INTENTION-BASED ECONOMIC THEORIES OF RECIPROCITY

Darwin Cortes*
Received 18 September 2002, accepted 10 February 2003

ABSTRACT

In recent years, several experiments have shown individuals exhibit authentic reciprocal behaviour in anonymous one-shot interactions. As reciprocity has been shown to be relevant in several economic fields, there have also been several attempts to model reciprocal behaviour. I review the intention-based models of reciprocity and present an example of teacher management in the public sector in which the government offers an incentive scheme to implement a program. The incentive scheme has a prisoner's dilemma structure. In both simultaneous and sequential games, equilibrium results may differ from those predicted by standard theory.

Key words: Game theory, psychological games, Intention-based models, reciprocity.
JEL Classification: C700.

RESUMEN

Recientemente, varios experimentos han mostrado que los individuos exhiben un comportamiento auténticamente recíproco en interacciones anónimas que se dan una sola vez ('one-shot'). En tanto que se ha mostrado que la reciprocidad es relevante en múltiples campos de la economía, también ha habido varios intentos por modelar el comportamiento recíproco. Este documento revisa los modelos de reciprocidad que se fundamentan en las intenciones y presenta un ejemplo para el caso del manejo de los profesores en el sector público, en el que el gobierno ofrece un esquema de incentivos para la implementación de un programa. Este esquema tiene la estructura del dilema del prisionero. Tanto en los juegos simultáneos como en los secuenciales, los resultados de equilibrio pueden ser distintos a los que predice la teoría convencional.

Palabras clave: teoría de juegos, juegos sicológicos, modelos basados en intensiones, reciprocidad.
Clasificación JEL: C700.

---

* Université de Toulouse 1, MPSE, GREMAQ. A previous version of this paper was presented as a DEA mémoire to the MPSE - Ecole Doctorale de Science Economique of the Université de Toulouse 1. I want to thank Paul Seabright and Emmanuelle Auriol for their comments.
E-mail: darwin.cortes@univ-tlse1.fr

I. INTRODUCTION

For several years, other social sciences different from economics, including psychology, sociology and anthropology, have pointed out that human beings tend to engage in reciprocal behaviour. Until recently it had not been clear whether this behaviour was only caused by some expectations of future rewards or, at least in some cases, if it was genuine reciprocal behaviour.

If the first explanation was true, the usual economic hypothesis that individuals behave in a self-interested manner could explain those behaviours. Nevertheless, in the last two decades several experiments have shown that individuals exhibit authentic reciprocal behaviour in anonymous one-shot interactions. For example, in the ultimatum game a pair of individuals has to distribute a fixed sum of money in a sequential move game. The "proposer" has to divide the amount between himself and the second subject. The "responder" can accept or reject the proposed division. If individuals were rational and self-interested, the responder would accept any quantity of money and the proposer would give the smallest possible quantity. However, evidence shows offers lower than 20% are atypical and rejected with a high probability, while offers close to 50% are very common and rarely rejected (Fehr and Fischbacher, 2001).

On the other hand, in the gift-exchange game the proposer (employer) offers a wage to the responder (worker). The worker can either reject or accept it. If the worker rejects it both players gain nothing. If the worker accepts she has to exert a costly effort. The higher the effort, both the lower the payoff she gets and the higher income the employer receives. Under standard assumptions, the worker will always choose the lowest effort and the employer will only offer the lowest possible wage. Evidence suggests wages are clearly higher than minimum levels and wages and effort have a positive relation (Fehr and Fischbacher, 2001).

Those and other experiments have shown individuals actually reciprocate each other. A reciprocal individual rewards kind behaviour and punishes unkind behaviour. The gift-giving game illustrates the former, sometimes called positive reciprocity, and the ultimatum game the latter (negative reciprocity). Additionally, it has been shown that reciprocity can have an important role in some economic fields. In labour economics, questionnaire studies have shown managers are unwilling to cut wages because it can adversely affect work morale. Effectively, wage cuts are considered as an insult by the workers (Bewley, 1995). Besides, Akerlof (1982) suggests reciprocal behaviour can explain why wages remain above the market clearing level. In fact, this is supported by some experiments that have shown how reciprocity contributes to the enforcement of contracts, as loyalty and trust are relevant in labour relationships. Further experiments show individuals punish free-riders in public good provision games even if it reduces their own payoffs; material incentives may crowd-out implicit incentives that rely on reciprocal behaviour and reciprocity can explain why in reality contracts are incomplete, among other facts.[1] All these phenomena cannot be explained assuming the self-interest hypothesis.

---

[1] Fehr and Gächter (2000) survey experimental evidence. Frey (2001) also surveys circumstantial and econometric evidence.

There have been several attempts to model reciprocal behaviour. In this document I review the so-called intention-based models of reciprocity, particularly the models proposed by Rabin (1993) and Dufwenberg and Kirchsteiger (2001). This approach emphasizes in the fact that reciprocal individuals want to reward kind intentions and to punish unkind intentions. To illustrate these theories I propose two examples in teacher management. The first one consists in a game that models teachers' strategic behaviour in the following situation: the government wants to improve the quality of public education for which it intends to implement a program to improve teacher abilities. The government offers an incentive scheme that has a prisoner's dilemma structure to enforce the program; in such a way that standard game theory will predict both teachers are going to participate. The second game slightly modifies the material payoffs of the first one. I obtain that, in both simultaneous and sequential games, reciprocal teachers may deviate from participation in equilibrium, as they consider participation as an unkind behaviour. Instead, no participation is regarded as kind behaviour. Of course, participation of both teachers may also be in equilibrium when each teacher believes the other is going to participate. In that case, both teachers punish the other's unkind intention.

The text is organized in three sections. In section one, I provide an overview of the economic theories about reciprocity in order to give context to the intention-based theories. Section two is divided in several subsections in which I present the examples and the theories mentioned. With expositive purposes I first introduce the example and show the results obtained using the standard theory, and then I provide the model of reciprocity and the new results. The last section offers conclusions.

II. MODELLING RECIPROCITY

In standard theory, the self-interest hypothesis is formalized by defining individual preferences solely based on the material resources the individual has. One way to model reciprocal behaviour is to enlarge the space in which individual preferences are defined to include others' material payoffs or welfare. "When an individual does not only care about the material resources allocated to her but also cares about the relevant reference agents", we will say she has *social preferences* (Fehr and Fischbacher, 2001 p. 2).

In fact, most of the theories that try to model reciprocity introduce it as a social preference. These theories have taken into account that reciprocity has two elements in its nature: it is not only related to the consequences of others' actions but also to the others' intentions. They have focused on one of those elements of reciprocal behaviour. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) stress the fact that people desire to maintain equity and provide models of inequity aversion. On the other hand, Rabin (1993) and Dufwenberg and Kirchsteiger (2001) emphasize that persons want to punish nasty intentions and to reward friendly intentions. Levine (1998) builds a model in which individuals do not respond to intentions but to the type of person they face. The type is determined by the degree of altruism the individual has. Charness and Rabin (2000) and Falk and Fischbacher (2000) develop theories that have elements from both intention-based reciprocity and inequity aversion models. Finally, Segal and

Sobel (1999) present an axiomatic treatment of reciprocity and altruism which is compatible with some of the social preferences models of reciprocal behaviour.

It is worth pointing out that some inequity aversion models, which are only concerned with payoffs distribution, can mimic some predictions of intention-based reciprocity models. However, although intention-regarding models can be much more difficult to handle than inequity aversion models, experimental evidence suggests that people punish others even if punishment does not reduce inequity.[2] In the following section I present the pure intention-based economic models of reciprocity.

III. MODELS OF INTENTION-BASED RECIPROCITY

These models take into account not only the individuals' material payoff but also their beliefs about others' kindness when deciding what action they are going to follow. Specifically, individual utility is composed of two parts: a *material payoff*, which is given in terms of some measurable quantity, e.g. money; and a *reciprocity payoff* that she obtains from assessing the others' kindness. So, individuals will engage in the action that gives them the highest utility regarding both payoffs.

For example, consider the game in Figure 1. It presents a prisoner's dilemma. As is usual when individuals only care about their own material payoff, the Nash equilibrium is no cooperation for both persons. However, notice that when an individual chooses no cooperation instead of cooperation she is reducing the other's material payoff. So when one of the agents decides to cooperate, it can be interpreted by the other as a kind action, since the former reduces his payoff and increases the latter's at the same time. If both players have high enough sensitivity to reciprocity concerns, cooperation can be the best option for them.

Figure 1

		Player 2	
		C	*NC*
Player 1	*C*	0.1, 0.1	-0.1, 0.2
	NC	0.2, -0.1	0, 0

It is worth pointing out that beliefs on kindness are formed when assessing the other's intentions. If player 1's action increases her payoff and player 2's payoff simultaneously, player 2 will probably not consider that action as kind. Furthermore, it is possible that even if one player "sacrifices" her material payoff she will be considered

[2] For a complete discussion in this regard, look at Falk and Fischbacher (2000).

as not kind. For instance, in the game depicted in Figure 1, assume player 2 has no option different from cooperation. Somehow this player is forced to cooperate. So, we have a degenerate game composed by the left column of the game. In this case, player 1 will not believe player 2 is being kind by cooperating, as the latter has no choice.

To illustrate the theories considered in this document, we are going to analyze a qualitative example from teacher management. The next section poses the basic problem.

1. A Qualitative Example of Teacher Management

Assume that a government uses two teachers to offer public education. There is a teacher trade union so that if both take the same decision with respect to government policies, the government cannot punish them. Assume as well that the government wants to improve the quality of education offered and hence decides to implement a program that raises teacher abilities.

In order to implement that policy, the government brings out an incentive scheme as follows: If neither teacher participates in the program, the government cannot fire them and they continue gaining the same payoff as before, say X. If both teachers enter the program, they engage in a higher effort and obtain the same payment X.[3] Payoffs cannot be lower than X because otherwise the trade union would impede implementation of the governmental program.[4] Finally, if teachers take different decisions, the trade union is not working properly any more, so the teacher who does not participate is fired and obtains his reservation utility while the teacher who participates receives a payoff $X + \delta$ higher than X.

It is also assumed that teachers make their decisions simultaneously. The game is depicted in Figure 2. It is easy to see that the incentive scheme has a prisoner's dilemma structure. The government persuades teachers to participate by offering a contingent reward δ to deter trade union obstructions. Thus, players have an incentive to participate in the program independently from the other's choice. In such a model, if teachers only care about their material payoff, the unique Nash Equilibrium in pure strategies is (*Participate, participate*).

Figure 2

		Teacher 2	
		np	*p*
Teacher 1	*NP*	X, X	$0, X + \delta$
	P	$X + \delta, 0$	X, X

[3] Moral hazard is not an issue here but it should be in a more realistic model.

[4] In fact, the game structure is preserved even if participation payoffs are higher than X. It would be enough to assume the participation premium to be lesser than δ.

2. Introducing Reciprocity

Suppose both teachers appreciate kind behaviour, so they draw utility from reciprocity concerns. Notice that in this example, as in the first one, when a player attempts to maximize her material payoff she reduces the other's payoff. As teachers are reciprocal, they will reward friendly actions and will punish hostile actions. Assume teacher 2 has chosen to participate, so she can obtain either $X + \delta$ or X. If teacher 1 chooses to participate as well, he not only minimizes teacher 2's payoff (she would obtain X instead of $X + \delta$) but also maximizes his (he would get X instead of 0). Thus, this action could be considered unkind by teacher 2, hence she would not be willing to deviate from participation because otherwise she would reward teacher 1.

Now suppose teacher 1 chooses not to participate, so teacher 2 gets $X + \delta$ instead of X. In this case, teacher 2 perceives teacher 1 is giving up X for giving her δ and hence she could believe teacher 1's action to be kind. In this situation, teacher 2 would be unkind to player 1 if she remains participating. So as teacher 2 is reciprocal she could change her decision (from participation to non-participation) if she is better off doing so.

Notice that one player's assessment of the other's kindness depends not only on what the former believes the latter is going to do, but also on what the former believes the latter believes the former is going to do. To form both beliefs, fairness of intentions is determined assessing the equitability of the final payoffs' distribution with regard to the feasible set of payoffs. By doing so, each player will compare the utility she gets in both situations: participation brings her a higher material payoff than non-participation. Instead, non- participation brings her a higher reciprocity payoff than participation. So, if her reciprocity sensitivity is high enough, teacher 2 will decide to give up δ of her own payoff for giving teacher 1 X. Performing the same analysis for the other teacher, we obtain that with reciprocal teachers we have two possible equilibria: (*Not Participate, not participate*) and (*Participate, participate*).[5]

But, when will (*Not Participate, not participate*) be chosen? It depends on both the notion of fairness and reciprocity sensitivity players have, and the amount of the material payoffs. To see this one needs to introduce a formal model of reciprocity.

3. The Rabin (1993) model

Rabin (1993) models reciprocity based on psychological games proposed by Geanakoplos, Pearce and Stacchetti (1989) (hereafter GPS). In such games, players' payoffs depend not only on players' actions but also on their beliefs. GPS show that many standard concepts have useful analogues in the framework they develop.

Rabin's goal is to derive psychological games from "material" games. Let us consider a normal form game with two players, player 1 and player 2, who have mixed

[5] In the next section, we will call them fairness equilibria.

strategy sets A_1 and A_2, respectively, obtained from pure finite strategy sets S_1 and S_2. Player *i*'s material payoff is given by the function $\pi_i: A_1 \text{ x } A_2 \rightarrow \Re$.

In order to construct the psychological game, let us assume that when a player chooses her strategy, her subjective utility function will depend on three things: her strategy, her belief about the other's strategy and her belief about the other's belief about her strategy.[6] Let us call $a_1 \in A_1$ and $a_2 \in A_2$ the strategies of player 1 and player 2, respectively; $b_1 \in A_1$ and $b_2 \in A_2$ player 2's belief about player 1's strategy and player 1's belief about player 2's strategy, respectively; and $c_1 \in A_1$ and $c_2 \in A_2$ player 1's belief about player 2's belief about player 1's strategy and player 2's belief about player 1's belief about player 2's strategy. Observe that although a_i, b_i, and c_i belong to the same set, they are different in nature as a_i is a player *i*'s decision, b_i is player *j*'s belief ($j \neq i$) and c_i is a player *i*'s belief.

To incorporate reciprocity (fairness in terms of Rabin) in the model we first need to define a kindness function $f_i(a_i, b_j)$ which measures how kind player *i* is to player *j*. If player *i* believes player *j* chooses b_j, how kind is player *i* by choosing a_i? When player *i* chooses a_i, he is selecting a payoff pair $(\pi_i(a_i, b_j), \pi_j(b_j, a_i))$ from the set of all the feasible payoffs to player *j* when he chooses b_j. Let us call this set $\Pi(b_j) = \{(\pi_i(a, b_j), \pi_j(b_j, a)) | a \in A_i\}$.

How kind player *i* is being depends on both the point she chooses from $\Pi(b_j)$ and on the notion of kindness players have. To express this notion in formal terms, we need to define a function for both player i's kindness to player j and player *i*'s belief about how kind player *j* is being to her. Rabin (1993) provides some general properties that sort of functions must have. The following payoffs are useful to do that: let $\pi_j^h(b_j)$ be player *j*'s highest payoff in $\Pi(b_j)$, $\pi_j^l(b_j)$ be player *j*'s lowest payoff among the Pareto-efficient points in $\Pi(b_j)$, and $\pi_j^e(b_j)$ be an "equitable payoff" in $\Pi(b_j)$.

The following properties for kindness functions are sufficient conditions for the main result Rabin obtains:[7]

Property 1: A kindness function must be bounded and increasing. A kindness function $f_i(a_i, b_j)$, is bounded and increasing if:

---

[6] Higher order beliefs can be considered but it is enough to take the first two.

[7] They are presented as definitions in Appendix A in Rabin (1993), p. 1297. For additional results one also needs to assume the kindness function to be affine, but that property is not relevant for our present purposes.

a) There exists a number *N* such that $f_i(a_i, b_j) \in [-N, N]$ for all $a \in A_j$ and;

b) $f_i(a_i, b_j) > f_i(a'_i, b_j)$ if and only if $\pi_j(b_j, a_i) > \pi_j(b_j, a'_i)$.

This property rules out the possibility of fairness to generate infinitely positive or infinitely negative utility and brings out a positive association between the player *j*'s payoff and player *i*'s kindness: given b_j, the higher player *j*'s payoff is, the kinder player *i* is.

Property 2: A kindness function must be a Pareto split. A kindness function $f_i(a_i, b_j)$ is a Pareto split if there exists some $\pi_j^e(b_j)$ such that:

a) $\pi_j(b_j, a_i) > \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) > 0$; $\pi_j(b_j, a_i) = \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) = 0$; and $\pi_j(b_j, a_i) < \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) < 0$;

b) $\pi_j^h(b_j) \geq \pi_j^e(b_j) \geq \pi_j^l(b_j)$; and

c) if $\pi_j^h(b_j) > \pi_j^l(b_j)$, then $\pi_j^h(b_j) > \pi_j^e(b_j) > \pi_j^l(b_j)$.

This property says that the *fair* payoff to player *j* is strictly between the best and the worst Pareto efficient payoffs in $\Pi(b_j)$, provided that the Pareto efficient set is not a singleton.

Among the class of functions defined by the previous properties, Rabin picks the following:

Definition 1: Player *i*'s kindness to player *j* is given by

$$f_1(a_i, b_j) \equiv \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{\min}(b_j)}$$

where $\pi_j^{\min}(b_j)$ is the worst possible payoff for player *j* in $\Pi(b_j)$ and $\pi_j^e(b_j)\frac{\pi_j^h(b_j) + \pi_j^l(b_j)}{2}$. If $\pi_j^h(b_j) - \pi_j^{\min}(b_j) = 0$ then $f_i(a_i, b_j) = 0$.

It is easy to check that this function has the general properties presented above: First, $f_1 = 0$ if and only if player *j* receives the equitable payoff. This is so because when $\pi_j^h(b_j) = \pi_j^{\min}(b_j)$ player *j* always gains the same payoff and there is no kindness

issue. Second, $f_1 < 0$ when player *j*'s payoff is lesser than the equitable payoff. This happens either when $\pi_j(b_j, a_i)$ is a Pareto-efficient point smaller than the equitable payoff or when $\pi_j(b_j, a_i)$ is not an efficient point. Finally, $f_i > 0$ only if player *j*' s payoff is greater his equitable payoff and the Pareto set is not singleton. Notice also that the functions take values in the interval [-1, ½].

Player *i*'s belief about how nice player *j* is to her, can be expressed as a function $\tilde{f}_j(b_j, c_i)$. This function is formally equal to the previous one but it relates the two levels of beliefs considered in the model.

Definition 2: Player *i*'s belief about how kind player *j* is to her is given by

$$\tilde{f}_j(b_j, c_i) = \frac{\pi_i(c_i, b_j) - \pi_i^e(c_i)}{\pi_i^h(c_i) - \pi_i^{\min}(c_i)}$$

where $\pi_i^{\min}(c_i)$ and $\pi_i^e(c_i)$ have analogue definitions. If $\pi_i^h(c_i) - \pi_i^{\min}(c_1) = 0$ then $\tilde{f}_j(b_j, c_i) = 0$.

Using both functions $f_j(a_i, b_j)$ and $\tilde{f}_j(b_j, c_i)$ we can define a utility function for player *i*. Doing so, we are assuming players have a shared notion of fairness. This utility function integrates the material payoff and the reciprocity payoff:[8]

$$U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + Y_i \tilde{f}_j(b_j, c_i)\left[1 + f_i(a_i, b_j)\right]$$

The first term is the material payoff and the second one is the reciprocity payoff. The constant Y_i reflects how sensitive player *i* is to reciprocity matters regarding player *j* and we will assume it is positive. This utility function gathers the main feature about reciprocal behaviour. If player *i* believes player *j* is treating her unkindly $\left(\tilde{f}_j(b_j, c_i) < 0\right)$, she will want to punish him for being unkind, that is choosing a_i such that $f_j(a_i, b_j)$ is low. On the contrary, if player *i* thinks player *j* is being nice $\tilde{f}_j(b_j, c_i) = 0$, she will be nice. Furthermore, the higher $\tilde{f}_j(b_j, c_i)$ is, the more material payoff player *i* is willing to give up to reward player *j*. Finally, this utility function has the property that whenever

---

[8] This utility function is slightly different from which Rabin uses. We have added the term Y_i in the reciprocity payoff.

player *j* is hostile to player *i*, player *i*'s utility is lesser than her material payoff. That is, an individual is not able to completely recover her welfare by taking revenge once other has treated her badly.

These preferences together with the elements already defined for the material game form a psychological game. Using the concept of psychological Nash Equilibrium defined by GPS, Rabin (1993) proposes the following definition,

Definition 3: The pair of strategies $(a_1, a_2) \in (A_1, A_2)$ is a *fairness equilibrium* if, for $i = 1,2,\ j \neq i$

a) $a_i \in \arg\max_{a \in A_i} U_i(a, b_j, c_i)$

b) $c_i = b_i = a_i$

This notion of equilibrium is analogous to Nash Equilibrium, but applied to psychological games. Condition b) of the definition requires all high-level beliefs to correspond actual behaviour.

Considering again our example, we can calculate the teachers' utility functions regarding reciprocity. Although the theory is posed for mixed strategies we only analyze equilibriums in pure strategies. In Figure 3, we can see the utility values once Condition b) of the fairness equilibrium is satisfied.

Figure 3

		Teacher 2	
		np	*p*
Teacher 1	*NP*	$X + \frac{3}{4} Y_1 \ , X + \frac{3}{4} Y_2$	$-\frac{3}{4} Y_1 \ , \ X + \delta + \frac{1}{4} Y_2$
	P	$X + \delta + \frac{1}{4} Y_1 \ , \ -\frac{3}{4} Y_2$	$X - \frac{1}{4} Y_1 \ , X - \frac{1}{4} Y_2$

First, note that when player *i* is being unkind to player *j*, player *j*'s reciprocity payoff is negative, which reduces his overall utility and introduces incentives to deviate. However, the profile of strategies (*Participate, participate*) is a fairness equilibrium for all values of *X* and Y_i because the response for unkindness is unkindness. Consider now, what happens if player *i* deviates to non-participation. This action increases player *j*'s reciprocity payoff because he considers player *i* is being kind. In fact, player *j* will deviate to a non-participation strategy if the loss in material payoff, d, is less than the gain in reciprocity payoff, $\frac{1}{2} Y_j$. The profile (*Not Participate, not participate*) will be a fairness equilibrium whenever $\delta < \frac{1}{2} Y_i$ for $i = 1,2$. This condition is satisfied when

either δ is low enough or Y_i is high enough. If the government gives too little reward when one teacher participates and the other does not or if both teachers have a strong feeling to reciprocate the other, deviating from participation will be an equilibrium.

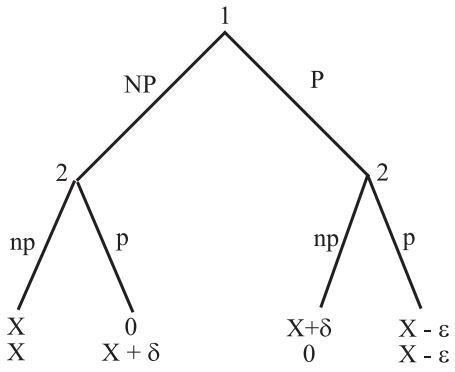
This model has been extended to include sequential actions. In principle, such a model would be more adequate to reality because reciprocal actions have an implicit delay. One is kind to somebody who *has been* kind. Besides, "extending the model to sequential games is also essential for applied research" (Rabin, 1993; p. 1296), as individuals can change their motivation due to information provided by past decisions.

4. Sequential Games

Consider now a slightly modified sequential version of our game of teachers, depicted in Figure 4. Assume there is no trade union anymore, so the government can offer a lesser material payoff if both teachers participate in the program, $X - \varepsilon$, $0 < \varepsilon < X$. In the first step, teacher 1 decides whether to participate or not in the program offered and once he has decided, teacher 2 has to take her decision. Assuming no further reciprocity (and perfect and complete information), one can see, solving by backward induction, that the profile (*Participate, participate*) will be the unique Subgame Perfect Nash Equilibrium. The government's strategy to implement the program is completely successful as teachers will always participate.

Let us introduce reciprocity. Suppose teacher 1 chooses not to participate (*NP*) in the program. Teacher 2 can choose either X or $X + \delta$ (or mix). Her choice will depend on both her kindness and the belief she has about teacher 1's intention to choose *NP*. When teacher 1 chooses *NP*, he gives teacher 2 a payoff of at least X and at most $X+\delta$. Instead, when teacher 1 chooses *P* he gives teacher 2 a payoff of at least 0 and at most $X - \varepsilon$. So teacher 2 will believe teacher 1 is being kind when he chooses *NP* and if the reciprocity payoff is high enough she will choose X instead of $X+\delta$ (or mix). To establish if (*NP*, *np*) will be an equilibrium, we have to evaluate what teacher 1 believes when teacher 2 chooses *np*.

Figure 4



It is convenient to point out one difference in the analysis of reciprocity in normal games and extensive games. In normal games teacher 2 will always choose do not participate, provided the reciprocity payoff supersedes material payoff. This does not happen in a sequential model because, for instance, once teacher 2 knows teacher 1 has chosen to participate, there is no reason to maintain the decision of not participating unconditionally. In that case teacher 2 would consider teacher 1 is being hostile and thus she would participate in the program as well. Unlike normal games, in sequential games unconditional *np* does not occur because player 2 is optimizing in each subgame.

On the other hand, in modelling reciprocity in sequential games it is not plausible to assume that players are going to keep their initial beliefs along the game. Player 2's belief about how kind player 1 is being once the latter has decided not to participate is different from the former's belief once the latter has decided to participate. It means it is necessary to analyze changes in beliefs in each node of the game in order to establish equilibrium conditions. Furthermore, it is not possible to consider each subgame separately. Player 2's belief about how nice player 1 is, given he has already decided not to participate, depends on which payoffs she would have had if player 1 had decided to participate. Therefore, backward induction cannot be used to obtain the equilibrium. Dufwenberg and Kirchsteiger (2001) (henceforth DK (2001)) provide a concept of sequential reciprocity which allows them to propose a new solution concept, Sequential Reciprocity Equilibrium (SRE).

5. The Dufwenberg and Kirchsteiger (2001) model

As we have said, when reciprocity is incorporated in sequential games it is necessary to distinguish between a player' initial and subsequent belief. Once a subgame has been attained, a player's belief can change and, as kindness depends on belief, kindness may therefore change as well. DK (2001) deal with this by keeping track of how beliefs change when a new subgame is reached and by assuming players' choices take into account the beliefs they hold in the most recently reached subgame. To do that, just like Rabin, they adopt the psychological games framework; but unlike GPS (1989), who limits his discussion to games where only initial beliefs can affect player assessments, DK (2001) proposes a notion of reciprocity in which player beliefs change in each subgame.

Formally, they pose a *t*-player extensive form game without nature and with perfect recall. Any such game Γ is described by a finite set of nodes organized in a tree, a collection of information sets, a set of choices available at each decision node, a function assigning each information set to the player who moves at the decision nodes in that set, and a collection of payoff functions assigned to each endnode (Mas-Colell, Whinston and Green (1995)). Let $T = \{1,...,t\}$ be the set of players where $t \geq 2$. It is convenient to add new notation to that used in section 3, as there are now several players. Let A_1 be the set of player *i*'s behaviour strategies, a_i; B_{ij} be the set of possible player *i*'s beliefs about player *j*'s strategies, b_{ij}; and C_{ijk} be the set of player *i*'s belief about player *j*'s belief about player *k*'s strategies, C_{ijk}. As in Rabin's model, beliefs are

mixed strategies, so we have B_{ij} and $C_{ijk} = B_{jk}$. Besides, player *i*'s material payoff is now given by the function $\pi_i : A \to \Re$ where $A = \Pi_{i \in T} A_i$.

Now, let us proceed to formalize how the player's beliefs change when new subgames are reached. To keep track of how each player's behaviour, niceness and perception of other's niceness differ across subgames, let *R* be the set of nodes that are starting nodes of all possible subgames in Γ, and let Γ^r be the subgame whose starting point is $r \in R$. Let us define the *r*-part of Γ^r as the set of nodes in Γ^r that do not belong to some proper subgame of Γ^r. For a strategy $a_i \in A_i$, let $a_i(r)$ be the strategy that has the same choices as a_i but assigning a probability equal to 1 to the choices that lead to node *r*. In an analogous way, define $b_{ij}(r)$ and $c_{ijk}(r)$ for $b_{ij} \in B_{ij}$ and $c_{ijk} \in C_{ijk}$, respectively. Thus, player *i* decides to play a_i believing other players are playing $(b_{ij})_{i \neq j}$ and believing $(c_{ijk})_{j \neq k}$, whereas in the *r*-part of the subgame Γ^r, player *i* is playing $a_i(r) \in A_i$ and believing other players will play $(b_{ij}(r))_{i \neq j}$ and believe $(c_{ijk}(r))_{j \neq k}$. This means that "even if players initially believe that others mix their choices, the subsequent perception of kindness is triggered by the actual choice" (DK (2001), p. 8). In terms of our example, consider the proper subgame starting in the player 2's right side node and call that node *r*. Player 2 believes player 1 is choosing his strategy as $b_{21} = p'NP + (1 - p')P$. Before 1 plays, *at node r* if *p*' is big (1 or near 1) player 2 will think player 1 is kind. However, once *r* is reached, player 2 does not consider player 1 to be kind anymore. At *r*, player 2's belief is $b_{21}(r) = P$.

DK (2001) also change the notion of efficiency suggested by Rabin (1993), which says that the lowest efficient strategy is chosen from $\Pi(b_j)$. They argue that in a sequential game framework, the set of Pareto-efficient strategies relevant to establish the equitable payoff cannot depend on beliefs, as this can drive us to non-existence of equilibrium.[9] DK (2001)'s efficiency notion can be formulated as

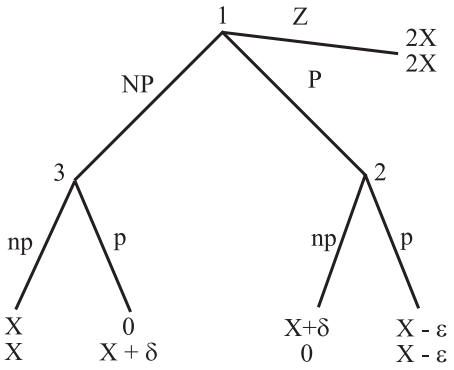
$E_i = \{a_i \in A_i |$ there exists no $a_i' \in A_i$ such that for all $r \in R(a_j)_{j \neq i} \in \Pi_{j \neq i} A_j$, $k \in T$ it holds that $\pi_k(a_i'(r), (a_j(r))_{j \neq i}) \geq \pi_k(a_i(r), (a_j(r))_{j \neq i})$, with strict inequality for some $(r(a_j)_{j \neq i})\}$

The concept of efficiency has a central role in intention-based theories. To illustrate this point consider the game depicted in Figure 5. We have the same game of Figure 4

[9] Look at DK (2001) p. 29 for an example of no existence of equilibria due to a belief dependent concept of efficiency.

but now player 1 can engage in an action Z in which both players obtain a payoff -2*X*. Let us suppose player 1 believes with probability one that player 2 is playing the strategy (*np*, *p*). It can be seen that player 1 believes he selects the material payoff $\pi_2(P, np, p) = X - \varepsilon$ from the feasible set $\{X, X-\varepsilon, -2X\}$. In the game of Figure 4, player 1 would be considered unkind. Now are we willing to accept player 1 is being kind due to the mere possibility of Z being chosen? To rule out this unreasonable consideration we restrict our attention to efficient payoffs in order to determine the equitable payoff. DK (2001) propose the notion of efficiency above to do that.[10]

Figure 5



We can define the equitable payoff as $\pi_j^{e_i}\left(\left(bi_j\right)_{j\neq i}\right) = \frac{1}{2}\left(\pi_j^h\left(a_i\left(b_{ij}\right)_{j\neq i}\right) + \pi_j^l\left(a_i, \left(bi_j\right)_{j\neq i}\right)\right)$, which is essentially the same defined in section 3. Unlike that one, a subindex *i* has been added to *e* to indicate that this is the equitable payoff for *i* and $\pi_j^l\left(a_i\left(bi_j\right)_{j\neq i}\right)$ is now the lowest payoff in E_i.

In turn, kindness, belief in kindness and utility functions can also be defined in a similar way as before.

Definition 4: Kindness of player *i* to another player $j \neq i$ in the *r*-part of a subgame Γ^r is given by the function $f_{ij} : A_i \times \Pi_{j\neq i} Bi_j \to \Re$ defined by

$$f_{ij}\left(a_i(r), \left(b_{ij}(r)\right)_{j\neq i}\right) = \pi_j\left(a_i(r), \left(bi_j(r)\right)_{j\neq i}\right) - \pi_j^{e_i}\left(\left(b_{ij}\right)_{j\neq i}\right)$$

[10] However, this distinction does not make any difference with respect to our example, because all the strategies are efficient under both concepts.

Apart from differences already mentioned, Definition 4 is analogous to Definition 1. f_{ij} differs formally from f_i in that f_{ij} is not normalized and thus, in principle, it may take extremely high or low values. However, due to the fact that we are analysing central points (as we subtract an average from the chosen payoff), we cannot expect to obtain an extreme number, so Property 1a in section 3 can be held without large inconveniences. On the other hand, it is straightforward to check that Definition 4 holds for Properties 2 and 1b.

Definition 5: Player *i*'s beliefs about how kind player $j \neq i$ is to *i* in the *r*-part of a subgame Γ^r is given by the function $\tilde{f}_{iji} : B_{ij} \text{ x } \Pi_{k \neq j} C_{ijk} \rightarrow \Re$ defined by

$$\tilde{f}_{iji}\left(bi_j(r), \left(c_{ijk}(r)\right)_{k \neq i}\right) = \pi_i\left(bi_j(r), \left(c_{ijk}(r)\right)_{k \neq j}\right) - \pi_i^{e_j}\left(\left(c_{ijk}(r)\right)_{k \neq j}\right)$$

This definition is formally equal to the previous one. The same comments for f_{ij} with respect to f_i can be done for f_{iji} in relation to $\tilde{f}_j$.

Definition 6: Player *i*'s utility in the *r*-part of a subgame Γr is a function

$U_i : A_i \text{ x } \Pi_{j \neq i}\left(B_{ij} \text{ x } \Pi_{k \neq j} C_{ijk}\right) \rightarrow \Re$ defined by

$$U_i\left(ai_j(r), \left(c_{ijk}(r)_{k \neq j}\right)_{j \neq i}\right) =$$

$$\pi_i\left(a_i(r), b_{ij}(r)_{j \neq i}\right) + \sum_{j \in T/\{i\}} \left(Y_{ij}\, f_{ij}\left(a_i(r), \left(b_{ij}(r)\right)_{j \neq i}\right) \tilde{f}_{iji}\left(b_i j(r), \left(c_{ijk}(r)\right)_{k \neq j}\right)\right)$$

The Utility function in Rabin's model has the term $(1 + f_1)$ instead of f_{ij}. For comparative purposes, it has been preferred to keep the functions as alike as those the authors propose. As sensitivity for reciprocity Y_{ij} is a nonnegative number, reciprocity payoff increases utility if player i believes player *j* is kind $\left(\tilde{f}_{iji} > 0\right)$, and reduces utility if player *i* believes player *j* is unkind $\left(\tilde{f}_{iji} < 0\right)$.

Appending this kind of utility functions to an extensive game, we get the tuple $\Gamma^o = \left(\Gamma, (U_i)_{i \in T}\right)$. DK (2001) call Γ^o a psychological game with reciprocity incentives. There is a notion of equilibrium associated to these games that can be formulated as

Definition 7: The profile $\hat{a} = (\hat{a}_i)_{i \in T}$ is a *Sequential Reciprocity Equilibrium* (SRE) if for all $i \in T$ and for all $r \in R$ it holds that

a) $\hat{a}_i(r) \in \arg\max_{ai \in Ai(r,\hat{a})} U_i\left(a_i, b_{ij}(r), \left(c_{ijk}(r)_{k\neq j}\right)_{j\neq i}\right)$

b) $b_{ij} = \hat{a}_j$ for all $_{j\neq i}$

c) $c_{ijk} = \hat{a}_k$ for all $_{j\neq i, k\neq j}$

where $A_i(r,a)$ is the set of strategies player *i* can use if she behaves according to $a_i(r)$ at information sets outside the *r*-part of Γ^r, but is free to choose any alternative in the *r*-part of Γ^r.

Condition a) says player *i* maximizes his utility at node *r* given his beliefs and given that he follows his equilibrium strategy outside the *r*-part of Γ^r. This entails beliefs to assign a probability of one to the sequence of choices that allow *r* to be reached. Conditions b) and c) say that in the equilibrium, beliefs are correct and correspond to the actual strategy. DK show that every psychological game with reciprocity incentives has at least one SRE. To do that, they first define the size of a subgame as the number of its subgames, then they simultaneously determine equilibrium choices of the subgames with the same size, starting from the smallest (size equal one) until arriving to the complete game.[11]

In the game that appears in section 4, first teacher 1 decides whether to participate or not in the program offered by the government and then teacher 2 does so. As we showed there, non- reciprocity implies a profile of (*participate, participate*) is the sole Subgame Perfect Nash Equilibrium. How is the analysis affected when teachers are reciprocal? We can find it out using the theory developed in this section. When there is reciprocity between agents, the game becomes a psychological game with reciprocity incentives. Examining SRE for different levels of reciprocity sensitivity we can say:[12]

1. If teacher 2's sensitivity to reciprocity, Y_2, is low enough, the profile (participate, participate) is an equilibrium behaviour. Specifically this occurs when $Y_2 < \frac{2\delta}{x(\delta+\varepsilon)}$.

   In this case, each player will believe the other is going to participate, which will in turn be considered as unkind. Those beliefs render a negative reciprocity payoff to both players and therefore each teacher prefers to participate in the government program. From the previous inequality it can also be seen that given a sensitivity to reciprocate level for both teachers, Y_1 and Y_2, the higher δ relative to *X* and ε is, the more likely both teachers will participate. The Government should take this into account in order to make teachers participate in the proposed program.

2. If teacher 2's inclination to reciprocate, Y_2, is high enough, profile (*do Not Participate, do not participate*) holds in all SRE. Regardless of Y_1, when teacher 2 has a

---

11 Demonstration appears in DK (2001) p. 35.

12 Detailed calculations are included in the Appendix. As the game has two players, we simplify notation, so Y_i, i = 1,2, is agent *i*'s sensitivity to reciprocity.

strong inclination to reciprocate, she will obtain a high reciprocity payoff if teacher 1 decides not to participate, so she will play *np* (instead of *p*) when teacher 1 does so. Notice player 2 would also get a higher material payoff doing so than that she would have obtained if teacher 1 played *P* (instead of *NP*). Teacher 1 knows all of this, and thus he will choose to play *NP* to get a higher material payoff than that he would get if he had played *P*. This equilibrium behaviour cannot be predicted when we assume non-reciprocity. The scheme proposed by the government does not work in the way the government expects due to reciprocity between teachers.

3. Given teacher 2's strong leaning towards reciprocating, this leads (*Participate*, *participate*) to be an equilibrium behaviour when teacher 1 also has a strong tendency to reciprocate. This arises when each player thinks the other one is going to play *p*, as each one expects an unkind action from the other. There are "self-fulfilling expectations".

4. For intermediate values of Y_1 and Y_2, equilibrium behaviours are mixed strategies. In equilibrium, probability of no participation, *p*, for player 2 is given by $p = 1 + \frac{\varepsilon}{\delta} - \frac{2}{Y_2 x}$. As it can be inferred from the previous analysis, this probability increases when Y_2 increases. In addition, *p* decreases if the ratio between ε and δ decreases and increases if *X* increases. $\frac{\varepsilon}{\delta}$ can be viewed as the inverse of the incentive the government provides player 2 to participate. Player 2 tries to gain δ (she gains δ if (*NP, p*) is chosen) but she loses ε if (*P, p*) is chosen. She evaluates how much she can obtain and lose from participation. This evaluation affects *p* in the way described. On the other hand, an increase in *x* increases *p* because *ceteris paribus* it makes less attractive to participate. For player 1, it is not possible to do the same kind of analysis due to parameters affecting his probability of non-participation, *q*, in a complex way. In fact, for a given Y_2, player 2's equilibrium behaviour is unique whereas, in general, 1's equilibrium behaviour is not unique for a given Y_1.[13]

Finally, from the results obtained for this game we can analyze a sequential version of the teacher's game with trade union. In that game $\varepsilon = 0$, so payoffs in profiles (*NP, np*) and (*p, p*) are equal to *X*. The most interesting result in this case is that non-participation is an equilibrium behaviour only in mixed strategies. The analysis is as follows. We know teacher 2 will always play *p* when teacher 1 plays *P*, so teacher 2 would get *X* in this profile.[14] On the other hand, if teacher 1 plays *NP*, teacher 2 can get either *X* or $X+\delta$. For (*NP, p*) to be possible in equilibrium, player 2 has to believe with probability one that player 1 believes player 2 will choose *p*. But in this situation, player 2 would obtain *X* from both (*NP, np*) and (*p, p*) and hence there would be no reciprocity issue (reciprocity payoff equal to zero). Therefore, player 2 would prefer to play another strategy, as

---

[13] See Remarks 4 and 5 in the Annex.

[14] See Remark 1 in the Annex.

profile (*NP, p*) offers player 2 a higher material payoff. A similar analysis can be done for player 1.

IV. CONCLUSIONS

Evidence has shown that sometimes people behave in different ways from what is predicted by assuming individuals are self-interested. Furthermore, when persons deviate from self-interested behaviour they do not always try to increase the well-being of others. On the contrary, it has been found that individuals usually respond in a kind manner to kind actions and in an unkind manner to unkind actions. In response to these findings, several economic theories have attempted to model reciprocity behaviour. In this document, we have reviewed the so-called intention-based theories of reciprocity, specifically the models made by Rabin (1993) and Dufwenberg and Kirchsteiger (2001).

These theories have received this name because they emphasize that people want to punish hostile intentions and reward nice intentions. To do that, they adopt the psychological games framework developed by Geanakoplos, Pearce and Stacchetti (1989). In this framework individual utility depends not only on strategies but also on beliefs. Rabin (1993) develops a theory for 2-players normal form games and introduces a new equilibrium notion called *fairness equilibrium*. Dufwenberg and Kirchsteiger (2001) in turn extend Rabin's theory dealing with t-players sequential games and present the notion of *sequential reciprocity equilibrium*. The main innovation they achieve is to keep track of beliefs about intentions as the game evolves. Players maximize their behaviour in each subgame taking into account beliefs about intentions formed in the previous stages. In a particular subgame, players use beliefs that come from the most recently reached subgame.

There are other differences between these models. Rabin (1993) uses a kindness function neutral to units of measure of the stakes, so that kindness cannot infinitely increase or decrease utilities. This also allows for individual kindness to decrease as long as payoffs become larger. Instead Dufwenberg and Kirchsteiger (2001) measure kindness in the same units of material payoffs (i.e. money), which has the advantage that kindness does not disappear when payoffs rise, but the disadvantage that it also makes utility sensitive to linear transformations as reciprocity payoff is measured in "money squared". Moreover, they differ in the efficiency notion used to define the equitable payoff. Rabin (1993)'s notion depends on beliefs and then it only considers strategies on the equilibrium path; whereas DK (2001) defines inefficient strategies as those that yield a weakly lower payoff for all player (strictly lower for some) than other alternatives in all the subgames. Finally, Rabin (1993) specifies kindness in the utility function in such a way to capture the idea that whenever a player is treated unkindly, her overall utility will be lower than her material payoff (her ability to take payback is not perfect). DK (2001)'s specification does not capture that.

We have also illustrated the theories studied with a simple example in teacher management. We have proposed an implementation mechanism for a governmental policy when there is a teachers' trade union. Both teachers have to decide to participate (*p*) or not (*np*) in a governmental program. In order to implement the policy, the government proposes a

game with a prisoner's dilemma structure. Without reciprocal teachers, in both games (normal and sequential forms) there is a unique equilibrium: teachers participate in the governmental program. With reciprocal teachers, we obtain additional results. In the normal form game, there are two fairness equilibria: one in which each one teacher is kind to the other and other in which both teachers are unkind. If in equilibrium both teachers are kind to each other, the government cannot implement the program.

In the sequential game, in turn, we have multiple equilibria. We considered two games: a sequential version of the previous one and a game in which there is no trade union and hence the government can give a lesser payoff if both teachers participate in the program. Now teacher 2 does not choose *np* unconditionally as in the normal form, since teacher 2 behaves optimally off the equilibrium path. In both games, conditional "cooperation" can be part of a SRE. However, under trade union *np* is an equilibrium behaviour only in mixed strategies.

One limitation of the intention-based approach is that one individual only has reciprocal behaviour when other individuals have demonstrated kind intentions. Suppose in our example player 2 is constrained to "choose" not to participate. Player 1 will not consider this action as kind because player 2 has no option. In fact, although nowadays there is almost consensus about the existence of reciprocal behaviour, there is still disagreement about the foundations of that behaviour. For instance, other theoretical approaches focus on inequity aversion (Bolton and Ockenfels, 2000) or the type of persons one faces (Levine, 1998). Hence, in an inequity model player 1 will behave kindly when there is an inequity issue even if player 2 is forced to choose not to participate. Discussion regarding this point is open.[15]

Another limitation of this approach is that equilibrium analysis is rather complex and there are multiple equilibria due to self-fulfilling beliefs. In the normal form game suggested, for example, both equilibria emerge for this reason, so it is not possible to establish which one is going to occur. On the other hand, even though treatment of beliefs in the sequential model is very innovative it makes it difficult to build tractable models.

Finally, despite the simplicity of our examples, they suggest it will be worth to take into account reciprocity in theories that try to model government - teachers' relationships. On one hand, a significant part of the literature on reciprocity has shown reciprocal behaviour is relevant in the analysis of employer-employee relationships. There is documentation on how employers are reluctant to decrease wages in crisis times because they do not want to reduce employee' morale to work. In particular, it would be interesting to find out how reciprocity affects the main results of multiagent settings.[16] On the other hand, some empirical research has shown teachers' trade unions can affect negatively student performance (quality of education) (Hoxby, 1996).

---

[15] Falk and Fischbacher (2000) show evidence that supports intentions matter. Fehr and Schmidt (2001) survey existing models on fairness and reciprocity

[16] One of the main results in these settings is that under moral hazard, principal can use relative performance of agents to elicit a higher effort (yardstick competition). Cf. Laffont and Martimort (2002).

V. REFERENCES

Akerlof, G. (1982), "Labor contracts as a partial gift exchange", *Quarterly Journal of Economics* 97, 543 - 569.

Benabou, R. and J. Tirole (2002), Intrinsic and extrinsic motivation, mimeo

Bewley, T. (1995), "A depressed labor market as explained by participants", *American Economic Review* 85, Papers and Proceedings, 250-254.

Bolton, G. and Ockenfels, A. (2000), "ERC - A theory of Equity, Reciprocity and Competition", *American Economic Review* 90, 166-193.

Charness, G. and Rabin, M. (2000), Social preferences: some simple tests and a new model, University of California at Berkeley, mimeo.

Dufwenberg, M. and Kirchsteiger, G. (2001), "A theory of sequential reciprocity", Discussion paper. CentER, Tilburg University.

Falk, A. and Fischbacher, U. (2000), "A theory of reciprocity. Institute for Empirical Research in Economics, University of Zurich", Working Paper 6.

Fehr, E. and Falk, A. (2001), Psychological Foundations of Incentives, University of Zurich, mimeo.

Fehr, E. and Fischbacher, U. (2001), Why social preferences matter - The impact of non-selfish motives on competition, cooperation and incentives, University of Zurich, mimeo.

Fehr, E. and Gächter, S. (2000), "Fairness and retaliation: The economics of reciprocity", *Journal of Economic Perspectives* 14, 159-181.

Fehr, E. and Schmidt, K. (2001), Theories of fairness and reciprocity - Evidence and economics applications, University of Zurich, mimeo.

Frey, B. and Jegen, R. (2001), "Motivation crowding-out theory", *Journal of Economic Surveys* 15 (5), 589- 611.

Geanakoplos, J.; Pearce, D. and Stacchetti, E. (1989), "Psychological games and sequential rationality", *Games and economic behaviour* 1, 60-79.

Hoxby, C. (1996), "How teachers' unions affect education production", *Quarterly Journal of Economics*, 671-718.

Laffont, J. and Martimort, D. (2002), *The theory of Incentives: The principal - agent model*, Harvard University Press.

Levine, D. (1998), "Modelling altruism and spitefulness in experiments", *Review of Economic Dynamics* 1, 593-622.

Mas-Colell, W. and Green (1995), *Microeconomic Theory*, Oxford University Press.

Rabin, M. (1993), "Incorporating fairness into game theory and economics", *American Economic Review*, 83 (5), 1281-1302.

Seabright, P. (2002), Blood, Bribes, and the crowding-out of altruism by financial incentives, University of Toulouse, mimeo.

Segal, U. and Sobel, J. (1999), Tit for tat: Foundations of preferences for reciprocity in strategic settings, University of California at San Diego, mimeo.

ANNEX. EQUILIBRIUM ANALYSIS OF THE SEQUENTIAL GAME

Remark 1: If teacher 1 participates, teacher 2 also participates in every SRE.

Note that only the reciprocity payoff can make 2 choose *np*, as the material payoff *per se* dictates a choice of *p* for 2. However, for any possible strategy of 2, teacher 2 gets less when 1 chooses *P* than when he chooses *NP*. Whatever 1 believes about 2's strategy, 1's choice of *P* is unkind, and hence 2 must believe that 1 is unkind. Thus the reciprocity payoff as well as the material payoff makes teacher 2 to choose *p*.

Remark 2: If teacher 1 does not participate, the following holds in all SRE:

a) If $Y_2 > \frac{2\delta}{\varepsilon x}$, teacher 2 does not participate

b) If $Y_2 < \frac{2\delta}{x(\delta+\varepsilon)}$, teacher 2 participates

c) If $\frac{2\delta}{x(\delta+\varepsilon)} < Y_2 < \frac{2\delta}{\varepsilon x}$, teacher 2 does not participate with a probability of

$$p = 1 + \frac{\varepsilon}{\delta} - \frac{2}{Y_2 x}$$

Notice that if player 1 does not participate, player 2 can give player 1 a material payoff of at least 0 and at most *x* so the "equitable" payoff of player 1 is *x*/2. If player 2 chooses no participation, player 1 receives *x*. Therefore, player 2's kindness of non-participation is *x*/2. Similarly, player 2's kindness of participation is -*x*/2. In order to calculate how kind player 2 believes player 1 is after choosing *NP* we have to specify player 2's belief of player 1's belief about player 2's choice after *NP*.[17] Denote this by *p"*. Then player 2's belief about how much payoff player 1 intends to give to player 2 by choosing *NP* is $p" x + (1 - p")(x + \delta)$, and since player 2's payoff resulting from player 1's choice of *P* would be *x*,[18] player 2's belief about player 1's kindness from choosing *NP* is:

$$p" x + (1 - p")(x+\delta) - 0.5\,(p" x + (1 - p")\,(x + \delta) + x - \varepsilon) = 0.5\,((1 - p")\,\delta + \varepsilon).$$

This implies that when player 1 does not participate and the second order belief is *p"*, player 2's utility of non-participation is given by $x + 0.5\,Y_2\,(x/2)\,((1 - p")\,\delta + \varepsilon)$,

---

[17] In principle we also need 2's belief about 1's behavior. However, after 1 has already chosen *NP*, 2 already knows what 1 has done, and 2's belief has to be in accordance with her knowledge.

[18] In any SRE player 2 participates after a participation of 1 (Remark 1).

whereas player 2's utility of participation is $(x + \delta)$ -0.5 $Y_2(x/2)((1 - p'') \delta + \varepsilon)$. The former is larger than the latter if $Y_2(x/2)((1 - p'') \delta + \varepsilon) > \delta$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium player 2 does not participate, the condition must hold for $p'' = 1$. This is the case if $Y_2 > \frac{2\delta}{\varepsilon x}$. On the other hand, if in equilibrium player 2 participates, that condition must not hold for $p'' = 0$; This implies that $Y_2 < \frac{2\delta}{x(\delta+\varepsilon)}$ For intermediate values of $Y_2\left(\frac{2\delta}{x(\delta+\varepsilon)} < Y_2 < \frac{2\delta}{\varepsilon x}\right)$, neither non-participation nor participation can be of an equilibrium. In order to have a mixed equilibrium, the utility of non-participation must be equal to the utility of participation. This is the case when $Y_2(x/2)((1-p)\delta+\varepsilon) = \delta$. Since in equilibrium the second order belief is correct, the actual probability of non-participation, p, must be such that the condition is fulfilled. This implies that $p = 1 + \frac{\varepsilon}{\delta} - \frac{2}{Y_2 x}$.

Notice that probability p equals zero for $Y_2 = 2\delta/(x(\delta+\varepsilon))$, and p equals one $Y_2 = 2\delta/\varepsilon x$. Hence, Remarks 1 and 2 together imply that for a given parameter Y_2 2's equilibrium behaviour is unique. This is, however, in general not true for 1's behaviour which can be characterized by three observations:

Remark 3: If $Y_2 < \frac{2\delta}{x(\delta+\varepsilon)}$, participation is the unique 1's equilibrium behaviour.

Notice that for $Y_2 < \frac{2\delta}{x(\delta+\varepsilon)}$ teacher 2 always participates (Remarks 1 and 2). Hence, only the reciprocity part of the utility function can make 1 to choose *NP* (the material payoff alone would dictate for 1 to choose *P*). However, for any second order belief about 1's behaviour 2's strategy of always participating is unkind. Hence, the reciprocity payoff as well as the material payoff makes teacher 1 chooses *P*.

Remark 4: If $Y_2 > \frac{2\delta}{\varepsilon x}$, teacher 1's equilibrium behaviour is typified by one of the following possibilities:

a) Teacher 1 does not participate (regardless of Y_1)

b) $Y_1 > 2/(\varepsilon + \delta)$ and teacher 1 participates

c) $Y_1 > 2/(\varepsilon + \delta)$ and teacher 1 does not participate with probability
$q = 1 - ((xY_1 + 2)/(Y_1)(\varepsilon + x + \delta))$

Note that $Y_2 > \frac{2\delta}{\varepsilon x}$ implies that teacher 2 does not participate when teacher 1 does not participate and participates when teacher 1 participates (Remark 1 and 2). Hence, teacher 1 can give teacher 2 a material payoff of at least $(x - \varepsilon)$ and at most x. Thus, the "equitable" payoff of teacher 1 is $x - (\varepsilon/2)$. If teacher 1 chooses non-participation, teacher 2 receives x. Therefore, teacher 1's kindness of non-participation is $\varepsilon/2$. Likewise, teacher 1's kindness of participation is $-(\varepsilon/2)$. In order to calculate how kind teacher 1 believes teacher 2 is, we have to specify teacher 1's belief about what teacher 2 believes teacher 1 will do. Denote by q" this second order belief of teacher 1 choosing *NP*. Then teacher 1 believes that teacher 2 believes that she gives teacher 1 a material payoff of $q"\, x + (1 - q")(x - \varepsilon)$ by choosing her equilibrium strategy. If teacher 2 always chooses not to participate, teacher 1's payoff is $q"\, x + (1 - q")(x - \delta)$, whereas if teacher 2 always participates, teacher 1's payoff is $q"\, 0 + (1 - q")(x - \varepsilon)$. Hence, teacher 1's belief about teacher 2's kindness from choosing *np* after *NP* and *p* after *P* is given by:

$$q"\, x + (1 - q")(x - \varepsilon) - 0.5\,(q"\, x + (1 - q")\,(x + \delta) + q"\, 0 + (1 - q")\,(x - \varepsilon))$$
$$= 0.5\,(-\varepsilon - \delta + q"\,(\varepsilon + x + \delta))$$

This implies that when teacher 2 plays the equilibrium strategy and the second order belief is q", teacher 1's utility of non-participation is given by:

$$x + 0.5Y_1\,(\varepsilon/2)\,(-\varepsilon - \delta + q"\,(\varepsilon + x + \delta))\,,$$

whereas teacher 1's utility of participation is:

$$x - \varepsilon - 0.5Y_1\,(\varepsilon/2)\,(-\varepsilon - \delta + q"\,(\varepsilon + x + \delta))\,.$$

The former is larger than the latter if $\varepsilon + Y_1\,(\varepsilon/2)\,(-\varepsilon - \delta + q"\,(\varepsilon + x + \delta)) = 0$. In equilibrium, the second order belief must be correct. Hence, if in equilibrium teacher 1 does not participate, the condition must hold for $q" = 1$, which is always the case.

On the other hand, if teacher 1 participates in equilibrium, the condition must not hold for $q" = 0$. This implies that $Y_1 > 2/(\varepsilon + \delta)$. In order to have a mixed equilibrium, the utility of non-participation must be equal to the utility of participation.

This is the case when $\varepsilon + Y_1\,(\varepsilon/2)\,(-\varepsilon - \delta + q"\,(\varepsilon + x + \delta)) = 0$. Since in equilibrium the second order belief must be correct, the actual probability of non-participation, q, must be such that the condition is fulfilled. This implies that $q = 1\,((xY_1 + 2)/(Y_1(\varepsilon + x + \delta)))$.

Next we turn to the equilibrium behaviour when teacher 2 is moderately motivated by reciprocity and hence answers a non-participating choice of teacher 1 with mixing.

Remark 5: if $Y_2 > \frac{2\delta}{x(\delta+\varepsilon)} < Y_2 < \frac{2\delta}{\varepsilon x}$, teacher 1's equilibrium behaviour is characterized by one of the three following possibilities:[19]

a) $Y_2 > \frac{4\delta}{x(\delta+x)}$ and teacher 1 does not participate

b) $Y_1 > \frac{x\left(-4\delta + Y_2\left(\delta x + x^2\right)\right)}{\delta^2(2\delta+x)}$ and teacher 1 participates

c) $Y_1 > \frac{x\left(-4\delta + Y_2\left(\delta x + x^2\right)\right)}{\delta^2(2\delta+x)} > 0$ and teacher 1 does not participate with probability

$$q = \frac{Y_2\left(2\delta^2 Y_1 + 4x + \delta Y_1 x - x^2 Y_2 - \frac{x^3 Y_2}{\delta}\right)}{Y_1\left(-8\delta + 2\delta^2 Y_2 + 3\delta x Y_2 + 2x^2 Y_2\right)}$$

To see this, notice that $\frac{2\delta}{x(\delta+\varepsilon)} < Y_2 < \frac{2\delta}{\varepsilon x}$ implies that teacher 2 does not participate with probability $p = 1 + \frac{\varepsilon}{\delta} - \frac{2}{Y_2 x}$ when teacher 1 does not participate, and teacher 2 participates when teacher 1 participates (see Remarks 1 and 2). Hence, teacher 1 can give teacher 2 a material payoff of at least $(x + \varepsilon)$ and most $px + (1 - p)(x + \delta)$. Hence, the "equitable" payoff of teacher 1 is $0.5\left((1-p)\delta - \varepsilon + 2x\right)$. If teacher 1 chooses no participation, teacher 2 receives $px + (1 - p)(x + \delta)$. Therefore, teacher 1's kindness of no participation is $0.5\left((1-p)\delta + \varepsilon\right)$. Similarly, teacher 1's kindness of participations is $0.5\left((1-p)\delta + \varepsilon\right)$. In order to calculate how kind teacher 1 believes teacher 2 is, we have to specify teacher 1's belief about what teacher 2 believes that teacher 1 will do. Denote by q" this second order belief of teacher 1 choosing *NP*. Then teacher 1 believes that she gives teacher 1 a material payoff of q"$(px + (1 - p)0) + (1 - q")(x - \varepsilon)$ by her equilibrium strategy. If teacher 2 always chooses not to participate, teacher 1's payoff is q" $x + (1 - q")(x + \delta)$, whereas if teacher 2 always participates, teacher 1's payoff is q" $0 + (1 - q")(x - \varepsilon)$. Hence, teacher 1's belief about teacher 2's kindness of her equilibrium strategy is

---

[19] To obtain the specific right-hand side values of these inequalities we assume $\varepsilon = 1/2\ x$. This assumption is no essential and it is made to simplify calculations. Analogous results can be obtained without use it.

$q''(px+(1-p)0)+(1-q'')(x-\varepsilon)-0.5(q''x+(1-q''(x+\delta)+q''0+(1-q''(x-\varepsilon))=$
$q''px-0.5((1-q'')\varepsilon+(1-q'')\delta+q''x)$

This implies that when teacher 2 plays the equilibrium strategy and the second order belief is q'', teacher 1's utility of non-participation is given by

$$px+\frac{1}{2}Y_1((1-p)\delta+\varepsilon)\left(q''px-\frac{1}{2}((1-q'')\varepsilon+(1-q'')\delta+q''x)\right),$$

whereas teacher 1's utility of participation is

$$x-\varepsilon-\frac{1}{2}Y_1((1-p)\delta+\varepsilon)\left(q''px-\frac{1}{2}((1-q'')\varepsilon+(1-q'')\delta+q''x)\right).$$

The former is larger than the latter if

$$-(1-p)x+\varepsilon+Y_1((1-p)\delta+\varepsilon)\left(q''px-\frac{1}{2}((1-q'')\varepsilon+(1-q'')\delta+q''x)\right)>0\ .$$

In equilibrium, the second order belief must be correct. Hence, if in equilibrium 1 does not participate, the condition must hold for $q''=1$, that is

$$px-(x-\varepsilon)+Y_1((1-p)\delta+\varepsilon)\left(px-\frac{1}{2}x\right)>0$$

In general we have a solution for p finding out the roots of the left-hand side quadratic equation. To simplify calculations, let us assume $\varepsilon=\frac{1}{2}x$. In this case the condition holds if $p>0.5$. This in turn implies that $Y_2>\frac{4\delta}{x(\delta-x)}$ (see the calculation of p in Remark 2c).

On the other hand, if in equilibrium 1 participates, the condition must not hold for $q''=0$. Inserting for p and rearranging terms this leads to $Y_1>\frac{x\left(-4\delta+Y_2\left(\delta x+x^2\right)\right)}{\delta^2(2\delta+x)}$.

In order to have a mixed equilibrium, the utility of non-participation must be equal to the utility of participation. This is the case when

$$-(1-p)x+\varepsilon+Y_1\left((1-p)\delta+\varepsilon\right)\left(q''px-\frac{1}{2}\left((1-q'')\varepsilon+(1-q'')\delta+q''x\right)\right)=0$$

Since in equilibrium the second order belief must be correct, that actual probability of non-participation, *q*, must be such that the condition is fulfilled. Substituting for *p* this implies that

$$q=\frac{Y_2\left(2\delta^2Y_1+4x+\delta Y_1x-x^2Y_2-\frac{x^3Y_2}{\delta}\right)}{Y_1\left(-8\delta+2\delta^2Y_2+3\delta xY_2+2x^2Y_2\right)}.$$

The other conditions of Remark 5c are necessary to guarantee that *q* is larger than zero and smaller than 1.

On the other hand, we can also derive the solution for the sequential version of a teachers game with a trade union. It is enough to assume $\varepsilon=0$. Doing so, Remarks 1 and 3 remain unaffected. As Remark 2a does not hold, Remarks 4 and 5a neither do. Results are summarized in the following remarks.

Remark 1': If teacher 1 participates (chooses *P*), teacher 2 also participates in every SRE

Remark 2': If teacher 1 does not participate, the following holds in all SRE:

a) If $Y_2<\frac{2}{x}$, teacher 2 participates

b) If $Y_2<\frac{2}{x}$, teacher 2 does not participate with probability of $p=1-\frac{2}{Y_2x}$

Remark 3': If $Y_2<\frac{2}{x}$, participation is teacher 1's unique equilibrium behaviour.

Remark 4': if $Y_2>\frac{2}{x}$, teacher 1's equilibrium behavior is characterized by: $Y_1>\frac{x(-2+xY_2)}{\delta^2}$ and teacher 1 participates

Remark 5': if $Y_2>\frac{4}{x}$, teacher 1's equilibrium behavior is characterized by:

$Y_1 < \frac{2xY_2}{\delta(xY_2 - 4)}$ and teacher 1 does not participate with probability $q = \frac{Y_2(\delta^2 Y_1 + 2x)}{Y_1(-4\delta + \delta^2 Y_2 + \delta x Y_2)}$.