ELSEVIER

Journal of Economic Psychology 23 (2002) 27–48

JOURNAL OF Economic Psychology

www.elsevier.com/locate/joep

Is it rational to internalize the personal norm that one should reciprocate?

Volker Grossmann [*]

Socioeconomic Institute, University of Zurich, Rämisstrasse 62, CH-8001 Zurich, Switzerland

Received 1 June 2000; received in revised form 23 August 2001; accepted 24 September 2001

Abstract

This paper shows in a simple game-theoretic model that it can be rational for non-altruistic individuals to adopt a personal value-based norm to reciprocate. Moreover, it is argued that such a behavioral commitment is feasible and thus self-binding. Reciprocal behavior has become a stylized fact in experimental labor markets. Our analysis suggests that in laboratory experiments "workers" may provide high effort either because they adopted the norm to behave reciprocally fair or because they fear to "work" with an "employer" who adopted the norm to punish unkind behavior. © 2002 Elsevier Science B.V. All rights reserved.

PsycINFO classification: 3120

JEL classification: C70

Keywords: Reciprocal behavior; Personal norms; Experimental labor market; Fair wage-effort hypothesis

1. Introduction

Why do we treat kind persons kindly and unkind persons unkindly? *Reciprocity* is part of everyone's experience in daily social interactions. Various sociological theories have been proposed to explain reciprocal behavior. First, *equity theory* suggests that people aim to equalize the ratio of inputs to outcomes in social interactions (e.g. Adams, 1963). Second, according to *social exchange theory*, people are kind to others

[*] Corresponding author. Tel.: +41-1-634-2288; fax: +41-1-634-4996.
E-mail address: volker.grossmann@wwi.unizh.ch (V. Grossmann).

PII: S0167-4870(01)00066-6

for purely selfish reasons, e.g. they want the recipient to become obliged to return the favor at some later date, to gain friendship, to impress others, or they hope to gain social approval and social acceptance (e.g. Blau, 1964; Homans, 1961). According to this theory, it is a *social norm* that one should reciprocate, i.e. a norm which is anchored in social groups and sustained by people's anticipation of social sanctions when violating this norm (see also Gouldner, 1960). Related to social exchange theory is the notion of "reciprocal altruism" (Trivers, 1971), according to which individuals aim to build a *reputation* to reciprocate. A "reciprocal altruist" thus reciprocates only if this generates future rewards. Third, the theory of *repeated games* has shown that strategies based on reciprocal behavior can lead to efficiency-enhancing cooperation among non-altruistic individuals. Placed in competition with a variety of other computer programs modeling sets of exchange rules, Axelrod (1984) has shown that the "tit-for-tat" rule of social exchange (i.e. cooperate only if you observe cooperation of your exchange partner, defect otherwise) yields the best pay-offs. Fourth, evolutionary game theory has suggested specific *social learning* models to explain cooperative behavior. In these models, natural or cultural selection of types determines the equilibrium share of cooperative individuals (e.g. Gale, Binmore, & Samuelson, 1995; Güth & Yaari, 1992). Finally, it has been argued that people have internalized a value-based *personal norm* that one should reciprocate. In contrast to a social norm, a personal norm is anchored in the self, and sustained by self-evaluation and self-sanctioning (e.g. Schwartz & Howard, 1984).

This paper shows that it can be both rational and feasible for non-altruistic individuals to adopt such a personal norm. It can be *rational* if a behavioral commitment can be credibly (even though imperfectly) signaled, yielding on average more attractive offers from partners in bilateral bargaining games. For instance, visible emotions like facial expressions, gestures or the unconscious choice of words in a conversation may indicate behavioral commitments (Frank, 1987, 1988). The commitment to reciprocate is also *feasible* and thus self-binding, if deviating from it induces (psychological) costs which outweigh the (material) benefits from deviation. As Hirshleifer (1987, p. 316) argues, emotions like gratitude or anger triggers emotions which help to "guarantee execution of threats and promises and thereby promote achievement of mutually beneficial solutions".

The game-theoretic model in this paper incorporates both ideas: imperfect signaling of personal norms and psychological costs of deviating from behavioral commitments. It is assumed that individuals can decide whether or not to commit to a behavioral norm. Formally, this is represented as choice among two probability density functions of signals. That is, signaling is random, but the probability of being recognized as someone who internalized the norm to reciprocate depends on the choice of a signal distribution (Frank, 1987, 1988). A commitment to reciprocate can be rational in the sense that the expected utility function (representing purely non-altruistic preferences) in a subsequent bilateral bargaining game is maximized. If commitment is ruled out, standard non-cooperative game theory suggests that individuals do not reciprocate. This is illustrated in Section 2.1. In Section 2.2, we allow the utility function of individuals to be conditional on their decisions about personal norms with respect to reciprocal behavior, by incorporating costs from deviating from an adopted norm. The latter game is analyzed in Section 3. Section 4 illustrates which kind of equilibria exist.

It has been widely recognized that reciprocal behavior can substantially affect outcomes of economic transactions, even with macroeconomic implications. For instance, fairness considerations in the labor market between employers and employees have been suggested to affect both unemployment and the wage distribution (Akerlof, 1982; Akerlof & Yellen, 1988, 1990).[1] Empirical studies support the economic relevance of fairness perceptions and reciprocal behavior. For instance, survey evidence shows that price rigidities and wage setting behavior is motivated by these factors (e.g. Agell & Lundborg, 1995; Bewley, 2000; Kahneman, Knetsch, & Thaler, 1986; Levine, 1993). Evidence from laboratory experiments shows that reciprocal behavior is prevalent even in anonymous one-shot labor market games (e.g. Fehr, Gächter, & Kirchsteiger, 1996; Fehr, Kirchsteiger, & Riedl, 1993, 1998 for a comprehensive survey of experimental evidence about fairness and reciprocity, see Fehr & Gächter, 2000).

The game analyzed in this paper is motivated by these experimental labor markets. Obviously, evidence from such one-shot experiments cannot be explained by social exchange theory or the analysis of repeated games, which would refer to long-term employer–employee relationships. (Section 5 discusses how our model can contribute to the understanding of the evidence from experimental labor markets.)

The following explanations of reciprocal behavior in such experiments have been proposed. In line with equity theory, it has been assumed that individuals are motivated by interdependent preferences, i.e. by the final distribution of the material payoff (Bolton & Ocksenfels, 2000; Fehr et al., 1998; Fehr & Schmidt, 1999). Moreover, so called "reciprocity utility functions" have been suggested to explain both fair and punishing behavior of agents (Dufwenberg & Kirchsteiger, 1998; Falk & Fischbacher, 1999; Levine, 1998).[2] Whereas both types of models show that one can *find* preferences which explain reciprocal behavior, this paper deals with the question if it is rational to *have* a norm to reciprocate. An alternative approach to the one in this paper is the evolution of norms (i.e. social learning) in economies populated by bounded-rational individuals on basis of evolutionary success (e.g. Güth, 1995; Güth & Yaari, 1992; Höffler, 1999). Section 6 discusses the relationship between this approach and the one pursued in this paper. Section 7 summarizes the paper.

2. A simple labor market model without and with norm internalization

Akerlof and Yellen (1990) formulate gift exchange in employer–employee relationships by the so-called *fair wage-effort hypothesis*. According to this hypothesis,

[1] Other examples of economic consequences of reciprocity are the private provision of public goods (Sugden, 1984) and the enforcement of contracts (Fehr, Gächter, & Kirchsteiger, 1997).

[2] For a simultaneous move version of these kinds of psychological games, see the seminal paper of Rabin (1993).

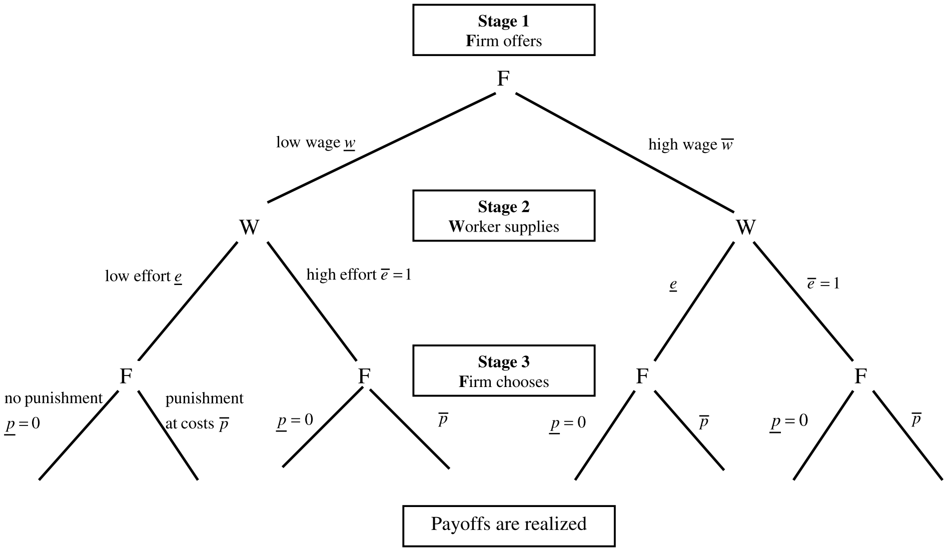

Fig. 1. Timing of events in game Γ.

effort e provided by a worker is a non-decreasing function of the ratio of her actual wage w to the wage level $\bar{w}$ she perceives as fair. Formally, this kind of reciprocal behavior of workers can be stated as

$$e = \min\{w/\bar{w}, 1\}, \tag{1}$$

i.e. full effort is normalized to unity. Section 2.2 presents a simple labor market model which rationalizes this kind of behavior by allowing individuals to internalize the personal norm to reciprocate.

2.1. A labor market model without norm internalization

As a benchmark case, first consider a simple finite game (Γ) of perfect information with players $i \in \{F, W\}$. A productive unit consists in exactly one firm F and one worker W. Action sets in each stage of the game are assumed to be finite. As depicted in Fig. 1, the timing of events is as follows.

At the initial decision node a firm owner F announces the wage payment to her employee. For simplicity, she can only choose among two wages, a low wage $\underline{w}$ and a high wage $\bar{w}$. After the wage has been announced, the worker W can either supply a low effort level $\underline{e}$ with $0 < \underline{e} < 1$ or a high effort $\bar{e} = 1$. At the last stage the employer can either punish the worker at low costs $\underline{p} = 0$ (i.e. no punishment) or at high costs $\bar{p} > 0$.[3] For simplicity, a punished worker bears the same cost p as her employer.

[3] This setup is motivated by the experimental labor market in Fehr et al. (1996).

Reciprocal behavior in this context means that the worker supplies full effort $\bar{e} = 1$ in response to a high wage offer $\bar{w}$ (and $\underline{e}$ in response to $\underline{w}$) and the employer punishes if she offered $\bar{w}$ but observed low effort $\underline{e}$ (and does not punish otherwise).

Payoff functions of workers and firm owners, respectively, are given by

$$u^W = w - e - p \quad \text{and} \quad u^F = \beta e - w - p, \tag{2}$$

where $\beta > 0$ is a productivity parameter, $w \in \{\underline{w}, \bar{w}\}, e \in \{\underline{e}, 1\}$ and $p \in \{0, \bar{p}\}$. Wages are assumed to be a share α of "output" βe, where $0 < \alpha < 1$. That is, $\underline{w} = \alpha\beta\underline{e}$ and $\bar{w} = \alpha\beta$, where the latter can be viewed as the "fair" wage. Hence, if $e = 1$ is supplied in response to a wage payment $\bar{w}$, and $\underline{e}$ in response to $\underline{w}$, the fair-wage effort hypothesis (1) exactly holds.

The unique subgame perfect equilibrium (SPE) of this game is given by

$$((\underline{w}, \underline{p} \text{ no matter if } W \text{ plays } \underline{e} \text{ or } \bar{e}),\ (\underline{e} \text{ no matter if } F \text{ plays } \underline{w} \text{ or } \bar{w})) \tag{3}$$

(the first component is the SPE strategy of player F, the second one of player W). The corresponding payoffs are $((u^F)^{\mathrm{SPE}}/(u^W)^{\mathrm{SPE}}) \equiv (\underline{e}\beta(1-\alpha)/\underline{e}(\alpha\beta - 1))$. It is easy to see that this equilibrium is inefficient. All strategy pairs in which the firm owner pays the high wage and the worker supplies full effort (i.e. reciprocates), e.g. $((\bar{w}, \bar{p}$ if W plays $\underline{e}$ and $\underline{p}$ if W plays $\bar{e})$, $(\bar{e}$ if F plays $\bar{w}, \underline{e}$ if F plays $\underline{w}))$, would be Pareto-improving. The reason for the inefficiency of the SPE is that both players anticipate that the partner will be selfish at any decision node. Hence, this model cannot explain reciprocal behavior in a one-shot situation.

2.2. A labor market model with norm internalization

Now consider the following game Γ_C. Prior to the three-stage "bargaining process" depicted in Fig. 1, individuals are now able to adopt the personal norm that one should reciprocate. (Throughout the paper, we often refer to this as "commitment to reciprocate".) According to the assumptions below, such a commitment is feasible. That is, a worker can commit herself to supply full effort $\bar{e} = 1$ whenever she is paid the fair wage $\bar{w}$. An employer can commit herself to punish an exploiting worker at costs $\bar{p}$ if the worker provides minimum effort $\underline{e}$, despite having received a "fair" wage $\bar{w}$. It is reasonable to assume that whether or not an individual has internalized the norm to reciprocate is not perfectly observable. Rather, both types of players imperfectly signal this decision, e.g. by visible emotions. This idea closely follows Frank (1987, 1988), who provides an extensive discussion about the role of emotions by signaling behavioral attitudes.[4] Formally, the individual choice whether or not to adopt the norm to reciprocate is represented by the choice between two probability density functions (p.d.f.'s) of signals. Which one is chosen is unobservable for the other player. After this decision, an observable signal from the chosen p.d.f. is realized. Each player only knows the p.d.f. of her signal in advance, but

[4] Frank (1987) is concerned with cooperative behavior in a one-shot simultaneous-move game, whereas we are concerned with reciprocal behavior in a sequential game.

not its realization. After the realization of the signals, players enter the three-stage bargaining process depicted in Fig. 1. Table 1 summarizes the timing of events.

Denote the p.d.f. of a signal S of a committed worker with $f_C(\cdot)$ and of an uncommitted worker with $f_N(\cdot)$. It is assumed that the expected realization of S is higher for a worker who has chosen $f_C(\cdot)$, i.e. for a committed worker. Similarly, denote by $g_C(\cdot)$ and $g_N(\cdot)$ the p.d.f.'s of the signal T of a committed and uncommitted firm owner, respectively. Again, the expected realization of T is higher for an employer who has chosen $g_C(\cdot)$ rather than $g_N(\cdot)$.

These assumptions ensure that the choice to adopt the behavioral norm to reciprocate can be *rational*. In order to ensure that it is also *feasible*, such a behavioral norm is assumed to affect the players' utility derived from the outcome of the subsequent three-stage bargaining process. It has been convincingly argued that consistent behavior is sustained by emotional dispositions. In our model, this would mean that, for instance, not being kind to kind persons may trigger feelings of guilt, and unkind behavior may trigger anger and thus the desire to seek vengeance. In other words, individuals anticipate that (not) reciprocating would make them feel good (bad) (see e.g. Elster, 1996; Frank, 1988; Hirshleifer, 1987; Hoffman, 1984). In our context, these discussions suggest that emotions help to behave according to self-based and consciously chosen norms of behavior. Formally, deviating from the norm to reciprocate in the bilateral bargaining process implies some costs c (of feeling guilty) for a worker and d (of suppressing anger) for an employer, respectively. That is, utility functions are now given by:

$$U^W = \begin{cases} u^W - c & \text{if } f_C(\cdot) \text{ is chosen, but } W \text{ does not reciprocate,} \\ u^W & \text{otherwise,} \end{cases} \tag{4a}$$

$$U^F = \begin{cases} u^F - d & \text{if } g_C(\cdot) \text{ is chosen, but } F \text{ does not reciprocate,} \\ u^F & \text{otherwise,} \end{cases} \tag{4b}$$

where u^W and u^F are given by (2). Both the signal p.d.f.'s and the utility functions (4a) and (4b) are common knowledge among the players.

Feasibility of a commitment to reciprocate requires that the costs of deviating from a norm to reciprocate outweigh the (material) benefits of deviating from it. Using (2), it is easy to see that a worker who does not provide high effort in response to

Table 1
Timing of events in game Γ_C

No.	Event
1	Signal p.d.f.'s are chosen, imperfectly signaling if committed to reciprocate and determining utility functions: Worker chooses f_C or f_N, employer chooses g_C or g_N
2	Signals S and T of worker and employer, respectively, are realized
3	Employer offers wage $\bar{w} = \alpha\beta$ or $\underline{w} = \underline{e}\alpha\beta$
4	Worker supplies effort $\bar{e} = 1$ or $\underline{e} < 1$
5	Employer chooses whether or not to punish worker ($\bar{p} > 0$ or $\underline{p} = 0$)

a high wage offer gains $1 - \underline{e}$, provided that she is not punished. (If she is punished, the payoff difference of not reciprocating equals $1 - \underline{e} - \bar{p}$.) Similarly, an employer who does not punish (although she is exploited) gains $\bar{p}$. Thus, commitment to reciprocate is feasible if $c > 1 - \underline{e}$ and $d > \bar{p}$, which is assumed from now on.

Finally, let $f_N(S)/f_C(S)$ and $g_N(T)/g_C(T)$ be non-increasing functions of S and T, respectively. These conditions are known as *monotone likelihood ratio properties* in the principal–agent literature (see Milgrom, 1981). Adopting these standard assumptions will lead to plausible behavior of both employers *after* receiving signals S and T, respectively.

Two remarks are at order. First, due to the assumptions $c > 1 - \underline{e}$ and $d > \bar{p}$, respectively, the decision whether or not to adopt the behavioral norm to reciprocate fully determines the strategy profile for both players in the subsequent subgame, conditional on the realization of signals. That is, if a player is committed and the received signal is above a certain threshold (which is derived below), she will reciprocate. Without commitment, a worker will always supply minimum effort $\underline{e}$ and an employer will never punish (i.e. will always choose $p = 0$). Second, note that we generally allow for mixed strategies with respect to the choice whether or not to commit to a behavioral norm. At the end of next section, Nash equilibria of Γ_C in mixed strategies are interpreted and discussed.

3. Equilibrium analysis

3.1. Threshold signal values

Denote the mixed strategies of a worker and an employer to reciprocate, i.e. the probabilities assigned to the pure strategies $f_C(\cdot)$ and $g_C(\cdot)$, with s_C and t_C, respectively. As the payoff functions in (4a) and (4b) indicate, both workers and employers are risk-neutral.

Employers. Consider an employer who has chosen whether or not to commit herself to punish unfair workers; i.e. given her mixed strategies t_C and $(1 - t_C)$ she has flipped a coin reflecting this randomization and made her choice according to the outcome. According to (4b), a committed employer will find it prudent to pay the fair wage $\bar{w}$ after receiving a signal S from a worker if and only if

$$\mathrm{prob}\{C|S\} \cdot \beta(1-\alpha) + (1 - \mathrm{prob}\{C|S\}) \cdot (\beta(\underline{e} - \alpha) - \bar{p}) \geqslant \underline{e}\beta(1-\alpha). \tag{5}$$

According to Bayes' theorem, the probability $\mathrm{prob}\{C|S\}$ that a worker with signal S reciprocates (i.e. supplies full effort when $\bar{w}$ is paid) is given by

$$\mathrm{prob}\{C|S\} = \frac{\mathrm{prob}\{C\} \cdot \mathrm{prob}\{S|C\}}{\mathrm{prob}\{S\}} = \frac{s_C f_C(S)}{s_C f_C(S) + (1 - s_C) f_N(S)}. \tag{6}$$

Using (6), it is easy to check that condition (5) is equivalent to

$$s_C \geqslant \frac{f_N(S)(\alpha\beta(1-\underline{e}) + \bar{p})}{f_C(S)\beta(1-\alpha)(1-\underline{e}) + f_N(S)(\alpha\beta(1-\underline{e}) + \bar{p})}, \tag{7}$$

i.e. the higher the unconditional probability s_C that a worker reciprocates, the "more likely" it is that a committed firm owner pays the high wage to a worker for a given signal S. Similarly, an uncommitted employer will pay the fair wage after receiving S if and only if

$$\mathrm{prob}\{C|S\} \cdot \beta(1-\alpha) + (1-\mathrm{prob}\{C|S\}) \cdot \beta(\underline{e}-\alpha) \geqslant \underline{e}\beta(1-\alpha), \tag{8}$$

which is equivalent to

$$s_C \geqslant \frac{f_N(S)\alpha\beta(1-\underline{e})}{f_C(S)\beta(1-\alpha)(1-\underline{e}) + f_N(S)\alpha\beta(1-\underline{e})}, \tag{9}$$

according to (6). It is easy to show that according to the monotone likelihood ratio property assumed above, the right-hand side of (7) and (9), respectively, are non-increasing in S. Thus, we can define $\bar{S}^*$ and $\underline{S}^*$ as the minimum values of S such that (7) and (9) hold, respectively. That is, a committed (an uncommitted) employer pays the high wage whenever she receives a signal $S \geqslant \bar{S}^*$ $(S \geqslant \underline{S}^*)$. $\bar{S}^*$ and $\underline{S}^*$ may thus be called *threshold signal values*. According to (9) and the monotone likelihood ratio property, these threshold signals have the following plausible properties. First, we have $\bar{S}^* \geqslant \underline{S}^*$, i.e. a committed employer is not less cautious than an uncommitted employer, since the former has more to lose due to costly punishment when her wage payment is exploited. Note that the costs $d > \bar{p}$ prevent her from suppressing her anger. Second, both $\bar{S}^*$ and $\underline{S}^*$ are non-increasing in the unconditional probability s_C that the worker is fair, i.e. an employer is not less cautious when a worker is less likely to reciprocate.

Workers. Similarly, consider a worker who has chosen whether or not to commit herself to be kind when treated kindly, again, by flipping a coin reflecting her mixed strategies s_C and $(1-s_C)$. Due to the costs $c > 1-\underline{e}$ of not reciprocating, a committed worker would always supply full effort after receiving a fair wage payment. Moreover, she can never receive less than $(u^W)^{\mathrm{SPE}} = \underline{e}(\alpha\beta-1)$, her utility if not recognized as fair. In contrast, an uncommitted worker only provides high effort to an employer who credibly signals to be punishing. Formally, if the signal value of an uncommitted worker is high enough such that her employer is willing to pay a high wage to her, she would supply the low effort $\underline{e}$ to an employer signaling T if and only if

$$\mathrm{prob}\{C|T\} \cdot (\alpha\beta-\underline{e}-\bar{p}) + (1-\mathrm{prob}\{C|T\}) \cdot (\alpha\beta-\underline{e}) \geqslant \alpha\beta-1. \tag{10}$$

If the punishment $\bar{p}$ would be sufficiently low, (9) would always hold, i.e. an uncommitted worker would always exploit a high wage payment. Thus, to consider a nontrivial case it is assumed that $\bar{p} > 1-\underline{e}$.

Conditional on a signal T, a worker who does not reciprocate in response to a fair wage offer $\bar{w}$ receives punishment with probability

$$\mathrm{prob}\{C|T\} = \frac{t_C g_C(T)}{t_C g_C(T) + (1-t_C)g_N(T)}. \tag{11}$$

Using (11), (10) can be equivalently stated as

$$t_C \leqslant \frac{g_N(T)(1-\underline{e})}{g_C(T)(\bar{p}-(1-\underline{e})) + g_N(T)(1-\underline{e})}. \tag{12}$$

According to the monotone likelihood ratio property assumed above, the right-hand side of (12) is non-increasing in T. Hence, an uncommitted worker tries to exploit her employer if and only if the signal value received is sufficiently low, i.e. if and only if $T \leqslant T^*$, where T^* denotes the maximum value of T such that (12) holds. Stated differently, the lower the unconditional probability t_C that a firm owner punishes, the "more likely" it is for a given signal T that an uncommitted worker supplies low effort in response to $\bar{w}$. Also note that T^* is non-increasing in both t_C and $\bar{p}$ due to the monotone likelihood ratio property, i.e. an uncommitted worker does not become less cautious if the unconditional probability t_C to be punished or the punishment $\bar{p}$ is higher.

3.2. Expected utility functions and Nash equilibria

In order to find Nash equilibria of the extensive form game Γ_C one has to compute the expected utility functions of each player conditional on the decision whether or not the agent reciprocates, for a given (mixed) strategy of the other agent.

Depending on the decision whether or not to adopt the personal norm that one should reciprocate, workers will signal $S \geqslant \bar{S}^*$ ($S \geqslant \underline{S}^*$) with probability

$$\bar{Q}_j \equiv \int_{\bar{S}^*}^{\infty} f_j(S)\,\mathrm{d}S \left(\underline{Q}_j \equiv \int_{\underline{S}^*}^{\infty} f_j(S)\,\mathrm{d}S\right), \quad j = C, N. \tag{13}$$

Analogously, employers will signal $T > T^*$ with probability

$$R_j \equiv \int_{T^*}^{\infty} g_j(T)\,\mathrm{d}T, \quad j = C, N. \tag{14}$$

Workers. Note that $\bar{S}^* \geqslant \underline{S}^*$ implies both $\underline{Q}_C \geqslant \bar{Q}_C$ and $\underline{Q}_N \geqslant \bar{Q}_N$. The expected utility of a committed worker equals

$$\begin{aligned} E_C^W = {} & \bar{Q}_C(\alpha\beta - 1) + (\underline{Q}_C - \bar{Q}_C)(t_C\underline{e}(\alpha\beta - 1) + (1 - t_C)(\alpha\beta - 1)) \\ & + (1 - \underline{Q}_C)\underline{e}(\alpha\beta - 1). \end{aligned} \tag{15}$$

This can be seen as follows. $\bar{Q}_C$ is the probability that any firm owner will pay the high wage $\bar{w} = \alpha\beta$ to a committed worker, yielding a utility $(\alpha\beta - 1)$. With probability $(\underline{Q}_C - \bar{Q}_C)$ a committed worker does not receive the high wage if she is matched with a committed employer, whereas an uncommitted employer still pays a high wage (provided uncommitted employers are less cautious, i.e. $\underline{Q}_C > \bar{Q}_C$). The probability to meet a committed employer is t_C. With probability $(1 - \underline{Q}_C)$ a committed worker will always get a low wage $\underline{w} = \underline{e}\alpha\beta$, yielding a net payoff $(\underline{e}\alpha\beta - \underline{e})$.

An uncommitted worker will exploit her employer after receiving a high wage, only if she receives a signal $T \leqslant T^*$, because of the punishment she might face. The probability θ that a firm owner with a mixed strategy t_C signals $T \leqslant T^*$ equals

$$\theta \equiv \mathrm{prob}\{T \leqslant T^*\} = t_C(1 - R_C) + (1 - t_C)(1 - R_N). \tag{16}$$

Thus, the expected payoff of an uncommitted worker is given by

$$\begin{aligned} E_N^W = {} & \overline{Q}_N\Big[\theta\Big(\text{prob}\{C|T\leqslant T^*\}(\alpha\beta-\underline{e}-\bar{p})+(1-\text{prob}\{C|T\leqslant T^*\})(\alpha\beta-\underline{e})\Big) \\ & +(1-\theta)(\alpha\beta-1)\Big]+\Big(\underline{Q}_N-\overline{Q}_N\Big)(t_C\underline{e}(\alpha\beta-1)+(1-t_C)(\alpha\beta-\underline{e})) \\ & +(1-\underline{Q}_N)\underline{e}(\alpha\beta-1), \end{aligned} \tag{17}$$

which can be understood as follows. $\bar{Q}_N$ is the probability for an uncommitted worker to get a high wage by any employer. With probability θ she is matched with a firm owner who signals to be an appropriate victim for exploitation. However, with a conditional probability of

$$\text{prob}\{C|T\leqslant T^*\}=\frac{\text{prob}\{C\}\cdot\text{prob}\{T\leqslant T^*|C\}}{\text{prob}\{T\leqslant T^*\}}=\frac{t_C(1-R_C)}{\theta} \tag{18}$$

her employer will nevertheless punish her in response to a low effort supply. Moreover, with probability $(1-\theta)$ an uncommitted worker will be cautious enough not to exploit a high wage payment even if she has $S\geqslant\overline{S}^*$. An uncommitted worker signaling $S\in[\underline{S}^*,\overline{S}^*)$ [with probability $(\underline{Q}_N-\overline{Q}_N)$] receives a fair wage less often than a worker signaling $S\geqslant\overline{S}^*$. But she does not have to fear to be punished *if* she gets the opportunity to exploit, since only uncommitted firm owners will pay the high wage. Thus, with probability $(\underline{Q}_N-\overline{Q}_N)(1-t_C)$ she has $S\in[\underline{S}^*,\overline{S}^*)$ and will get away with unfair behavior. However, with probability t_C the employer is committed, and pays a low wage if she receives $S\in[\underline{S}^*,\overline{S}^*)$. If a worker signals $S<\underline{S}^*$ [with probability $(1-\underline{Q}_N)$] she is not trustworthy enough to get a high wage from any employer.

Employers. In order to compute the expected utility functions of firm owners, note that the total probability that a worker with a mixed strategy s_C signals $S\geqslant\bar{S}^*(S\geqslant\underline{S}^*)$ is given by

$$\bar{\pi}\equiv s_C\bar{Q}_C+(1-s_C)\bar{Q}_N\Big(\underline{\pi}\equiv s_C\underline{Q}_C+(1-s_C)\underline{Q}_N\Big). \tag{19}$$

Moreover, the conditional probability that a worker signaling $S\geqslant\bar{S}^*$ $(S\geqslant\underline{S}^*)$ does not exploit a fair wage payment equals

$$\text{prob}\{C|S\geqslant\bar{S}^*\}=s_C\bar{Q}_C/\bar{\pi}\Big(\text{prob}\{C|S\geqslant\underline{S}^*\}=s_C\underline{Q}_C/\underline{\pi}\Big). \tag{20}$$

Thus, the expected utility for an employer who has committed to punish is given by

$$\begin{aligned} E_C^F = {} & R_C\beta(1-\alpha)+(1-R_C)\ \Big[\bar{\pi}\Big(\text{prob}\{C|S\geqslant\bar{S}^*\}\beta(1-\alpha) \\ & +(1-\text{prob}\{C|S\geqslant\bar{S}^*\})(\beta(\underline{e}-\alpha)-\bar{p})\Big)+(1-\bar{\pi})\underline{e}\beta(1-\alpha)\Big], \end{aligned} \tag{21}$$

whereas the expected utility of an uncommitted employer equals

$$\begin{aligned} E_N^F = {} & R_N\beta(1-\alpha)+(1-R_N)\ \Big[\underline{\pi}\Big(\text{prob}\{C|S\geqslant\underline{S}^*\}\beta(1-\alpha) \\ & +(1-\text{prob}\{C|S\geqslant\underline{S}^*\})\beta(\underline{e}-\alpha)\Big)+(1-\underline{\pi})\underline{e}\beta(1-\alpha)\Big]. \end{aligned} \tag{22}$$

A committed firm owner deters an uncommitted worker from exploiting a high wage payment with probability R_C. However, if a committed employer signals $T \leqslant T^*$ [with a probability $(1 - R_C)$] a seemingly fair worker will exploit a fair wage payment with probability $\bar{\pi} \cdot (1 - \text{prob}\{C|S \geqslant \bar{S}^*\})$. So far the argumentation with respect to uncommitted firm owners is completely analogous. However, a committed firm owner bears punishment costs $\bar{p}$ if she is exploited, whereas an uncommitted employer never punishes. Finally, the probability that a committed firm owner does not deter unfair effort supply and is not matched with a trustworthy worker is given by $(1 - R_C)(1 - \bar{\pi})$. Thus, in this case she will pay the low wage. (An analogous argument holds for an uncommitted employer.)

In a Nash equilibrium of the extensive form game Γ_C, given the other player's randomization, neither player can raise her expected utility by changing her own randomization. Thus, in a Nash equilibrium (s_C^*, t_C^*) it holds that $E_C^F = E_N^F$ if $0 < s_C^* < 1$ and $E_C^W = E_N^W$ if $0 < t_C^* < 1$.

3.3. *Interpretation of mixed Nash equilibria as population shares*

In order to interpret Nash equilibria of our simple labor market model with norm internalization, assume there is a *large population* of both employers and workers, and both populations have an equal size. Firm owners and workers are *randomly pairwise matched* to a productive unit. Like in evolutionary game theory, individuals are confined to use pure strategies only (e.g. Fudenberg & Levine, 1998). That is, each player either adopts the behavioral norm to reciprocate or not, i.e. randomization in choosing signal p.d.f.'s is ruled out (which is the only reasonable assumption in our context). A mixed strategy profile of the extensive form game Γ_C is interpreted as representing the *population share* of individuals who have chosen the corresponding pure strategy. Thus, a Nash equilibrium in mixed strategies (s_C^*, t_C^*) of the extensive form game Γ_C is interpreted to reflect the equilibrium population share of workers and employers, respectively, who are committed to reciprocate. Let's call this a *reciprocity equilibrium*. In reciprocity equilibrium, no individual wants to change her personal norm with respect to reciprocal behavior.

Efficiency of a productive unit is enhanced compared to the SPE of Section 2.1 if the employer pays the high wage and the worker provides full effort. The latter will occur either when the worker is committed to reciprocate or when an uncommitted worker fears punishment. In order to rationalize the "fair-wage effort" hypothesis (1), it suffices to show that efficiency is enhanced at least in some employer–employee relationships.

4. Which kind of reciprocity equilibria do exist?

The following analysis reveals which kind of reciprocity equilibria of Γ_C do exist. Will there always be a positive fraction of employers or workers, respectively, who deliberately internalized the norm to reciprocate? Can there be an equilibrium in which *all* workers or employers, respectively, consistently reciprocate?

These questions are examined in an example with uniform signal distributions. That is,

$$f_C(S) = \begin{cases} \left(z_C^f - y_C^f\right)^{-1} & \text{if } S \in \left[y_C^f, z_C^f\right], \\ 0 & \text{otherwise,} \end{cases}$$
$$f_N(S) = \begin{cases} \left(z_N^f - y_N^f\right)^{-1} & \text{if } S \in \left[y_N^f, z_N^f\right], \\ 0 & \text{otherwise,} \end{cases} \tag{23a}$$

$$g_C(T) = \begin{cases} (z_C^g - y_C^g)^{-1} & \text{if } T \in [y_C^g, z_C^g], \\ 0 & \text{otherwise,} \end{cases}$$
$$g_N(T) = \begin{cases} (z_N^g - y_N^g)^{-1} & \text{if } T \in [y_N^g, z_N^g], \\ 0 & \text{otherwise,} \end{cases} \tag{23b}$$

where $y_N^f < y_C^f < z_N^f < z_C^f$ and $y_N^g < y_C^g < z_N^g < z_C^g$. Furthermore, assume for the sake of simplicity

$$a \equiv \left(z_C^f - y_C^f\right)^{-1} = \left(z_N^f - y_N^f\right)^{-1} \quad \text{and} \quad b \equiv (z_C^g - y_C^g)^{-1} = (z_N^g - y_N^g)^{-1}, \tag{24}$$

respectively.

4.1. Threshold signal values with uniformly distributed signals

Employers. According to (7), (23a) and (24), if $s_C < (\alpha\beta(1-\underline{e}) + \bar{p})/(\beta(1-\underline{e}) + \bar{p}) \equiv \tilde{s}_C$ (i.e. if the share of committed workers is sufficiently low), a committed employer will pay the high wage if and only if she can be completely sure that the worker she is matched with will not exploit her (i.e. if she receives a signal $S > z_N^f$). (Note that $\alpha < \tilde{s}_C < 1$.) Uncommitted firm owners are less cautious since they will never be angry enough to bear the costs of punishing workers. According to (7), (23a) and (24), their threshold signal for paying the fair wage equals z_N^f if and only if $s_C < \alpha$. In contrast, if s_C is sufficiently high, it looks attractive for firm owners to pay the fair wage unless knowing for sure they will be exploited, i.e. unless they receive $S < y_C^f$. Hence,

$$\bar{S}^* = \underline{S}^* = z_N^f \qquad \text{if and only if } 0 \leqslant s_C < \alpha, \tag{25a}$$

$$\bar{S}^* = z_N^f \quad \text{and} \quad \underline{S}^* = y_C^f \quad \text{if and only if } \alpha \leqslant s_C < \tilde{s}_C, \tag{25b}$$

$$\bar{S}^* = \underline{S}^* = y_C^f \qquad \text{if and only if } \tilde{s}_C \leqslant s_C \leqslant 1. \tag{25c}$$

Together with (12), (23a) and (24), one obtains from (25a)–(25c) that the respective shares in total population of workers above those threshold signals are given by

$$\bar{Q}_N = \underline{Q}_N = 0 \quad \text{and} \quad \bar{Q}_C = \underline{Q}_C = a\left(z_C^f - z_N^f\right) \quad \text{for } 0 \leqslant s_C < \alpha, \tag{26a}$$

$$\bar{Q}_N = 0,\ \underline{Q}_N = a\left(z_N^f - y_C^f\right) \quad \text{and} \quad \bar{Q}_C = a\left(z_C^f - z_N^f\right), \quad \underline{Q}_C = 1 \quad \text{for } \alpha \leqslant s_C < \tilde{s}_C, \tag{26b}$$

$$\bar{Q}_N = \underline{Q}_N = a\left(z_N^f - y_C^f\right) \quad \text{and} \quad \bar{Q}_C = \underline{Q}_C = 1 \quad \text{for} \quad \tilde{s}_C \leqslant s_C \leqslant 1. \tag{26c}$$

Workers. Committed workers always reciprocate after receiving a high wage payment. For uncommitted employees, using (12), (23b) and (24) one obtains (in an analogous way to the threshold values of employers derived above)

$$T^* = z_N^g \quad \text{if and only if } 0 \leqslant t_C \leqslant (1 - \underline{e})/\bar{p}, \tag{27a}$$

$$T^* = y_C^g \quad \text{if and only if } (1 - \underline{e})/\bar{p} < t_C \leqslant 1. \tag{27b}$$

(Remember $(1 - \underline{e})/\bar{p} < 1$.) Hence, if the share of punishing employers in the economy is sufficiently high, an uncommitted worker never tries to exploit a high wage payment unless the probability that her employer does never punish equals unity. From (27a) and (27b), the respective shares in total population of firm owners above those threshold signals are given by

$$R_N = 0 \quad \text{and} \quad R_C = b(z_C^g - z_N^g) \quad \text{for } 0 \leqslant t_C \leqslant (1 - \underline{e})/\bar{p}, \tag{28a}$$

$$R_N = b(z_N^g - y_C^g) \quad \text{and} \quad R_C = 1 \quad \text{for } (1 - \underline{e})/\bar{p} < t_C \leqslant 1. \tag{28b}$$

4.2. Reciprocity equilibria

Now reciprocity equilibria can be derived analytically. According to (12), uncommitted workers do not try to exploit a high wage payment if receiving a signal above a threshold value T^*. Thus, the interval $t_C \in [0, 1]$ generally has to be divided in at least two subintervals in order to compute expected utility functions. Moreover, with possibly different signal thresholds for committed and uncommitted employers (i.e. $\bar{S}^* > \underline{S}^*$ for some s_C), one has to consider at least three subintervals of $s_C \in [0, 1]$. Hence, generally one has to look at least at $2 \times 3 = 6$ cases. Considering uniformly distributed signals is the simplest example since one obtains just this minimum number of six cases, according to (25a)–(25c) and (27a) and (27b).

In Appendix A, propositions about the existing reciprocity equilibria are formally stated and proven. The derivation comes from a systematic analysis of the six cases outlined above. Since it is not necessary to be engaged in this formal analysis to grasp the nature and the intuition of the results one obtains, the main implications of the propositions in Appendix A are summarized and discussed in the following.

Result 1. *In any reciprocity equilibrium, a strictly positive share of workers is committed (see Proposition 1 in Appendix A).*

Result 2. *There always exists a reciprocity equilibrium in which all individuals (i.e. employers and workers) are committed (see Proposition 7 in Appendix A). More generally, if the share of committed employers in reciprocity equilibrium is sufficiently large, all workers are committed to reciprocate (see Propositions 2, 5 and 7 in Appendix A). But there may also exist reciprocity equilibria with committed workers only, in which the share of committed employers is low (see Proposition 4 in Appendix A).*

Result 3. *Reciprocity equilibria with only uncommitted employers may exist (see Propositions 3, 4 and 6 in Appendix A).*

Result 4. *If the fraction of non-punishing employers is sufficiently small, there may be a reciprocity equilibrium in which both types of workers, and, in addition, both types of employers coexist (see Propositions 3 and 6 in Appendix A).*

The intuition of these results are as follows. First, if the share of committed workers s_C would be close to zero, firm owners would be reluctant to pay a high wage unless they are matched with a fair worker with probability one, according to (25a). In other words, uncommitted workers would never be paid the high wage. Thus, those workers could have unambiguously raised their expected utility by adopting consistently fair behavior. Thus, a situation without any committed workers can never be an equilibrium. The second result says that there is at least one reciprocity equilibrium with fair workers only, given that the share of committed employers is large. This is because, on the one hand, uncommitted workers have a high probability to be punished and, on the other hand, they receive a high wage less often than committed workers. Third, reciprocity equilibria with only uncommitted firm owners cannot be ruled out. This is simply because, although commitment to punish may deter uncommitted workers from exploiting a high wage, such a commitment is costly if it does not deter unfair behavior. Finally, if the share of committed employers is small, so is the risk for exploiting workers to be punished. Thus, not all workers are necessarily committed in a reciprocity equilibrium. Moreover, if this is the case, employers have to weigh the benefits (deterring unfair workers) and the punishment costs when deciding whether or not to commit to punish unfair workers. This gives rise to the possibility of interior reciprocity equilibria.

5. Explaining reciprocal behavior in experimental labor markets

According to Result 1, a strictly positive share of workers internalize the personal norm that one should provide full effort $\bar{e} = 1$, if the fair wage $\bar{w}$ is paid. According to Result 2, in one type of equilibrium *all* workers commit to reciprocal behavior. Moreover, there are equilibria in which some or even all employers are committed to punish unfair behavior, thus deterring uncommitted workers to exploit fair wage payments. In sum, the results confirm the fair wage-effort hypothesis (1) at least for some, if not all, productive units. In the following, it is argued that our model is also capable to explain reciprocal behavior in anonymous laboratory experiments of one-shot games.

A basic premise of the above model is that agents can signal their behavioral commitment to reciprocate. One may thus object that in laboratory experiments communication (i.e. an exchange of signals) is usually precluded. So how can the model explain the observed reciprocal behavior in experimental labor markets? The following interpretation of the results from these experiments suggests an answer to this question. For both an "employer" to pay a high wage and for a "worker" to provide full effort the players have to attribute a positive probability to the possibility that there partners have internalized the personal norm that one should reciprocate. Usually, the two groups of experimental subjects ("workers" and "employers") are located in different rooms and do not directly communicate through signals. Hence, wage offers of "employers" can only depend on their beliefs about the share of fair "workers" in the other room. Similarly, giving "employers" the possibility to punish unfair "workers" (as in Fehr et al., 1996), effort choices of uncommitted "workers" depend on their beliefs about the share of vengeance-seeking "employers". Experimental subjects form these beliefs from their everyday experience. The above analysis provides a rational basis for such beliefs. The experimental subjects thus act according to both their beliefs *and* their internalized personal norms. Deviating from the commitment to reciprocate may imply psychological costs (e.g. due to emotional dispositions) which outweigh potential material gains. In our model, this has been the reason why such a commitment could serve as self-binding in the first place. Moreover, given that a positive share of "employers" in experimental labor markets is committed to punish unfair behavior, even "workers" without a behavioral norm to reciprocate may nevertheless behave kindly in order to avoid punishment. [5]

6. Social learning versus rational choice of norms

Evidence from laboratory experiments not only suggests that a substantial fraction of individuals reciprocate in one-shot games, even though this implies immediate material losses, but also that behavioral attitudes are altered at some point of life. In a series of experiments, Hannan, Kagel, and Moser (1999) find that there are large differences in reciprocity between US undergraduates and MBA students. The extent of (costly) reciprocation in one-shot labor market games is found to be substantially higher in experiments with MBA students who are older on average and already have gathered work experience.

Standard economic modeling takes preferences as given. For instance, Becker and Stigler (1977) defend the usual assumption that preferences are stable and identical among individuals in arguing that both changes in prices and differences in income among individuals alone can explain even substantial changes in behavior and behavioral differences among individuals, respectively. In contrast, Hirschman (1984, p. 90), following the philosopher Frankfurt (1971), argues that "autonomous, reflective changes in values [...] do occur from time to time in the lives of individuals" and objects to Becker and Stigler (1977) insisting that "de valoribus est disputandum".

In line with the so-called "indirect evolutionary approach" (e.g. Güth, 1995; Güth & Yaari, 1992), this paper takes an intermediate view by hypothesizing stable (non-altruistic) meta-preferences (see (2)). In our model, individuals

[5] Of course, reciprocal behavior in one-shot encounters is not restricted to experimental *labor* markets. See e.g. Fehr and Schmidt (1999) for a survey of other experiments with a sequential game structure.

consciously choose their type (i.e. behavioral norm) by taking into account these meta-preferences. Formally, this choice is reflected in their utility functions (see (4a) and (4b)), in turn determining behavior in the subsequent game. Also in the indirect evolutionary approach, individuals behave according to a utility function which depends on a certain type (or norm, respectively), however, without considering whether their behavior also serves their meta-preferences (which are represented by so-called "evolutionary success functions"). Rather, natural or cultural selection (i.e. evolutionary success) eventually determines one's type in this kind of model.[6] This process is then interpreted as social learning of behavioral attitudes, substituting the choice whether or not to commit to a self-binding personal norm pursued in our approach. In fact, evolutionary approaches do not (have to) rationalize why individuals do not deviate from a norm when deviation would be beneficial regarding their meta-preferences, simply because of the "bounded-rational" behavior of not considering them. In contrast, our approach suggests that emotional dispositions are *necessary* to sustain a norm in a concrete situation, and that it may be rational to adopt such dispositions. This has been formalized as sufficiently high costs of deviating from an adopted behavioral norm, as reflected in (4a) and (4b).

Frank (1988) argues that the term "rational choice" in such a context can be understood as self-reflection which eventually drives emotional dispositions. Human-beings would have the inherited cognitive skills to form habits and norms through a self-reflection process. (Admittedly, here this process has been formalized in a very stylized way.) In fact, a non-altruistic person could *wish* to be motivated by emotions (like feeling guilty or angry) and could consciously choose a kind of conditioning and social environment which raises the probability to be emotionally disposed. Obviously, this view is very different from evolutionary learning, in which individuals are eventually "programmed" to neglect their non-altruistic meta-preferences.[7]

Alternatively to both views, one can argue that parents, teachers etc. play an important role in shaping value-based personal norms. As Hoffman (1984, p. 119) states: "If a person has internalized a moral norm, then when it is activated it is usually experienced as deriving autonomously from within the self. That is, cognitive dimensions of the norm, the associated affect (guilt), and the disposition to act in accord with the norm are experienced as self-generated. The original source of the norm, for example, the parent, has lost most or all of its motive force and may be forgotten". However, even if parents play an important role in the development of

[6] In contrast to the indirect evolutionary approach, behavior in "direct" evolutionary games (e.g. Höffler, 1999) is determined without utility maximization, i.e., directly by the type of individuals. Again, types can be observed by individuals and evolve, for instance, through imitation of successful types ("replicator dynamics"). (For a discussion, see Königstein & Müller, 1999.)

[7] Whether or not preferences or norms, respectively, can be consciously chosen to some extent is a lively debate among philosophers (e.g. Moody-Adams, 1990; Nagel, 1979). According to Frankfurt (1987, p. 38), "what the person really wants [...] is incorporated into himself by virtue of the fact that he has it *by his own will*" (italics original).

personal norms, a reflective evaluation of these norms by adolescents or adults may still occur.

7. Summary

This paper has shown that the adoption of a self-sustained behavioral norm that one should be *kind to kind persons and unkind to unkind persons* can be understood as rational choice of non-altruistic individuals. Two assumptions are crucial for this result. First, signaling about personal norms is possible (although doubtlessly imperfect), e.g. through emotions (e.g. Frank, 1987, 1988; Hirshleifer, 1987). Second, the commitment to reciprocal behavior is feasible (i.e. self-binding) due to psychological costs, outweighing the material benefits, of deviating from behavioral norms.

Reciprocal behavior in employer–employee relationships have become a stylized fact in one-shot laboratory experiments. Based on our theoretical results, the following suggestions have been made to explain this behavior. First, in these experiments (as well as in real life) workers may provide high effort either because they have chosen to behave reciprocally fair or because they fear to work with employers who have chosen to punish unkind behavior. Second, subjects in experimental labor markets hold beliefs about the share of the population who consistently reciprocate (on basis of everyday experience), and realize that their exchange partners will stick to their norms even in a one-shot game.

The theoretical approach in this paper should be viewed as alternative to "social learning" theories, which are formalized in evolutionary games. In these models, natural or cultural selection substitutes rational choices of self-sustained behavioral norm pursued in this paper. Both kinds of forces may be important to understand why people reciprocate even in one-shot situations, but are unlikely to be the only explanations. This should be a bright prospect for future research.

Acknowledgements

I'm especially grateful to Josef Falkinger, Simon Gächter, the editor Simon Kemp of this Journal and two anonymous referees for very valuable and detailed comments. I'm also indebted to Wolfgang Buchholz, Ernst Fehr, Martin Summer and seminar participants at the University of Regensburg for helpful discussions. Of course, all errors are my own.

Appendix A

In this appendix, formal propositions about reciprocity equilibria with uniform signal distributions, which are summarized and discussed in Section 4, are stated and proven. First, the expected utility functions (15), (17), (21) and (22) are given in rewritten form. Using (16), (18), (19) and (20), respectively, we obtain the following expressions:

$$E_C^W = (\alpha\beta - 1)(\bar{Q}_C + (\underline{Q}_C - \bar{Q}_C)(t_C\underline{e} + 1 - t_C) + (1 - \underline{Q}_C)\underline{e}), \tag{A.1}$$

$$\begin{aligned} E_N^W = {} & \overline{Q}_N(t_C(1 - R_C)(\alpha\beta - \underline{e} - \bar{p}) + (1 - t_C)(1 - R_N)(\alpha\beta - \underline{e}) \\ & + (1 - t_C(1 - R_C) - (1 - t_C)(1 - R_N))(\alpha\beta - 1)) \\ & + (\underline{Q}_N - \overline{Q}_N)(t_C\underline{e}(\alpha\beta - 1) + (1 - t_C)(\alpha\beta - \underline{e})) + (1 - \underline{Q}_N)\underline{e}(\alpha\beta - 1), \end{aligned} \tag{A.2}$$

$$\begin{aligned} E_C^F = {} & R_C\beta(1 - \alpha) + (1 - R_C) \cdot (s_C\bar{Q}_C\beta(1 - \alpha) + (1 - s_C)\bar{Q}_N(\beta(\underline{e} - \alpha) - \bar{p}) \\ & + (1 - s_C\bar{Q}_C - (1 - s_C)\bar{Q}_N)\underline{e}\beta(1 - \alpha)), \end{aligned} \tag{A.3}$$

$$\begin{aligned} E_N^F = {} & R_N\beta(1 - \alpha) + (1 - R_N) \cdot (s_C\underline{Q}_C\beta(1 - \alpha) + (1 - s_C)\underline{Q}_N\beta(\underline{e} - \alpha) \\ & + (1 - s_C\underline{Q}_C - (1 - s_C)\underline{Q}_N)\underline{e}\beta(1 - \alpha)). \end{aligned} \tag{A.4}$$

We are now ready to derive the following propositions.

Proposition 1. *$s_C \in [0, \alpha)$ cannot be part of a reciprocity equilibrium.*

Proof. If $0 \leqslant s_C < \alpha$, then (A.1), (A.2) and (26a) imply $E_C^W = (\alpha\beta - 1)(\bar{Q}_C + (1 - \bar{Q}_C)\underline{e})$ and $E_N^W = \underline{e}(\alpha\beta - 1)$, where $\bar{Q}_C = a(z_C^f - z_N^f) \in (0, 1)$ according to (26a). Thus, $E_C^W > E_N^W$ if and only if $(\alpha\beta - 1)\bar{Q}_C(1 - \underline{e}) > 0$ which is fulfilled. This confirms Proposition 1. □

Proposition 2. *$s_C \in [\alpha, 1)$ and $t_C \in ((1 - \underline{e})/\bar{p}, 1]$ cannot be a reciprocity equilibrium.*

Proof. Proposition 2 is proven in two steps.

Step 1. If $\alpha \leqslant s_C < \tilde{s}_C$, it is easy to show by using (A.3), (A.4), (26b) and (28b) that $E_C^F > E_N^F$. In this case t_C would equal unity. However, one can see by using (A.1), (A.2), (26b) and (28b) that $t_C = 1$ would imply $E_C^W > E_N^W$. Thus, in this case $s_C \in [\alpha, \tilde{s}_C)$ cannot be a reciprocity equilibrium.

Step 2. If $\tilde{s}_C \leqslant s_C < 1$, using (A.3), (A.4), (26c) and (28b) yields $E_C^F > E_N^F$. This would imply $t_C = 1$. Moreover, if $t_C = 1$, then $E_C^W > E_N^W$ according to (A.1), (A.2), (26c) and (28b). However, this would imply $s_C = 1$ which is a contradiction. This confirms Proposition 2. □

Proposition 3. (*i*) *Suppose $\alpha\beta(1 - \underline{Q}_N) \leqslant 1$, where $\underline{Q}_N$ is given in* (26b). *If*

$$\hat{s}_C \equiv \frac{(1 - \alpha)R_C + \alpha\underline{Q}_N}{(1 - \alpha)\left(1 - \bar{Q}_C(1 - R_C)\right) + \alpha\underline{Q}_N} \in [\alpha, \tilde{s}_C),$$

where $\bar{Q}_C$ *and* R_C *are given in* (26b) *and* (28a), *respectively*, *and*

$$\hat{t}_C \equiv \frac{1 - \alpha\beta(1 - \underline{Q}_N)}{\bar{Q}_C(\alpha\beta - 1) + 1 - \alpha\beta(1 - \underline{Q}_N)} \in [0, (1 - \underline{e})/\bar{p}],$$

then $s_C^* = \hat{s}_C$ *and* $t_C^* = \hat{t}_C$ *is a reciprocity equilibrium.* (*ii*) *If* $\alpha\beta(1 - \underline{Q}_N) > 1, s_C \in [\alpha, \tilde{s}_C)$ *and* $t_C \in [0, (1 - \underline{e})/\bar{p}]$ *cannot be a reciprocity equilibrium.*

Proof. Consider the case $\alpha \leqslant s_C < \tilde{s}_C$ and $0 \leqslant t_C \leqslant (1 - \underline{e})/\bar{p}$. (i) Using (A.1), (A.2), (26b) and (28a) one obtains after some tedious but straightforward manipulation that $E_C^W >, =, < E_N^W$ if and only if $t_C >, =, < \hat{t}_C$. Similarly, using (A.3), (A.4), (25b) and (27a) reveals $E_C^F >, =, < E_N^F$ if and only if $s_C <, =, > \hat{s}_C$, where $\underline{Q}_N$, $\bar{Q}_C$ and R_C are given in (26b) and (28a), respectively. For $\alpha\beta(1 - \underline{Q}_N) \leqslant 1$, we have $\hat{t}_C \geqslant 0$. Also note that $\hat{s}_C \in (0, 1)$. Thus, it is possible that $\hat{t}_C \in [0, (1 - \underline{e})/\bar{p}]$ and $\hat{s}_C \in [\alpha, \tilde{s}_C)$ simultaneously hold. (ii) If $\alpha\beta(1 - \underline{Q}_N) > 1$, then $\hat{t}_C < 0$. But this means $E_C^W > E_N^W$ for any $t_C \geqslant 0$. This concludes the proof. □

Proposition 4. *If*

$$\bar{t}_C \equiv \frac{(1 - \underline{e})\Big(1 - \alpha\beta(1 - \bar{Q}_N)\Big)}{(1 - \underline{e})R_C + \bar{p}(1 - R_C)} \leqslant (1 - \underline{e})/\bar{p},$$

where $\bar{Q}_N$ *is given in* (26c) *and* R_C *is given in* (27a), *then* $s_C^* = 1$ *and any* $t_C^* \in [\max\{0, \bar{t}_C\}, (1 - \underline{e})/\bar{p}]$ *is a reciprocity equilibrium.*

Proof. If $s_C = 1$ and $0 \leqslant t_C \leqslant (1 - \underline{e})/\bar{p}$, then using (A.1), (A.2), (26c) and (28a) reveals $E_C^W >, =, < E_N^W$ if and only if $t_C >, =, < \bar{t}_C$. One can show that $\bar{t}_C \leqslant (1 - \underline{e})/\bar{p}$ if and only if $\alpha\beta(1 - \bar{Q}_N) \geqslant (1 - (1 - \underline{e})/\bar{p})R_C$. Also note that if $\alpha\beta(1 - \bar{Q}_N) > 1$, then $\bar{t}_C < 0$ and thus $E_C^W > E_N^W$ for all $t_C \in [0, (1 - \underline{e})/\bar{p}]$. If $s_C = 1$, then $E_C^F = E_N^F$ according to (A.3), (A.4), (26c) and (28a) which concludes the proof. □

Proposition 5. *If*

$$\bar{\bar{t}}_C \equiv \frac{\bar{Q}_N(1 - R_N)(1 - \underline{e}) - (\alpha\beta - 1)\Big(1 - \underline{e}(1 - \bar{Q}_N)\Big)}{\bar{Q}_N(1 - R_N)(1 - \underline{e})} > \frac{1 - \underline{e}}{\bar{p}},$$

where $\bar{Q}_N$ *and* R_N *are given in* (26c) *and* (28b), *respectively*, *then* $s_C^* = 1$ *and any* $t_C^* \in [\bar{\bar{t}}_C, 1)$ *is a reciprocity equilibrium.*

Proof. If $s_C = 1$ and $(1 - \underline{e})/\bar{p} < t_C \leqslant 1$, then $E_C^F = E_N^F$ for all t_C within the presumed interval according to (A.3), (A.4), (26c) and (28b). Moreover, using (A.1), (A.2), (26c) and (28b) one obtains $E_C^W >, =, < E_N^W$ if and only if $t_C >, =, < \bar{\bar{t}}_C$, and $\bar{\bar{t}}_C > (1 - \underline{e})/\bar{p}$ may hold. Also note that $\bar{\bar{t}}_C < 1$. This confirms the result. □

Proposition 6. *Suppose that* $\alpha\beta(1-\bar{Q}_N) \in [(1-(1-\underline{e})/\bar{p})R_C, 1]$. (*i*) *If* $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) = (1-R_C)\bar{Q}_N\bar{p}$, *where* $\bar{Q}_N$ *and* R_C *are given in* (26c) *and* (28a), *respectively, then* $t_C^* = \bar{t}_C$ *and any* $s_C^* \in [\tilde{s}_C, 1)$ *is a reciprocity equilibrium.* (*ii*) *If* $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) < (1-R_C)\bar{Q}_N\bar{p}$ *and* $\alpha\beta(1-\bar{Q}_N) = 1$, *then* $t_C^* = 0$ *and any* $s_C^* \in [\tilde{s}_C, 1)$ *is a reciprocity equilibrium.*

Proof. Note that $\alpha\beta(1-\bar{Q}_N) \in [(1-(1-\underline{e})/\bar{p})R_C, 1]$ implies $\bar{t}_C \in [0, (1-\underline{e})/\bar{p}]$. If $\tilde{s}_C \leqslant s_C < 1$ and $0 \leqslant t_C \leqslant (1-\underline{e})/\bar{p}$, one obtains $E_C^F >, =, < E_N^F$ if and only if $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) >, =, < (1-R_C)\bar{Q}_N\bar{p}$ for all $s_C \in [\tilde{s}_C, 1)$ according to (A.1), (A.2), (26c) and (28a). (i) If $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) > (1-R_C)\bar{Q}_N\bar{p}$, then there is no reciprocity equilibrium in this case. If $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) = (1-R_C)\bar{Q}_N\bar{p}$, then $t_C^* = \bar{t}_C$ and any $s_C^* \in [\tilde{s}_C, 1)$ is a reciprocity equilibrium. (ii) If $R_C\beta(1-\underline{e})(1-\alpha(1-\bar{Q}_N)) < (1-R_C)\bar{Q}_N\bar{p}$, then one obtains reciprocity equilibria only in the special case in which $\bar{t}_C = 0$ holds, i.e. if and only if $\alpha\beta(1-\bar{Q}_N) = 1$, where $\bar{Q}_N$ is given by (26c). In this case $t_C^* = 0$ and any $s_C^* \in [\tilde{s}_C, 1)$ is a reciprocity equilibrium. □

Proposition 7. $s_C^* = 1$ *and* $t_C^* = 1$ *is always a reciprocity equilibrium.*

Proof. Remember that in case of $s_C = 1$, we have $E_C^F = E_N^F$ for all $t_C \in ((1-\underline{e})/\bar{p}, 1]$ according to the proof of Proposition 5. Thus, it remains to show that if $t_C = 1$, then s_C would equal unity. In fact, using (A.1), (A.2), (26c) and (28b), it is easy to see that $E_C^W > E_N^W$ if $t_C = 1$. □

References

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal Social Psychology*, *67* (5), 422–436.

Agell, J., & Lundborg, P. (1995). Theories of pay and unemployment: Survey evidence from Swedish manufacturing firms. *Scandinavian Journal of Economics*, *97* (2), 295–307.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *97* (4), 543–569.

Akerlof, G. A., & Yellen, J. L. (1988). Fairness and unemployment. *American Economic Review Papers and Proceedings*, *78* (2), 44–49.

Akerlof, G. A., & Yellen, J. L. (1990). The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics*, *105* (2), 255–283.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Becker, G. S., & Stigler, G. (1977). De gustibus non est disputandum. *American Economic Review*, *67* (2), 76–90.

Bewley, T. F. (2000). *Why wages don't fall during a recession*? Cambridge, MA: Harvard University Press.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

Bolton, G. E., & Ocksenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90* (1), 166–193.

Dufwenberg, M., & Kirchsteiger, G. (1998). A theory of sequential reciprocity. University of Stockholm, Discussion Paper 1998:1.

Elster, J. (1996). Rationality and the emotions. *Economic Journal*, *106* (September), 1386–1397.

Falk, A., & Fischbacher, U. (1999). A theory of reciprocity. Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, *14* (3), 159–181.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1996). Reciprocal fairness and noncompensating wage differentials. *Journal of Institutional and Theoretical Economics*, *152* (4), 608–640.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, *65* (4), 833–860.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, *108* (2), 437–459.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998). Gift exchange and reciprocity in experimental markets. *European Economic Review*, *42* (1), 1–34.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, *114* (3), 817–868.

Frank, R. H. (1987). If homo economicus could choose his own utility function, would he want one with a conscience? *American Economic Review*, *77* (4), 593–604.

Frank, R. H. (1988). *Passions within reason. The strategic role of the emotions*. New York: W.W. Norton and Company.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5–20.

Frankfurt, H. G. (1987). Identification and wholeheartedness. In F. Schoeman (Ed.), *Responsibility, Character, and the Emotions* (pp. 27–45). Cambridge: Cambridge University Press.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. Cambridge, MA: MIT Press.

Gale, J., Binmore, K., & Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, *8* (1), 56–90.

Gouldner, A. W. (1960). The norm of reciprocity. *American Sociological Review*, *25*, 161–178.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory*, *24* (4), 323–344.

Güth, W., & Yaari, M. E. (1992). Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In U. Witt (Ed.), *Explaining process and change: Approaches to evolutionary economics* (pp. 23–34). Ann Arbor, MI: University of Michigan Press.

Hannan, R. L., Kagel, J. H., & Moser, D. V. (1999). Partial gift exchange in experimental labor markets: Impact of subject population differences and effort requests on behavior, University of Pittsburgh (mimeo).

Hirschman, A. O. (1984). Against parsimony: Three easy ways of complicating some categories of economic discourse. *American Economic Review Papers and Proceedings*, *74* (2), 89–96.

Hirshleifer, J. (1987). On the emotions as guarantors of threats and promises. In J. Dupré (Ed.), *The latest on the best essays on evolution and optimality* (pp. 307–326). Cambridge, MA: MIT Press.

Höffler, F. (1999). Some play fair, some don't. Reciprocal fairness in a stylized principal–agent problem. *Journal of Economic Behavior and Organization*, *38* (1), 113–131.

Hoffman, M. L. (1984). Parent discipline, moral internalization, and development of prosocial motivation. In E. Staub et al. (Eds.), *Development and maintenance of prosocial behavior* (pp. 117–137). New York: Plenum Press.

Homans, G. C. (1961). *Social behavior: Its elementary forms*. New York: Hartcourt.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, *76* (4), 728–741.

Königstein, M., & Müller, W. (1999). Combining rational choice and evolutionary dynamics: The indirect evolutionary approach. *Metroeconomica*, *51* (3), 235–256.

Levine, D. I. (1993). Fairness, markets, and ability to pay: Evidence from compensation executives. *American Economic Review*, *83* (5), 1241–1259.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1* (3), 593–622.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, *12* (1), 380–391.

Moody-Adams, M. (1990). On the old saw that character is destiny. In O. Flanagan, & A. Rorty (Eds.), *Identity, character and morality* (pp. 111–131). Cambridge, MA: MIT Press.

Nagel, T. (1979). *Mortal questions*. Cambridge, MA: Cambridge University Press.

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

Schwartz, S. H., & Howard, J. A. (1984). Internalized values as motivators of altruism. In E. Staub et al. (Eds.), *Development and maintenance of prosocial behavior* (pp. 229–255). New York: Plenum Press.

Sugden, R. (1984). Reciprocity: The supply of public goods through voluntary contributions. *Economic Journal*, *94* (December), 772–787.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.