms. Section 7 concludes.

2. Japanese vs "West": making sense of experimental evidence

The research by Yamagishi and colleagues (see Yamagishi et al., 1998 and references therein) points out that the idea of the Japanese as people sharing a genuine tendency to enjoy grouping together and cooperate is misleading. In Yamagishi (1988b), in particular, we find that Japanese subjects cooperate substantially less than their American counterparts in a standard dilemma games, i.e. in a game where individual payoff maximation leads to defection and to payoff allocations that are Pareto dominated by the ones which would be achieved if everybody cooperated. These difference vanishes once a sanctioning system is introduced: cooperation increases only slightly among U.S. subjects, substantially in Japan. In Yamagishi (1988a) it is also observed that Japanese subjects are more prone to leave a group when they find other members to be less inclined or less able to cooperate, where in this case cooperation consists in completing a test correctly, as rewards are to be shared among team members. This evidence undermines the common perception of a higher group orientation in Japan, which seems consistent with the frequently observed feature of organizing production in teams. Yamagishi proposes an interesting explanation: it is exactly because of the strength of the ties that Japanese individuals develop in their life, in the family and in their firm, that they are not accustomed to trust others when the ties are weak. In particular, long term employment favors this tendency and the related low ability to form trustful relationships with strangers. The link between these "Japanese" traits and the experimental evidence lies precisely in the sanctioning possibilities, which are more readily available in "strong ties" relationships. On the other hand,

Yamagishi argues that this same sanctioning systems are responsible for undermining a voluntary basis for cooperation.

Explaining the frequency of cooperation in working groups as motivated by individual utility functions in which the argument to maximize is group welfare appears to be inconsistent with the experimental evidence just mentioned. A more plausible explanation could be suggested by the frequency of punishment as leading to cooperation. In fact, this would also be consistent with the cultural origins of the large use of team production and related collective norms inside the Japanese firm, pointed out by Aoki (1988 and 2001). Aoki traces these origins in the egalitarian conventions fostering cooperation within and among villages - and the willingness to engage in bloody clashes in case of violations - for the control of water cultivation necessary for the cultivation of rice.

In fact, a more recent experiment addresses this aspect more directly. Cason et al. (2004) propose a non-linear public game made of two stage. In the first, each of the two players have to decide whether to commit to zero cooperation or not. In the second, the player(s) who did not commit has to decide his cooperation level. The non-linear payoff function is such that own-payoff maximization implies that if both players reach the second stage, they both pick 8 as contribution level in the Nash equilibrium. If only one player reaches that stage, he should pick a higher level, namely 11. Notice that payoffs are equal in the former case, and largely unequal in the latter, the zero contributor being obviously better off. The two pure strategy Nash equilibria involve one player's commitment to zero contribution, the other picking 11 in the second stage; in the mixed strategy equilibrium they both decide to commit to zero contribution with 0.32 probability. American data are consistent with the pattern of this last equilibrium, both in terms of the rate of participation to public good production (68%) and in second stage behavior: a player "should" contribute more when left as the only producer of a public good than when engaged in joint production. However, data show that when a Japanese player is left alone in the second stage, he typically decides to produce at a substantially lower level than the one maximizing his own payoff. As a consequence, the average payoff of players who commit to zero contribution is in fact lower than the average payoffs of players once they jointly reach the second stage. As the game is played repeatedly, this pattern causes the frequency of commitment to zero contribution to be lower in Japan, and this in turn makes average production of "public good" closer to the efficient levels and average payoffs higher than in the U.S.

This result is consistent with the idea that Japanese people are relatively more willing to sacrifice their own interest to punish people who contribute less and to make payoff distribution more egalitarian.

Among the model of social preferences, the ones which focus on the latter aspect have the advantage to be able to organize a large deal of experimental evidence in a very parsimonious way, in terms of parameter involved. In particular, the model in Fehr and Schmidt (1999) is based on the following utility function:

$$u_i(\pi_i, \pi_{-i}) = \pi_i - \frac{1}{N-1}\alpha_i \sum_{j \neq i}^{N} \max\{\pi_j - \pi_i, 0\} - \frac{1}{N-1}\beta_i \sum_{j \neq i}^{N} \max\{\pi_i - \pi_j, 0\} \quad (1)$$

As π denote payoffs, α measures the aversion to unfavorable inequality, or distaste to be "behind", β to favorable inequality, or distaste to be "ahead". We are going to explore in the following sections the relation between those parameters and the level of cooperation which can be attained in team production when a temptation of free-riding exists. Notice that, as the

model is presented by Fehr and Schmidt as incorporating inequity aversion, here the concept of inequity is reduce to inequality or, in other words, equitable outcomes are the ones where material payoffs are the same. This coincidence makes sense in reference to results in experiments, where subjects are ex ante equal (moreover they are typically all undergraduate students) and are randomly assigned their roles in the game to be played. In order to make the formulation in (1) valid in the more general contexts, where real effort decisions are taken, we are going to assume that effort can be expressed in monetary terms, so that the disutility it entails is equivalent to how much an individual would be willing to pay to lower his effort without any consequence on material payoffs.

In our comparative perspective, it is crucial to understand how to interpret the low levels of cooperation among Japanese subjects when playing anonymously public good games in terms of parameter levels. One possibility would be to assume that pure selfishness (i.e. $\alpha=\beta=0$) is more frequent among Japanese subjects. However, this idea would be in contradiction with Japanese agents' willingness to sacrifice their own interest in order to hurt defectors. The latter phenomenon can be explained in terms of higher frequency of high levels of α parameters. In fact, if a subject with a given payoff gets disutility from being worse off than another player, he is willing to punish her as long as the decrease in her payoff is higher than the amount he has to spend, if his exceeds a critical level. On the other hand, in standard dilemma games high levels of α in the population do not foster cooperation, but rather reinforce the tendency to defect, as being the only cooperator implies disutility from being poorer than the opponents.

Large presence of high-β individuals would appear to foster cooperation. This comes from the fact that defection may lead to disutility from being richer than those who cooperate. However, such type of individuals may still defect if they expect others to do the same; in fact, low levels of trust is, as we have seen, Yamagishi's explanation for his results in Japan.

The elements forming peer pressure, a phenomenon typical of Japanese firms, high α's are a necessary condition for the "shame" effect, which occurs when individuals monitor each other and express contempt for free-riders; β is linked to "guilt", which may be felt by who cooperates less than his peers. Overall, we argue that experimental evidence speaks in favor of a comparatively higher frequency in Japan of individuals characterized by high levels of α, while we cannot at the moment draw definitive conclusions about cross-cultural differences in levels of β. As a consequence, we are going to concentrate especially on the possible effects of this difference in team production settings, characterized as "dilemma games".

3. The team production game and peer sanctioning

We are going to represent team production as characterized by the following payoff formulation for group members:

$$\hat{\pi}_i = 1 - g_i + \gamma \sum_{j=1}^{N} g_j \qquad (2)$$

where g_j (constrained to be non-negative) is the contribution of a given player j and γ the public good production technology.

As usual, it is assumed that the following holds:

$$\frac{1}{N} < \gamma < 1$$

so that under standard assumptions about rationality and selfishness no positive g is chosen and this leads to Pareto inefficiency.

This basic representation permits to highlight the free-rider problem in a somewhat more extreme fashion with respect to the case where the production and the effort cost functions are non-linear, so that individual effort is suboptimal but not absent. We choose this representation for sake of simplicity, as it captures the essence of the team production problem. The zero effort choice, moreover, should not be taken literally but rather as the minimum effort compatible with keeping the job.

We employ a one-shot interaction setting both, again, for sake of simplicity and to represent an extreme case where cooperation is difficult to achieve. Of course in many situations workers do repeat their interaction; however, as long as the time horizon is finite, which is typically the case when working on a specific project or in a firm employing job rotation, backward induction leads to the emergence of equilibria characterized by free-riding.

Job rotation is in fact one of the characteristics of the typical Japanese firm; for instance, Sako (1994) points out that "horizontal job transfers are also abundant, facilitated partly by the relative absence of occupational consciousness guarding task demarcation". Although job rotation does also take place within the same group, it also occurs, as Itoh (1994) mentions, that "extensive rotations enable them [the employees] to work in various departments, regional offices, or factories, and as a result there is little chance that they will work together in the same place, particularly in a large firm". So the rationale for the existence and effectiveness of mutual monitoring and sanctioning can hardly be based on "folk theorem" type of argument, as the team production game is not "infinitely" nor "indefinitely" repeated among the same group members.

At the same time we argue that long term employment, a distinctive Japanese feature (see, among the other, Ouchi, 1981), does make a difference in the effectiveness of peer sanctioning. In particular, if an individual earns the reputation of being a free-rider, the damages he is going to suffer are obviously larger than what would be the case were he given the possibility to leave the firm at a low cost. This provides high possibility for hard working people to hurt fellow workers who shirked, even when this occurs in the last period of permanence in the same working team. Other features of the Japanese firm seem to be going in the same direction. Quality circles and team meeting, for instance, are indicated in Kandel and Lazear (1992) as mechanisms of investing in the peer pressure function. Also the participative decision making and the encouragement and organization by the company of social activities contribute to create an environment where peers can substantially affect each other's well-being. This whole system of attributes contribute to create an effective "punishment technology" which is not available in standard "Western" organizations. As with the production of the public good, we take a linear form similar to the one proposed by other authors (see, for instance, Fehr and Gaechter, 2002, and Sefton et al., 2002), as it allows a rapid derivation of the basic intuitions about the consequences of social preferences.

The punishment technology implies that a player can decrease the payoff of one or more opponents at his choice. We assume that to implement a reduction p of an opponent's payoff, a

player has to spend a quantity *cp*, with *0<c<1*. So the final payoff function is determined as follows:

$$\pi_i = \hat{\pi}_i - c\sum_{j=1}^{N} p_i^j - \sum_{j=1}^{N} p_j^i \qquad (3)$$

where p_r^s are the punishment points inflicted by player *r* upon player *s*. The final payoff, then, is obtained subtracting the cost for punishing opponent and the damage of being punished by them. Naturally, all *p* terms are non-negative, as we do not allow transfers among agents.

4. Unfavorable inequality aversion: α-positive players' behavior in a random matching environment

In this section we focus on the effects of variations of α parameter in the preferences of players in the population, who are randomly matched in couples to play a one-shot game according to the rules just described above. We assume that $\beta=0$ always holds for all players and that each player is informed about the frequency of types in the population, but cannot observe his specific opponent's type. We will discuss later the consequence of allowing for positive β's and for larger groups. We also assume that a player's type does not affect his expectation about his opponent's, which is determined by the frequency in the population, which is approximately true if the population is large[2].

If payoffs are determined as in 2), and if players are selfish (i.e. $\alpha=\beta=0$), all players choose 0. The equilibrium does not change if any number of players display positive α's. The best response when all opponents choose 0 remains 0: indeed, a positive contribution would give the player choosing it not only a lower payoff level, but also disutility from being behind with respect to his opponents. The best response to any positive numbers is also 0, as no disutility is derived from being ahead in the payoff distribution. So 0 is a dominant strategy and (0,0) is the Nash equilibrium.

Things change if we assume that, after playing the voluntary contribution game, each player is given the possibility to reduce the payoff of his opponents according to the formulation in 3). As specified above, punishing entails a cost: a player *i*, in order to decrease the opponent's payoff by p_i^j, has to spend cp_i^j, with *0<c<1*. This constraint on the cost function, also present in the above mentioned studies on the effects of punishment, is very important. On one hand, it is obviously positive, which implies, together with the one-shot structure of the game, that punishing unambiguously reduces the punisher's payoff. On the other, the upper bound 1 is such that the sacrifice for the punisher is lower than the damage for the punished.

---

[2] In a finite population, the probability of being matched with a certain type would be lower for a player of the same type than for a player of a different one, in an environment where the frequency is known. For instance, a type *x* in an *N*-size population with has probability $\frac{N_x - 1}{N - 1}$ to meet his own type (whose frequency is N_x, while a different type has $\frac{N_x}{N - 1}$. Clearly, the difference between the two probabilities vanishes as N gets very large.

What happens in the punishment phase? Clearly, players characterized by $\alpha=0$ do not spend anything to punish. If a player is selfish, he does not take an action which reduces his own payoff, regardless of the possible effects on his opponent. As the only deviation from selfishness has to do with inequality aversion, punishment actions may only be taken a player who had contributed more than his opponent, as we rule out perverse equilibria of the punishment subgame, where for instance equality could be achieved with both players equally punishing each other after having chosen the same contribution level. An inequality averse player whose contribution was higher may decide to punish. Denoting players "1" and "2", we get the following definition of player 1's utility function (and a symmetric definition of player 2's):

$$u_1(g_1, g_2, p_1^2) = 1 - g_1 + \gamma(g_1 + g_2) - \alpha_1(\max\{0, g_1 - g_2 - p_1^2 + cp_1^2\}) - cp_1^2$$

and

$$\frac{\partial u_1}{\partial p_1^2} = \begin{cases} \alpha(1-c) - c & \text{if } g_1 - g_2 - p_1^2 + cp_1^2 > 0 \\ -c & \text{if } g_1 - g_2 - p_1^2 + cp_1^2 < 0 \end{cases} \quad (4)$$

Clearly, two possibilities exists: either the player does not punish, or he does enough to equalize payoffs, for which it must be that:

$$\alpha_1(1-c) - c \geq 0$$

or

$$\alpha_1 \geq \overline{\alpha} = \frac{c}{1-c} \quad (5)$$

This necessary condition becomes sufficient if we use the strict inequality. As could be expected, the threshold which makes players punish depends on the parameter *c*. Broadly speaking, if it is very costly to punish, a player does not do it unless his aversion to unfavorable inequality is very high. In particular, if *c* approaches the upper limit 1, which would imply that the cost for the punisher is almost the same as the damage for the punished, only an (implausible) infinite aversion to inequality would justify a punishment action. However, it is not the case that the higher *c* the better for free-riders without ambiguity. In fact, *ceteris paribus* a higher cost *c* implies that the magnitude of punishment required to equalize payoffs is greater, so if a player is punished, her payoff is reduced more than what would be the case with a lower cost. More specifically, equalizing payoffs implies that:

$$g_1 - g_2 - p_1^2 + cp_1^2 = 0$$

so that if $g_1 > g_2$ we get:

$$p_1^2 = \frac{g_1 - g_2}{1-c} \quad (6)$$

which is of course increasing in *c*. Therefore, the higher *c*, the higher the minimum α which makes a higher contributor punish is opponent, but the higher the magnitude of punishment once it happens.

4.1. Homogeneous players: subgame perfect equilibria and evolutionary analysis

If the whole population is made of individuals characterized by $\alpha < \bar{\alpha} = \frac{c}{1-c}$, no punishment occurs and, expecting it, all players do not contribute anything. Things change if the common α exceeds $\bar{\alpha}$.

Proposition 1. *If all players in the population are characterized by (5), then for every* $g \in [0,1]$ *there exists a subgame perfect Nash equilibrium where both players choose it as contribution level.*

Proof. First notice than any Nash equilibrium in the subgame following the contribution phase results in the equalization of payoffs. In fact the worse off player would otherwise have incentive to deviate from his strategy and increase his punishment action.

Now take a strategy profile $\left(s_1^*, s_2^*\right)$ in which both players pick a contribution level g_1^*, player 1 sets $p_1^2 = 0$ whenever $g_2 \geq g^*$ and $p_1^2 = \frac{g^* - g_2}{1-c}$ otherwise, player 2 sets $p_2^1 = 0$ whenever $g_1 \geq g*$ and $p_2^1 = \frac{g* - g_1}{1-c}$ otherwise. This strategy profile is a subgame perfect Nash equilibrium. First, it is indeed the case that in every possible subgame following the contribution phase payoffs are equalized, as required by subgame perfection. Moreover, given this punishment profile, no player has incentive to deviate from contributing $g*$. In fact, selecting $g > g^*$ cannot be profitable: it provides a lower payoff in the contribution game which gets lowered after punishing in order to avoid inequality. g^* is a better response as it gives a higher payoff in the contribution game and no inequality to be "corrected".

For any level $g < g*$, as punishment occurs with probability one we get the following expected utility:

$$U(g) = 1 - g + \gamma(g + g^*) - \frac{g^* - g}{1-c}$$

from which

$$\frac{\partial U}{\partial g} = -1 + \gamma + \frac{1}{1-c}$$

As *c<1*, $c < 1, \frac{\partial U}{\partial g} > 0$ always holds in this domain, so choosing $g < g^*$ cannot be part of an equilibrium as the player would always deviate by increasing his contribution. ***Q.E.D.***

This result implies, in particular, that full cooperation, i.e. contribution 1 by both players, can be sustained as a subgame perfect Nash equilibrium.

In the remainder of this section, we present an evolutionary analysis[3] based on the assumption that players are "programmed" to punish lower contributors as to make payoffs equal. Recall that we rule out perverse cases of punishment towards higher or equal contributors, that is equivalent to say that among the punishment profiles that lead to payoff equalization, they are able to coordinate to pick the least "destructive" one. With this assumption in mind, we can actually refer to contribution levels as "strategies". A first result stemming out of the previous proposition is that any contribution level is an evolutionarily stable strategy, as it is strictly best response to itself.[4]

So far we have found that, unlike what happens with purely selfish players, equilibria with positive contribution levels can be sustained on the basis of aversion to be behind which cannot be seen as a form of "altruism", or positive orientation towards peers as, instead, could be the case for β-positive players. Here the only way in which players can take others' welfare into account is "negative", in form of envy for the better off. The question remains about how to select a particular equilibrium. Let's assume, for simplicity, that players choose between two contribution levels, 0 or 1, as part of their strategy. Denoting $E\pi_x$ the expected payoff from choosing strategy x, and p_x the frequency with which it is chosen, we find:

$$E\pi_1 = \rho_1 2\gamma + (1-\rho_1)(\gamma - \frac{c}{1-c})$$

$$E\pi_0 = \rho_1(1+\gamma-\frac{1}{1-c}) + (1-\rho_1)1$$

so that:

$$E\pi_1 - E\pi_0 = \rho_1(\gamma - 1 + \frac{1}{1-c}) + (1-\rho_1)(\gamma - 1 - \frac{c}{1-c})$$

At this point we have a formal expression of the fact that a cooperator does better than a defector when meeting another cooperator, in fact $0 < c < 1$ implies that $\lambda - 1 + \frac{1}{1-c} > 0,$ as the latter punish defectors harshly enough to overcompensate savings on contribution. On the other hand, a defector does relatively better with another defector as not only he saves on contribution (as $\lambda - 1 < 0$), but also on the punishment cost $\frac{c}{1-c}$.

---

[3] See Maynard Smith (1982) and Maynard Smith and Price (1973) for a description of the basic foundations of evolutionary game theory.

[4] Indeed, if we allow strategies which do not include punishing, then we find that contributing the same quantity but not punishing can invade a population of players contributing g^* and punishing lower contributors. In fact, both strategies produce the same expected payoffs. But then, if the non-punishing strategy gets spread, it is possible for free-riding strategies to "invade" the population, if the frequency of punishers gets low.

After simple passages we get:

$$E\pi_1 - E\pi_0 = \frac{\rho_1(1+c) - c + (\gamma - 1)(1-c)}{1-c}$$

so that:

$$E\pi_1 - E\pi_0 > 0$$

if and only if:

$$\rho_1 > \rho_1^* = \frac{1 + c\gamma - \gamma}{1-c}$$

Notice that $0 < 1 + cy - y < 1 + c,$ so that $0 < p_1^* < 1.$ Therefore there always exists a threshold such that cooperation has a higher expected value than defection if and only if the frequency of cooperation exceeds that threshold. It is immediate to see that the higher γ, the lower the threshold (recalling $c < 1$). As we could expect, the higher the benefit from the public good, the lower the benefit from free-riding and the easiest to achieve the condition for making cooperation the best strategy. The following also holds:

$$\frac{\partial \rho_1^*}{\partial c} = \frac{2\gamma - 1}{(1+c)^2} > 0$$

as we assumed $\gamma > \frac{1}{2}.$ That is, the higher the cost of punishing, which is born by a cooperator when meeting a defector, the hardest it is to meet the condition for cooperation to be the best value choice.

Using the concept of replicator dynamics (Taylor and Jonker, 1978), we have:

$$\dot{\rho}_1 = \rho_1(1 - \rho_1)(E\pi_1 - E\pi_0)$$

So $p_1 > p_1^*$ is the condition for cooperation to spread across the population, until full cooperation, $p_1 = 1,$ is reached. On the contrary, defection spreads if $p_1 < p_1^*$ while if $p_1 = p_1^*$ the frequency of cooperation remains stable.

What we have found implies that if a population composed of players characterized by $\alpha \geq \frac{c}{1-c}$ is sufficiently cooperative, the availability of a punishment device makes cooperation spread over the population. In particular, an initial condition of full cooperation is sufficient to foster the reproduction of this state, although not necessary. In absence of a punishment device, instead, if a "mutant" strategy of defection entered the population it would provide higher payoffs and therefore be imitated and drive away cooperation. Clearly, the same would occur in presence of the punishment device, but with levels of disadvantageous inequality aversion not high enough to make players be willing to punish.

Summing up, we have seen that, unlike in standard prisoner dilemmas, initial levels of cooperation matter. Namely, as *ceteris paribus* cooperators (contributing 1) do relatively better the higher the cooperation frequency, the initial level of cooperation is decisive in determining towards which evolutionarily stable equilibrium the population will converge.

4.2. Heterogeneous players

We now consider the possibility that players differ in their α parameter, and find the following result:

Proposition 2. *If the proportion* θ_α *of players in the population characterized by* $\alpha \geq \frac{c}{1-c}$ *exceeds* $(1-\gamma)(1-c)$, *then there exists a subgame perfect Bayesian Nash equilibrium with equal contribution level, for any possible value of this level.*

Proof. Assume that all players characterized by $\alpha \geq \frac{c}{1-c}$ decide to punish according to 6), while players with $\alpha < \frac{c}{1-c}$ do not exert any punishment. These punishment strategies are compatible with subgame perfection. Then for any level g^* the best response for any player is to select the same contribution. We know that a level $g > g^*$ cannot be best response to g^* (see proof of Proposition 1 above). Focusing on contributions $g < g^*$, we find that the expected utility of replying g to g^* is given by:

$$U(g) = 1 - g + \gamma(g + g^*) - \theta_\alpha \frac{g^* - g}{1-c}$$

so that

$$\frac{\partial U}{\partial g} = -1 + \gamma + \frac{\theta_\alpha}{1-c}$$

So, if $\theta_\alpha > (1-\gamma)(1-c)$ the expected utility is always increasing in the contribution level, therefore a contribution level $g < g^*$ cannot be the expected value maximizer. We therefore find that the best response to g^* is g^* itself. ***Q.E.D.***

So we find that if the presence of players with a sufficiently high α reaches a critical level, the chances to be punished makes a lower contribution a "wrong" choice. Notice that, once we are excluding as possible best reply a higher contribution (which could only be compatible with some forms of altruism or efficiency oriented preferences), the expected utility *U(g)* is the same both for selfish and for α-positive players themselves. In other words, their presence "discipline" both selfish and their own type, discouraging free-riding if the opponent cooperates.

Under these conditions, we find that any combination of equal contribution levels constitute a subgame perfect Bayesian Nash equilibrium. This result is a natural complement to Proposition 1, where players were assumed to be all characterized by the same aversion to unfavorable inequality, which makes them punish in case of being worse off. Here we ask for less, as the threshold $(1-\gamma)(1-c)$ is obviously lower than one. Notice that it is decreasing in γ, which is

not surprising: the higher γ, the lower the temptation to free-ride. More interestingly, it is also decreasing in c. This is related to the fact that, as pointed out earlier, the higher c, the bigger the magnitude of punishment once it occurs. So a high c makes it harder to meet the condition of a sufficiently high α, i.e. $\alpha > \frac{c}{1-c}$, in a given individual, but it reduces the proportion of people sharing this feature that is necessary for the existence of cooperative equilibria, as punishment is heavier for lower contributors.

The analysis of evolutionary stability changes when not all players share the same high level of α. In particular, it makes it harder to meet the condition $E\pi_1 - E\pi_0 > 0$, as now the defector may go unpunished even if he meets a cooperator. The only modification in analysing this condition is, in fact, that the probability that a cooperator prefers to punish is lower than 1. Of course, the fact that a player who has defected may be averse or not to unfavorable inequality is not relevant, as he does not punish in any case.

Let us analyse the variations in expected value for a player characterized by $\alpha > \frac{c}{1-c}$.

The expected payoff from cooperation does not vary:

$$E\pi_1 = \rho_1 2\gamma + (1-\rho_1)(\gamma - \frac{c}{1-c})$$

The expected payoff from defection does change. Denoting by ω the probability that a cooperator is characterized by $\alpha \geq \frac{c}{1-c}$ we get:

$$E\pi_0 = \rho_1(1+\gamma-\omega\frac{1}{1-c}) + (1-\rho_1)1$$

As in principle any frequency of cooperation and any relationship between cooperation and type is admissible as initial, exogeneous condition, we remain agnostic about the specific value ω may take, but it is clear that if the lower ω the higher $E\pi_0$.

The condition for $E\pi_1 > E\pi_0$ is found to be:

$$\rho_1 > \rho_1^{**} = \frac{1+c\gamma-\gamma}{\omega+c}$$

Clearly, if $\omega < 1$ the new threshold is higher. As a matter of fact, if $\omega < (1-\gamma)(1+c)$, we get that the threshold is higher than 1, so that it cannot be reached by any initial cooperation level.

About low α players, it is easy to see that while their expected payoff of defecting is the same as for the high α players, their expected payoff of cooperating is higher, as they do not undergo the cost of punishing if they meet a defector. Therefore, $\rho_1 > \rho_1^{**}$ is sufficient to ensure that these players have a higher expected payoff from cooperation than from defection.

For what we have seen, if we assume that in a population where original preferences are stable, but the frequency of choices between cooperation and defection is allowed to vary according to

the replicator dynamics described above, the removal of the assumption that all players share a high α makes it harder, or even impossible, to meet the condition for spreading cooperation.

4.3. Playing sequentially

We have seen that the presence of a proportion of players willing to punish to avoid unfavorable inequality creates the existence of equilibria, where contribution levels are the same. This result is a relevant change with respect to dilemma games under standard assumptions of selfish preferences, but we are left with the problem of equilibrium selection. While we have showed the conditions under which evolutionary forces can lead to cooperation, here we propose an intuitive variation of our dilemma game, where one player moves first. In some settings both players interacting ignore the opponent's choice, but in many situations the timing is different: for instance, when the tasks are intrinsically sequential, which is not infrequent in team production.

In our game, we get the following:

Proposition 3. *If* $\theta_\alpha > (1-\gamma)(1-c)$, *and if players characterized by (5) punish according to (6), then the only equal contribution level which is part of subgame perfect Bayesian Nash equilibrium, when players choose contributions sequentially, is the maximum contribution.*

Proof. We have seen in Proposition 2 that if these assumptions are met, the best reply to any contribution level is to equalize it. Knowing that the second mover will do so, and that therefore no inequality will arise, the first mover's problem can be written, as $\max_g 1-g+2\gamma g$.

As $\gamma > \frac{1}{2}$, the argument is increasing in *g*. ***Q.E.D.***

Recall that we are assuming that the preferences in the population are common knowledge, unlike the ones of any specific individuals. So, as $\theta_\alpha > (1-\gamma)(1-c)$, i.e. the probability that a given player is willing to punish if the opponent contributes less than himself exceeds a critical level, the second mover prefers to equalize the first mover's contribution. That is, her belief about the chances that the first mover is a high α type leads to this strategy. As the first mover is sure about the fact that his choice will be equalized, he decides to contribute the maximum level *regardless of his own type*. What really counts is that he believes that the second mover assigns a sufficiently high probability to the fact that he is a high α type, so that she believes that contributing less is a bad idea.

5. Effects of group size and aversion to being ahead

5.1. The N players case

We now explore how things change in the more general case where group size can take different values.

The formulation in (1) shows that the disutility player *i* derives from the fact that another player is better off by a quantity *x* is $\frac{\alpha}{N-1}x$. As a consequence, as we move from the $N=2$ case to the possibility of larger groups, we find that a player suffers less for a single case of a

richer opponent. This, of course, has implications on his willingness to spend to reduce her payoff. Namely, a necessary condition for punishment to occur is that $\alpha \geq \frac{c}{1-c}(N-1)$. More precisely, in the punishment stage, for player *i* to be willing to punish it must be that the utility gain from inequality reduction $\frac{\alpha}{N-1}(1-c)$ compensates the cost *c*. However, this necessary condition is not sufficient: it might be that by spending *c* player *i* increases the disadvantageous inequality with respect to all other players. We will get back to this point shortly. As we could intuitively expect, cooperative equilibria do exist also in *N*-size groups formed out of a population with homogeneous players sharing a sufficiently high level of unfavorable inequality aversion:

Proposition 4. *Assume that all players in an N-size group are characterized by* $\alpha \geq \frac{c}{1-c}(N-1)$. *Then for any contribution level* g^* *there exists a subgame perfect Nash equilibrium where all players contribute* g^* *as part of their strategy.*

Proof. First, it is clear that no players have incentive to contribute more than $g*$. Consider a downward deviation by player *i*. In the subgame following the contribution phase, as $c<1<N-1$ there exists a strategy profile where $p_k^i = \frac{g^* - g}{N-1-c}$ for all $k \in \{1,...,i-1,i+1,...,N\}$ such that all final payoffs are equalized. This strategy profile constitutes a Nash equilibrium in the subgame. In fact, no player has incentive to deviate increasing his punishment towards i, as this would imply an additional cost and no higher utility comes from making *i* poorer. Consider a downward deviation, i.e. that a player *j* chooses $p_j^i = \frac{g^* - g}{N-1-c}$. We get:

$$\frac{\partial u^j(\bar{g},\bar{p})}{\partial p_j^i} = -c + \frac{\alpha}{N-1}(1-c) \geq 0$$

by assumption.

Clearly, if the players who contribute g^* play the profile just shown, the downward deviation by player *i* is not profitable. In fact, for any $g < g^*$ we get:

$$U(g) = 1 - g + \gamma(g + (N-1)g^*) - \frac{g^* - g}{N-1-c}(N-1)$$

from which

$$\frac{\partial U}{\partial g} = -1 + \gamma + \frac{N-1}{N-1-c} > 0$$

Q.E.D.

We already pointed out that, for a given *c*, the level of inequality aversion sufficient to justify punishment is higher as it must be that $\alpha \geq \frac{c}{1-c}(N-1)$. However, it is also clear that the "punishment profile" is not the unique equilibrium in the subgame following a defection. In

particular, no punishment is also an equilibrium as long as $c > \frac{1}{N-1}$: in this case, a single player has no incentive to be the only one engaging in punishment even if he is infinitely inequality averse. This comes from the fact that punishment by a single player makes him suffer from inequality towards all those players who do not punish and therefore do not bear a related cost. So, the marginal effect of increasing punishment on unfavorable inequality is $1-c-c(N-2)$ where the latter term involve the non-punishing players.[5] Indeed, if instead $c < \frac{1}{N-1}$, then the best response by a highly inequality averse player to zero punishment by all the other cooperative players is to punish also them, besides the defector, to equalize all payoffs.

5.2. The $\beta > 0$ case

In the two-player case where the contribution game is played simultaneously, cooperative equilibria can be sustained even without punishment opportunities when players are characterized by $\beta \geq 1-\gamma$. In fact, in terms of utility levels, this change makes our game lose the feature of a "prisoner dilemma" as players genuinely prefer to cooperate as much as their opponent do so, as they would suffer from advantageous inequality created by contributing less. Therefore, any contribution profile including the same contribution is part of a Nash equilibrium in a standard public good game. Of course the presence of punishment opportunities does not change this result: when agents pick the same contribution level no punishment occurs, irrespective of players' levels of α.

If instead $\beta < 1-\gamma$ there is no substantial change in the previous equilibrium analysis: free-riding dominates in the standard public good game while punishment opportunities can sustain cooperative equilibria if combined with sufficiently high α's.

In terms of evolutionary analysis nothing changes, as it is assumed that strategies spread according to payoff levels, rather than to utility levels.[6] So, in the game including punishment α does play a role in determining payoff differentials between cooperators and defectors, as the latter's payoff may be reduced in the punishment phase, but the possible internal feeling of guilt suffered by free-riders is irrelevant for the spreading out of cooperation or defection.

When players are assumed to be heterogeneous, it is still the case that a selfish player's best response to any cooperation level is to free-ride completely unless he expects with more than $(1-\gamma)(1-c)$ probability to encounter a player characterized by $(1-\gamma)(1-c)$, in which case he would rather equalise his contribution level. Whether he expects to meet a high β type is irrelevant for his choice. On the other hand, we stated above that a player with $\beta \geq 1-\gamma$ always prefers to meet his opponent's contribution level. The fact that, in a population with a

---

[5] In general, an equilibrium with punishment in the subgame following the first phase can only be sustained if (1) m players among cooperators are characterized by $\alpha \geq \frac{c}{1-c}(N-1)$ and (2) $c < \frac{1}{N-m}$.

[6] If that were the case, we would find again that cooperation sustains itself, as high players would be satisfied with cooperative choices if and only if they were matched with other cooperators.

proportion of high α players exceeding $(1-\gamma)(1-c)$, it is still the case that any profile including the same contribution level can be characterized as a subgame perfect Bayesian Nash equilibrium holds *a fortiori* if some agents are characterized by high β. If the high-α proportion is lower, the presence of high β players does not make cooperative equilibria arise in those interactions involving players with $\beta < 1-\gamma$, whose choice is to free-ride in any case.

When the game is played sequentially, instead, the presence of high β players can potentially affect the behavior of all types of agents. In the previous analysis we have seen that a player chooses the maximal level of contribution if he expects his opponent to believe that the probability that he (the first mover) would punish a lower contribution is high. Now we have an alternative route to full cooperation in the sequential public good game: a first mover could contribute 1 if he expects the second mover to be a high β player, namely that she has a $\beta \geq 1-\gamma$ with a probability exceeding $\frac{1-\gamma}{\gamma}$.[7] In fact, in that case the expected return from contributing 1 is the highest possible as a β player would "return the favor". Of course, only real β players do, while all the others free-ride irrespectively of first mover's choice. Summing up, in the sequential case we find that if the frequency of α players is high enough (recall we found $(1-\gamma)(1-c)$ as the threshold frequency) both first and second movers pick the maximum contribution in equilibrium, irrespectively of their actual type and of the β values in the population. If the frequency of high α players is not high enough, then first movers still pick 1 if the proportion of high β agents exceeds $\frac{1-\gamma}{\gamma}$, but only actual high β types equalise this contribution levels, while the others pick 0.

Interesting changes with respect to the previous analysis are created in the *N*-players case. The basic result follows:

Proposition 5. *If players' preferences are such that* $\frac{\alpha(1-c)+\beta(N-2)c}{N-1} \geq c$ *then for any value* $g^* \in [0,1]$ *there exists a subgame perfect Nash equilibrium where all players pick* $g^* \in [0,1]$ *as their contribution level.*

Proof. We show that in case all players pick g^* except for one player who selects a lower level, there exists a Nash equilibrium in the subgame following the contribution phase where all the others $N-1$ players punish the lower contributors in such a way that all final payoffs are equalised. Assume that player *j* has contributed g^*, player *i* a lower level and that all the other players are equally assigning punishing points $p_r^i = \frac{g^* - g}{N-1-c}$ such that if also *j* takes the same choice all payoffs are equalised. Then, if $p_i^j < p_r^j$ the following holds:

[7] In this case, it is easy to see that for the first mover $\frac{\partial U}{\partial g} = \gamma - 1 + q\gamma$, where q is the probability that the second mover is characterized by a high ß, which implies that she would equalize his contribution.

$$\frac{\partial u^j(\bar{g},\bar{p})}{\partial p^i_j} = -c + \frac{\alpha}{N-1}(1-c) + \frac{\beta}{N-1}(N-2)c \geq 0$$

by assumption. The term $\frac{\beta}{N-1}(N-2)c$ arises from the fact that, being $p_i^j < p_r^j$, all the other players except for the lower contributors are worse off than *i*, therefore any additional unit of p_i^j reduces the inequality between *i* and every agent among the $(N-2)$ "punishers" by *c*, the cost born by *i*. Therefore it is a best response by player *i* in the subgame to issue the same level of punishment p_r^j, and by extension no player among the $(N-1)$ who contribute g^* have incentive to deviate from the level of punishment p_r^j. A punishment profile where this is their choice (while of course the lower contributor does not punish anybody) is a Nash equilibrium in this subgame. But in this case it cannot be optimal for the *j* player to issue a lower level of contribution, as for sure the payoff after the punishment phase is lower than what he would achieve by contributing $g*$, and in both cases all payoffs end up at the same level (so no disutility from inequality is suffered).

So any strategy including g^* as contribution level is a best response to the strategy profile by the remainder of the players which include, for all of them, contributing g^* and then punishment p_r^j in case one player contributed less. In particular this holds for the strategy including g^* and p_r^j as punishment in case of one single lower contributor, which is then best response to itself and part of a symmetric subgame perfect Nash equilibrium. ***Q.E.D.***

We had seen previously how increasing the group size had unambiguously "bad" effects in achieving cooperation based on fear of punishment, when $\beta = 0$. Now we see that if we allow for the existence of aversion to advantageous inequality, or "guilt", we find that the perspective of a "collective punishment" equilibrium following free-riding by one player is less complicated. Notice in fact that we can rewrite the assumption above as:

$$\alpha \geq \frac{c}{1-c}(N-1) - \beta(N-2)c$$

which is clearly less restrictive than the $\alpha \geq \frac{c}{1-c}(N-1)$ we have when assuming $\beta = 0$.

Let us observe now that:

$$\frac{\partial^2 u^i(\bar{g},\bar{p})}{\partial p_i^j \partial N} = \frac{-\alpha + \alpha c + \beta c}{(N-1)^2}$$

Therefore, we get that the effect of group size on the marginal utility of punishment is positive if $-\alpha + c(\alpha + \beta) > 0$. This implies that if the cost of punishment is high then it may indeed be easier to achieve a cooperative equilibrium sustained by punishment possibilities in a large group. The reason is, that a high cost implies a big disutility from advantageous inequality when a player's coworker are punishing a free-rider. This result is somewhat in line with the intuition that when punishment is very costly it may be better to be able to share its burden with other coworkers; this structure makes it clear the relevance of "guilt" to sustain this aspect of punishment. Clearly, all these results are based on the assumption that contributions, or

effort levels, are perfectly observable, an assumption which could be hard to meet if the group gets very large.

6. Inequality aversion, cooperation, cultural differences and organizational practices: a discussion

We started off arguing that experimental evidence casts doubt on the general idea that Japanese people share a higher "cooperative attitude" when compared to Western subjects. In fact, evidence was found where cooperation rates in standard public good games where even lower in Japan. This seems at odd, at first sight, not only with "common wisdom" but especially with the fact that in Japan it is common to organize production in teams, a practice which may in principle create free-riding problems. It is also well known that Japanese team production goes hand in hand with mutual monitoring and peer pressure among group members. Indeed, additional experimental evidence shows a higher tendency in Japan to be willing to spend in order to punish free-riding and in general behavior which can be regarded as "exploitative". It remains to be explained why agents exert pressure when it is costly to do so. An explanation solely based on higher future benefits from stimulating cooperation by previous "defectors" is not compatible with the finiteness of the interaction among group members; indeed, punishment occurs also in experiments where the number of repetitions is announced.

We referred to the formulation in Fehr and Schmidt (1999) due to its simplicity and its compatibility with a wide array of experimental evidence. Then we interpreted the differences found in Japanese agents' behavior as based on higher aversion to unfavorable inequality (the α parameter). Our formal analysis showed that indeed cooperative equilibria exist even in one-shot games when punishment opportunities are provided, under suitable assumptions combining α and the "punishment technology". In two-player interactions, this is due to the threat of punishment which makes any contribution level the best response to itself. Moreover, we find that in an evolutionary setting the initial frequency of cooperation is critical for its diffusion across the population, again in case players' α is high enough. Then we extended our analysis to the heterogeneous players case, to the possibility that aversion to favorable equality also exists and to the *N*-player case. In all those cases we find conditions including inequality aversion as a necessary condition for achieving cooperative equilibria among the possible ones. It is then clear that providing punishment possibilities is not per se welfare enhancing. In a homogeneous players setting where the level of α is low punishment possibilities will not be used; the same holds in a heterogeneous environment if the proportion of high α players is not sufficient to make cooperation the best response to itself, and as a consequence free-riding prevails.[8] As it is possibly costly for a firm to set up meetings, social activities, a long term employment structure beyond optimality requirement, then it makes sense to set-up a "punishment technology" only when inequality aversion levels in the population are high. This could be an explanation for the Japanese tendency to organize production in group, not related to differences with respect to Western firms concerning the team production technology. Indeed, our results hold even if we assumed that $\gamma < 1/N$, with the exception of the equilibrium in the sequential contribution case. While it is clear that in general $\gamma > 1/N$ is a reasonable

[8] Notice, however, that in equilibrium it does not occur that inequality averse agents cooperate, so that we do not get welfare reduction due to punishment actions.

assumption when production is arranged in teams, what we find is that neither the efficiency nor the immediate consequence of team production on individual salaries are the driving forces of our results. What really matters is the fact that a cooperator dislikes an outcome where the free-rider is better off because of not having exerted a costly effort.[9]

7. Conclusions

Our analysis shows a simple way to reconcile seemingly contradictory evidence of low cooperation of Japanese subjects in dilemma experiments and the tendency by Japanese firms to organize production in teams. We argue that typical Japanese organizational practices, such as long term employment, team meetings and quality circles, various social activities in which workers are encouraged to participate and so forth, contribute to create opportunities to punish those agents who do not cooperate to team production as much as their peers. Punishment as an action implying a cost for the enforcer but a higher one for its recipient is a simple way of thinking about one aspect of "peer pressure", often referred to as another distinctive trait of the Japanese working environment. The idea of "shame" proposed by Kandel and Lazear (1992) is based on the monitoring effort by peers and an action of exerting pressure which creates a disutility for the recipient. Indeed evidence from "cross-cultural" experiments points out that Japanese subjects do punish more when confronted with uncooperative behavior. On the other hand, no clear evidence has been found on differentials in "guilt", the part of peer pressure related to internal negative feelings suffered by free-riders independently of being monitored by their coworkers.

One effect of different levels of cooperation is that, when the benefits of team production are equally shared, a free-rider is better off than a high contributor. We apply the idea of inequality aversion to express the distaste the latter feels when comparing his own payoff - with a monetary transformation of the disutility caused by effort - with the former's. Then we relate the tendency to punish, an action which reduces inequality, with the egalitarian culture Japanese agents are embedded in. We use the formulation in Fehr and Schmidt (1999) and assume that Japanese players are distinguishable for a high level of aversion to unfavorable inequality. This leads us to investigate the role of this parameter in making cooperative equilibria possible. We find the following results:

[9] Recall that we are assuming throughout the paper that we can express the cost of effort in monetary terms, as it is in principle possible to evaluate it as the money a player would be willing to spend not to exert it. Then the application of inequality aversion, more often related to income levels, becomes intuitive.

- In a random matching set-up with homogeneous preferences, high α lead to the existence of cooperative equilibria. Moreover, any contribution level, including the highest, is an evolutionary stable strategy once we assume players do punish in the second phase as their utility functions command.

- In the heterogeneous players case, any level of equal contribution can be part of a subgame perfect Bayesian Nash equilibrium, if the proportion of players with a sufficiently high α reaches a given threshold.

- If the contribution game is played sequentially, and the same conditions as in the previous points hold, then the only contribution profile which can be part of a subgame perfect Nash equilibrium is the maximum level for both players.

- Increasing group size makes the emergence of cooperative equilibrium more difficult as long as we assume that players are not averse to favorable inequality. However, if we drop this assumption then increasing the size can have positive effects under appropriate conditions, as long as all contribution levels are perfectly observable by all group members.

- The other way positive levels of β can affect previous results occurs when they reach levels high enough, such that cooperation is a best response to itself irrespective of punishment opportunities, which of course implies that utility functions change the "prisoner dilemma" aspect of the public good game.

Overall, we have shown a variety of ways in which the aversion to being worse off than opponents can favor cooperation, when punishment opportunities are given. We maintain that what we find is consistent with the combination of that part of the Japanese cultural aspect of egalitarianism which is confirmed by experimental evidence and relevant organizational aspects observed in Japanese firms. Indeed, also the observed low levels of cooperation in standard dilemma experiments among Japanese subjects is totally compatible with high levels of unfavorable inequality aversion, as a player with high α has indeed one more reason to free-ride, based on the risk which cooperation entails of being worse off than his opponent. If, instead, punishment opportunities are given, as in the case of the typical Japanese firm, then cooperation can be sustained even in absence of "positive" feelings towards peers. The latter here are included as a possibility in terms of aversion to favorable inequality, a phenomenon still under debate. We find that adding positive β's has indeed important implications in terms of the effects of group size, as increasing it may foster cooperation in the punishment phase, while when all players are averse only to unfavorable inequality increasing group size undermines punishment behavior.

This contribution proposes one aspect at the basis of cooperation which is culturally based and is sustained via organizational practices and a general system of attributes of the Japanese firm. A consequence of this analysis is that these practices and attributes are not the best solution for any social and cultural environments, so that differences among Japan and Western countries could resist for a very long time. However, if some sort of "cultural convergence" occurs as societies are increasingly less isolated, things may change. For instance, a system of control based on peer pressure may lose effectiveness in the Japanese firm if individualism spreads across its members creating, among other effects, an envy-free environment; on the other hand, the same system may be effective in those Western contexts where significant frequency and intensity of "social preferences" emerge.

References

Alchian, A. and H. Demsetz (1972), "Production, Information Costs and Economic Organisation", *American Economic Review,* 62, pp. 777-795

Andreoni, J., W. Harbaugh and L. Vesterlund (2002), "The Carrot and the Stick: Rewards, Punishments and Cooperation", *American Economic Review*, 93 (3), June 2003, pp. 893-902

Aoki, M. (1988), "Information, Incentives, and Bargaining in the Japanese Economy", Cambridge University Press, New York, N.Y., U.S.A.

Aoki, M. (2001), "Toward a Comparative Institutional Analysis", MIT Press, Cambridge Ma.

Aoki, M. and R. Dore (eds.) (1994), "The Japanese Firm - Sources of Competitive strength", Oxford University Press.

Barron, J.M. and K. Paulson Gjerde (1997), "Peer Pressure in an Agency Relationship", *Journal of Labor Economics*, 15, 2, pp. 234-254.

Bolton, G.E. and A. Ockenfels (2000), "A Theory of Equity, Reciprocity and Competition", *American Economic Review,* 100, pp. 166-193.

Brandts, J. and E. Fatás (2001), "Social information and social influence in an experimental dilemma game", working paper, Instituto de Analisi Economico.

Cason, T. T. Saijo, T. Yamato and K. Yokotani (2004), "Non-Excludable Public Good Experiments", *Games and Economic Behavior*, 49, 1, pp. 81-102.

Dore, R. (1994), "Equality-Efficiency Trade-offs: Japanese Perceptions and Choices", in Aoki, M. and R. Dore, (eds.), (1994).

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

Fehr, E. and S. Gaechter (2000), "Cooperation and Punishment in Public Goods Experiments", *American Economic Review*, 90, 4, pp. 980-994.

Fehr, E. and S. Gaechter (2002), "Altruistic Punishment in Humans", *Nature*, 415, pp. 137-140.

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

Fischbacher, U., S. Gaechter and E. Fehr (2001), "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment", *Economic Letters* 71, pp. 397-404.

Itoh, H. (1994), "Japanese Human Resource Management form the Viewpoint of Incentive Theory", in Aoki, M. and R. Dore (eds.), (1994).

Kandel, E. and P.E. Lazear, (1992), "Peer Pressure and Partnerships", *Journal of Political Economy*, 100.

Keser, C. and van F. Winden (2000), "Conditional Cooperation and Voluntary Cooperation to Public goods", *Scandinavian Journal of Economics*, 10 (1), pp. 23-39.

Ledyard, J. (1995), "Public Goods: a Survey of Experimental Research", in Kagel, J. and A.E. Roth (eds.), Handbook of Experimental Economics, Princeton University Press, Princeton, N.J.

Maynard Smith, J. (1982), "Evolution and the Theory of Games", Cambridge University Press, Cambridge.

Maynard Smith and Price (1973), "The Logic of Animal Conflict", *Nature*, 246, pp. 15-18.

Nash, J.F. (1950), "Equilibrium Points in N-person games", Proceedings of the National Academy of Sciences, 36, pp. 48-49.

Ouchi, W.G. (1981), "Theory Z", Avon Books, New York, N.Y., U.S.A.

Rabin, M. (1993), "Incorporating Fairness into Game Theory and Economics", *American Economic Review*, 83 (5), pp. 1281-1302.

Sako, M. (1994), "Training, Productivity, and Quality Control in Japanese Multinational Companies", in Aoki, M. and R. Dore (eds.), (1994).

Sefton, M., R. Shupp and J. Walker (2002), "The Effect of Rewards and Sanctions in Provision of Public Goods", CEDEX Research Paper.

Selten, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory*, 4, pp. 25-55.

Taylor, P. and L. Jonker (1978), "Evolutionary Stable Strategies and Game Dynamics", *Mathematical Biosciences*, 40, pp. 145-156.

Yamagishi, T. (1988), "Exit from the Group as an Individualistic Solution to the Free Rider Problem in the United States and Japan", *Journal of Experimental Social Psychology,* 24, pp. 530-542.

Yamagishi, T. (1988), "The Provision of a Sanctioning System in the United States and in Japan", Social Psychology Quarterly, 51, pp. 264-270

Yamagishi, T., K. Cook and M. Watabe, (1998), "Uncertainty, Trust, and Commitment Formation in the United States and Japan", *American Journal of Sociology*, 104, pp. 165-194.