Essays in behavioral economics

Essays in Behavioral Economics

Alexander Sebald

ISBN: 978-90-5278-752-7

Essays in Behavioral Economics

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
prof. mr. G.P.M.F. Mols
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen op
donderdag 25 september 2008 om 12.00 uur

door

Alexander Christopher Sebald

UNIVERSITAIRE
PERS MAASTRICHT

Promotores:
Prof. Dr. Georg Kirchsteiger
Prof. Dr. Arno Riedl

Copromotores:
Dr. Markus Walzl

Beoordelingscommissie:
Prof. Dr. Jean-Jacques Herings (voorzitter)
Prof. Dr. Mathias Dewatripont
Prof. Dr. Victor Ginsburgh
Dr. Ronald Peeters

To my parents

Table of Contents

Acknowledgments

Writing a doctoral thesis is never an endeavor that one accomplishes by oneself. Discussions with supervisors, co-authors, friends and family are as (if not more) important as the own contribution. Therefore, I would like to thank those that have been close to me in all those years and have accompanied me on this adventure.

First, I would like to express my deep gratitude to Georg who supported me from the first minute that I fearlessly walked into his office. The fact that his door was always open, he read and re-read every single paper that I worked on and he loves coffee as much as I do gave me an invaluable sense of security.

Thanks a lot also to my other supervisors and co-authors Arno, Markus, Martin, Charles and Gani. Our discussions were / are immensely important to me and I honestly hope that we will continue our inspiring collaboration.

Third, I would like to thank the 'Bruxelles-Gang' for the wonderful time that we spend together at ECARES. When I arrived in 2004, you made it incredibly easy for me to integrate and feel at ease in the new environment. I will always remember our football matches on the corridor that helped to briefly relax from seemingly irresolvable problems. In particular, I would like to thank Paolo with whom I have spend an amazing amount of invaluable time in the last few years. I learned a lot from your well thought-out opinions and your incredible knowledge about Jazz!

Even though further away, also Martin, Clivia, Anna, Andre, Simone, Andrea, Nils, Julia, Verena and many others were tremendously important to me in all this time. Our discussions have always helped me a lot to question myself and what I was doing. Honestly, I had never imagined that it would be so difficult for me to talk about my work, my research projects and ideas using 'normal' words that are also comprehensible to people less familiar with the literature. I thank you for your patience in all those situations in which I got lost in my own explanations. In particular, I would like to thank Andre and Simone with whom I share deep and confiding friendships that started in Maastricht. Andre, I am truly glad that we met between the supermarket shelfs of the Edah in 1997 and have been able to cultivate our friendship ever since. Simone, I very much hope that we always manage to live in the same city from time to time.

Furthermore, I would like to thank my parents and my sister who always greatly supported me. Without your help and trust I would not have been able to accomplish this thesis.

Last but not least, I would like to thank my girlfriend Sophia who had the strength and endurance to continuously overcome the distance between Cologne and Bruxelles. I am unbelievably happy to have you and am looking forward to our new life in Copenhagen.

Introduction

Traditionally economic theory is based on very narrow presumptions about human behavior. It is essentially assumed that people only care about their own monetary payoff, or, in other words, that people are selfish. However, in the last 20 years experimental research has accumulated overwhelming evidence that is at odds with this classical model of human behavior. It has been shown that people very often care about the distributional consequences of their actions as well as underlying motives and intentions. As a consequence the model of human behavior has been substantially widened. Models of distributional concerns as well as belief-dependent models of reciprocity, guilt aversion, regret and shame have been conceptualized. Against the background of the experimental findings and the associated new models of human behavior, the question arises whether the broadening of the behavioral presumptions impacts the conclusions drawn on the basis of our classical model. This question is the central point of my dissertation. More precisely, I study in four different papers the impact of these broader models of human behavior on decision making and human interactions.

In the first paper, 'Investments into education - Doing as the parents did' (with Georg Kirchsteiger), we study the impact of indirect reciprocity on the efficiency of private investments into human capital. The starting point of this project is empirical evidence suggesting that parents act indirectly reciprocal toward their children. Indirect reciprocity in this context means that parents that have received a lot (little) from their parents tend to give also a lot (little) to their children. More specifically, the paper focuses on parental investments into the education of their children, i.e. parents that have received a lot of education financed for by their parents do the same for their children and vice versa. This indirectly reciprocal behavior implies an intergenerational chain transmitting the attitude towards the formation of human capital from one generation to the next. In this paper we incorporate this 'chain' into an overlapping generations model with endogenous human capital formation and show that in absence of any state intervention such an economy might be characterized by multiple steady states. Interestingly, temporary public investments into human capital formation can move the economy from a steady state with low human capital levels to one with higher human capital levels. Nevertheless, even the best steady state that can be reached by temporary public investments is suboptimal when human capital is privately provided in the long run. This inefficiency can only be overcome by a permanent public subsidy for education. The analysis, hence, presents another good reason for government

intervention to support optimal private investments into the education of children.

The second project, 'Procedural Concerns and Reciprocity', concentrates on another issue concerning the model of human behavior in economic theory. 'Procedural concern' is a well established concept in psychology. Sparked by experimental evidence, economists have only lately started to ask the question why people often behave very differently in outcome-wise identical situations depending on the ways, i.e. procedures, which have led to them. In this second project I present a framework which allows to account for procedural concerns in economic analyzes. More specifically, building on Martin Dufwenberg and Georg Kirchsteiger (2004)'s 'theory of sequential reciprocity', I show how procedural concerns can be conceptualized assuming that agents are (also) motivated by belief-dependent reciprocal preferences. Already during my work on 'Procedural Concerns and Reciprocity' I came to the conviction that it actually represents only one step in a bigger theory. Reciprocity is only one type of motive through which procedural concerns can be rationalized.

In my job market paper, 'Procedural Concerns in Psychological Games', I generalize this idea to show that in the presence of all kinds of belief-dependent utilities (guilt, reciprocity, regret etc) procedural concerns arise. Hence, building on my second project, I generalize in this third project the results regarding procedural concerns to all kinds of belief-dependent motivations and demonstrate how the interaction of agents with belief-dependent psychological payoffs is influenced by procedural choices. More specifically, I use Martin Dufwenberg and Pierpaolo Battigalli (2007)'s framework of 'dynamic psychological games' and show that procedural concerns cannot only be conceptualised assuming reciprocal preferences, but inherently arise in the interaction of agents with all kinds of belief-dependent motivations. One of the main contributions, in my view, is the way that I define procedures and formalize 'procedural games' in which agents do not choose actions and strategies, as traditionally assumed in game theory, but procedures. I show that outcomes and procedures are inherently connected but nevertheless play distinct roles in the interaction of agents with belief-dependent utilities. In the context of the procedural games I clearly separate procedural choices from outcomes which allows to isolate the impact that procedural choices have on the strategic interaction of agents.

Lastly, in the paper 'How (too much) self esteem facilitates contracts with subjective evaluations' (with Markus Walzl) we analyze the impact of aggressive reactions to ego-threatening feedback on principal-agent relationships. More specifically, we show how peoples' desire to protect their self-esteem can explain

the existence of contractual relationships in environments with unobservable effort and subjective measures of performance. We concentrate on situations in which performance can only be measured subjectively as these constitute exactly the settings in which disagreements about effort and performance arise. This project is closely related to the recent works on self-esteem by Jean Tirole and Roland Bénabou (2002) and contracts with subjective performance signals by Bentley MacLeod (2003).

All in all, as said in the beginning, all papers analyze the implications of a broader model of human behavior in economic theory. It can be concluded that allowing for more complex human behavior in economic analyzes greatly impacts and alters conclusions that have been drawn on the basis of classical presumptions.

References

1. Battigalli, P. and Dufwenberg, M. (2007), Dynamic Psychological Games, forthcoming Journal of Economic Theory.

2. Bénabou, R. and Tirole, J. (2002), Self-Confidence and Personal Motivation, Quarterly Journal of Economics, 117(3), 871-915.

3. Dufwenberg, M. and Kirchsteiger, G. (2004), A theory of sequential reciprocity, Games and Economic Behavior, 47, 268-298.

4. MacLeod, B. (2003), Optimal Contracting with Subjective Evaluation, American Economic Review, 93(1), 216-240.

Investments into Education
-
Doing as the parents did

(with Georg Kirchsteiger)

Investments into education - Doing as the parents did*

Georg Kirchsteiger† and Alexander Sebald ‡

Abstract

Empirical evidence suggests that parents with higher levels of education generally attach a higher importance to the education of their children. This implies an intergenerational chain transmitting the attitude towards the formation of human capital from one generation to the next. We incorporate this intergenerational chain into an OLG-model with endogenous human capital formation. In absence of any state intervention such an economy might be characterized by multiple steady states with low or high human capital levels. There are also steady states where the population is permanently divided into different groups with differing human capital and welfare levels. Depending on the parameters of the model, a temporary or permanent public investment into human capital formation is needed to overcome steady states with low human capital and welfare levels. Furthermore, even the best steady state is suboptimal when the human capital is privately provided. This inefficiency can be removed by a permanent public subsidy for education.

Keywords: Human Capital Formation, Education Subsidy, Indirect Reciprocity.
JEL Classification: H23, H52, I2.

1 Introduction

In modern economies human capital is one of the most important determinants of economic progress and welfare. In contrast to the investment into physical capital the formation of human capital is to a large extent not financed by its owner. Rather, parents and the state cover most of the expenditures on education. The parental engagement has traditionally been explained by credit market

* We are grateful to Monika Bütler and to seminar participants at the Universities of Essex, Constance and Maastricht, and at ECARES/ULB for helpful comments.

† ECARES, Université Libre de Bruxelles, Avenue F D Roosevelt 50, CP114 1050 Brussels, Belgium, CEPR, and CESifo. Kirchsteiger is also member of ECORE, the recently created association between CORE and ECARES. E-mail: gkirchst@ulb.ac.be

‡ Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands, and ECARES, Université Libre de Bruxelles. E-mail: a.sebald@algec.unimaas.nl

imperfections, parental altruism (see e.g. [4] and [5]) and/or an exchange between education expenditures for the children and old-age support for the parents (see e.g. [7] and [6]).

Parental altruism is traditionally assumed to be exogenously given in economic theory, neglecting its source and evolutionary development. Among biologists and social-psychologists, on the other hand, there exists by now a large consensus that preferences, norms and cultural attitudes are endogenous with respect to our socio-economic system (see also [3], [8], [11], [16] and [19]). It is argued that two main channels exist through which preferences are transmitted across generations. Preferences are passed on genetically and/or through a process of socialization whereby e.g. children adopt parental preferences by means of imitation.

One area where the transmition of preferences through socialisation / imitation has been found particularly important is 'the attitude to education'. According to the socio-psychological literature parents have a pervasive influence in shaping young people's attitudes to education (see e.g. [24], [9] and [10]). More precisely, parents with higher levels of education transmit a more positive attitude towards education to their children (see e.g. [24]).[1]

This intergenerational transmission of attitudes can be viewed as an example of indirect reciprocity, which has been found to be particularly important within family relations (see e.g.[1], [2] and [20]). In contrast to direct reciprocity (see e.g. [13], [26]), we speak of indirect reciprocity when a person does not directly reciprocate to the behavior of another person, but rather reciprocates indirectly to a third party (see e.g. [1], [22], [23] and [15]). In the context of education financing this means people do not directly reciprocate for the education they have received from their own parents, but rather repay it by financing the education of their children. Hence, the more education parents have received themselves, the more they are willing to finance the education of their children. In this way investments into human capital do not only affect the immediate recipient, i.e. the next generation, but also future generations.

The intergenerational transmition of attitudes is in line with the empirical fact that for given family income, higher educated parents tend to spend more on the education of their children than parents with lower education (see [21]). Traditionally this has been explained by the so called 'home environment externality' [17], which states that not only private and public investments into education, but also innate abilities and the 'family environment' determine human capital formation. This strand of literature (see e.g. [5], [14], [17] and [18]) assumes that children's ability to acquire human capital depends on parental levels of education. Higher levels of parental education are assumed to increase the marginal product of investments into the human capital of children. Hence, the higher the level of education of the parents, the more effective investments in human capital become. If parents care about their children, this 'home environment externality' can explain the effect of parents' human capital on the education expenditures. If such a 'home environment externality' exists and parents only care about the

[1] A similar intergenerational attitude transmission mechanism has been analysed in the context of arts education (see [12])

educational level of their offsprings, Eckstein and Zilcha [14] also show that private investments into human capital are suboptimal. In their analysis the source of suboptimality is twofold. First, parents do not take into account the impact of their investment into the education of their children on their children's wages. Second, they do not take into account their impact on the relative effectiveness of their children's investments into the education of their grandchildren.

In contrast to the 'home environment externality', the intergenerational transmission of attitudes implies that parents directly affect children's preferences, rather than their production of human capital. Parents' preference for the human capital of their children depends on their own human capital, which was financed for by the grandparents. Our paper investigates the impact of this intergenerational chain on welfare and the optimal education policy. Using an OLG model we show that multiple steady states might exist. There always exists an illiterateness steady state, which is characterized by low incomes and no investments into formal education. Depending on the parameters of the model, a temporary or permanent public funding of education could be necessary to overcome this 'bad' steady state and to get the economy into a 'good' steady state with investments into formal education and higher welfare. Depending on the initial conditions there also exist steady states where the population is permanently split into a group with large human capital endowment and high welfare and a group with low human capital and welfare level. Again a temporary or permanent subsidy is necessary to overcome such a situation. Furthermore, even the best steady state is suboptimal, since the model investigated exhibits an externality. It is shown how a permanent, tax financed subsidy on human capital acquisition can internalize this inefficiency. The paper is organized as follows. In the next section we describe the model, followed by a characterization of the economy with private investments into human capital. In section 4 we analyze the welfare properties of this economy. Finally, we draw conclusions. All the proofs are delegated to the appendix.

2 The model

We assume a competitive economy, in which the output in period t, Y_t, does not only depend on physical capital used in t, K_t, and on labour L_t, but also on human capital, H_t. The economy is endowed with a Cobb-Douglas production technology. The normalized production function is given by

$$y_t = k_t^{\alpha} h_t^{1-\alpha} \tag{1}$$

where $\alpha \in (0,1)$ and $y_t = \frac{Y_t}{L_t}$, $k_t = \frac{K_t}{L_t}$, $h_t = \frac{H_t}{L_t}$. y_t denotes the output per worker in period t , and k_t and h_t are respectively physical and human capital per worker in t.

Every worker supplies inelastically one unit of labour, and for simplicity the number of workers is constant over time, i.e. $L_t = L$ for all t. Markets are assumed to be perfectly competitive, so that factors earn their marginal product:

$$r_t = f_k^{'}(k_t, h_t) = \alpha \left(\frac{h_t}{k_t}\right)^{1-\alpha} \tag{2}$$

$$w_t = f(k_t, h_t) - k_t f_k^{'}(k_t, h_t) = (1-\alpha) k_t^{\alpha} h_t^{1-\alpha} \tag{3}$$

with r_t being the interest rate and w_t the wage.

The capital stock depreciates fully in one period, so that the savings in period $t-1$ equal the capital stock in period t.

Human capital is produced by formal education, i.e. by schooling. We assume however, that even without any formal education everyone acquires some minimum human capital. We normalize human capital such that the minimum human capital is one. Human capital production is given by

$$h_{t+1} = (e_t)^{\beta} + 1 \tag{4}$$

with $\beta \in (0,1)$. $e_t \geq 0$ denotes the private expenditures into the formal education of a child born in t. Of course, the resulting human capital becomes productive in period $t+1$.

At each point in time three overlapping generations are alive in the economy.

Generation	Period		
	$t-1$	t	$t+1$
(1)	Education		
(2)		Work	
(3)			Retirement

Take a representative individual born at the beginning of period $t-1$. In this period he belongs to the youngest generation 1 which gets educated. The amount of his education is decided upon by his parent. In the next period t, the individual belongs to the working (parent) generation 2. In this period he works and has one child[2]. He divides his income between consumption in period t, savings for consumption in $t+1$ and spending for the education of his child. In period $t+1$, the individual belongs to the retired generation 3 and consumes his savings. At the end of this period, the individual dies.

Only the working generation has to make a decision. Individuals working in time t are assumed to maximize their utility function given by

$$U(c_{2,t}, c_{3,t+1}, h_{t+1}) = \ln c_{2,t} + \gamma \ln c_{3,t+1} + \varphi(h_t) \ln h_{t+1}, \tag{5}$$

where $c_{2,t}$ denotes the immediate consumption of an individual working in period t. $c_{3,t+1}$ is the consumption in the next period $t+1$ when the individual belongs to the retired generation 3. Since we assume full depreciation of the capital stock in one period, the savings in period t are the capital stock in period $t+1$, and the old generation only consumes the interest on their savings. Therefore, $c_{3,t+1} = k_{t+1} r_{t+1}$. h_{t+1} is the human capital of the child, which becomes effective in period $t+1$. γ and φ measure the individual's attitude towards future old-age consumption and towards the human capital of the child, respectively.

[2]For simplicity we assume that each adult has only one child, and each child has only one parent.

As explained in the introduction, there exists a lot of evidence that the importance parents attach to the education of their children is determined by indirect reciprocity. More precisely, the education a parent has received in his own childhood shapes his willingness to invest into the human capital of his own child. In order to capture this, we introduce an attitude function:

$$\varphi : [1, \infty) \to \Re_+^0,$$

with $\varphi(h_t)$ denoting the attitude of a parent with a human capital of h_t.[3] We assume that $\varphi(h_t)$ is continuous and differentiable. If the parent has not received any formal education himself, he is not willing to finance any formal education of his child. Furthermore, his attitude towards his child's education is positively correlated with his own human capital h_t, which was financed for by his own parent. These considerations lead to

$$\varphi(1) = 0$$

and

$$\varphi'(h_t) > 0.$$

In the next section we characterize the economy with pure private investments into human capital.

3 Private investments into human capital

Agents working in period t have to decide how much of their wage income w_t they want to spend on instantaneous consumption and on the education of their child. Furthermore, they save in order to finance consumption when they are retired. Recall that due to full depreciation of the capital stock, $c_{3,t+1} = k_{t+1} r_{t+1}$. Recall also that $e_t = (h_{t+1} - 1)^{\frac{1}{\beta}}$.

The maximization problem of a representative agent working in t can be written as:

$$\begin{aligned} \max_{c_{2,t}, k_{t+1}, h_{t+1}} U(c_{2,t}, k_{t+1}, h_{t+1}) &= \ln c_{2,t} + \gamma \ln k_{t+1} r_{t+1} + \varphi(h_t) \ln h_{t+1} \\ \text{s.t. } w_t &= c_{2,t} + k_{t+1} + (h_{t+1} - 1)^{\frac{1}{\beta}} \\ h_{t+1} &\geq 1 \\ c_{2,t}, k_{t+1} &\geq 0 \end{aligned}$$

Denote by $\widetilde{k}_{t+1}$, $\widetilde{h}_{t+1}$ the utility maximizing choice of the agent working in period t, when the human capital for the next generation is provided privately. The sequence of utility maximizing choices is denoted by $\{\widetilde{k}_t, \widetilde{h}_t\}_{t=2}^{\infty}$, and k_1 and h_1 denote the initial endowments with physical and human capital. The solution is characterized by the following lemma.

[3] Recall that even without formal education each individual is endowed with a minimum human capital normalized to 1. Hence, φ is defined for human capital levels not below 1.

Lemma 1 *If $\widetilde{k}_t > 0$ it holds that:*

i) The solution $(\widetilde{k}_{t+1}, \widetilde{h}_{t+1})$ fulfills the first order conditions

$$\frac{\partial U}{\partial h_{t+1}} = \frac{\varphi(\widetilde{h}_t)}{\widetilde{h}_{t+1}} - \frac{\frac{1}{\beta}\left(\widetilde{h}_{t+1} - 1\right)^{\frac{1}{\beta}-1}}{\widetilde{w}_t - \widetilde{k}_{t+1} - \left(\widetilde{h}_{t+1} - 1\right)^{\frac{1}{\beta}}} = 0 \tag{6}$$

and

$$\frac{\partial U}{\partial k_{t+1}} = \frac{\gamma}{\widetilde{k}_{t+1}} - \frac{1}{\widetilde{w}_t - \widetilde{k}_{t+1} - \left(\widetilde{h}_{t+1} - 1\right)^{\frac{1}{\beta}}} = 0 \tag{7}$$

ii) $\widetilde{k}_{t+1} > 0$.

iii) If $\widetilde{h}_t = 1$, *then* $\widetilde{h}_{t+1} = 1$.

iv) If $\widetilde{h}_t > 1$, *then* $\widetilde{h}_{t+1} > 1$.

Proof: see Appendix.

If $k_1 = 0$, no production, no consumption, and no formal education is possible in any future period. Since this case is not interesting, we restrict the analysis from now on to $k_1 > 0$.

We show next that there exists no unlimited expansionary path.

Proposition 2 *There exists a triple h^m, k^m, w^m such that for any initial conditions $\widetilde{k}_1$ and $\widetilde{h}_1$ there exists a t^m such that:*

$$\begin{aligned} \widetilde{h}_t &< h^m \text{ whenever } t > t^m \\ \widetilde{k}_t &< k^m \text{ whenever } t > t^m \\ \widetilde{w}_t &< w^m \text{ whenever } t > t^m \end{aligned}$$

Proof: See Appendix

Next, we turn to the analysis of the existence and of the stability properties of steady states. We first analyze the benchmark case where the attitude towards the children's education does not depend on parents' education. Then we analyze the steady states for endogenous education attitudes.

3.1 Exogenous education attitude

As a benchmark we first analyze the situation where the attitude towards education is not determined by the attitude function $\varphi(h_t)$, but exogenously determined at level $\overline{\varphi} > 0$. In this case, there exists a unique interior steady state with $h^* > 1$.

Proposition 3 *If the attitude towards education is exogenously fixed at level $\overline{\varphi} > 0$, there exists a unique steady state with formal education, i.e. with $h^* > 1$.*

Proof: See Appendix

Simulations suggest that the steady state is globally stable. See for example Figure 1 in which we graphically report simulation results for an exogenous education attitude $\overline{\varphi} = 4$ and parameter values $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$.[4]

[Figure 1 here]

Each line constitutes the optimal path of human and physical depending on the initial conditions h_1 and k_1. As one can easily see, for all initial conditions of physical and human capital the system converges towards $h^* = 1.18542$, and physical capital $k^* = 0.10281$.[5] In other words, from any initial values of human and physical capital the system converges towards the steady state.

Repeating the same simulation exercise for different values of the attitude parameter $\overline{\varphi}$ and other parameter values α, β, γ leads to different steady states (h^*, k^*) with $h^* > 1$. In Table 1 we report the steady states h^* and k^* for $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$ with varying levels of the exogenous education attitude parameter $\overline{\varphi}$. Not surprisingly the steady state level of human capital increases in the exogenous education attitude $\overline{\varphi}$.

	Exog. Attitude: $\overline{\varphi}$	Human Capital: h^*	Physical Capital: k^*
1.	0.1	1.00007	0.1250
2.	0.5	1.0033	0.1253
3.	2	1.0707	0.1229
4.	4	1.1854	0.1028

Table 1: Steady state levels h^* and k^* for α=0.3, β=0.7, γ=0.5 and varying degrees of $\overline{\varphi}$

Also for these parameter values we conducted simulations showing convergence to the steady state. In all the simulations with an exogenous attitude towards the education of the children the system converges towards the unique interior steady state. Hence simulations suggest that the steady state with agents investing into the formal education of their children is globally stable. As we will see in the next subsection, this result is in sharp contrast to the model with endogenous education attitudes.

3.2 Endogenous education attitude

Going back to our model with endogenous education attitude, note that $\varphi(1) = 0$. This implies that conditions (6) and (7) are always fulfilled by:

[4]Further simulations with different initial conditions and different parameters were conducted showing the robustness of the results. These simulations are available from the authors upon request.

[5]Plugging in the attitude parameter $\overline{\varphi} = 4$, the parameter values $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, $h^*$and k^* into condition (24) in Appendix 3 confirms that $h^* = 1.18542$ and $k^* = 0.10281$ constitutes the steady state.

$$
\begin{aligned}
h^* &= 1 \\
k^* &= \left(\frac{\gamma(1-\alpha)}{1+\gamma}\right)^{\frac{1}{1-\alpha}}
\end{aligned}
$$

In this steady state, no formal education takes place, and human capital is at its lowest possible level. This steady state, which we denote as illiterateness steady state, characterizes a situation where the economy is trapped in a vicious chain in which formal education is neglected: Since parents have no formal education, they are not willing to finance the formal education of their children, and hence the children are not interested in the education of the grandchildren, and so on.

Whether this illiterateness trap poses a severe problem depends crucially on the stability properties of this steady state and on the existence of other steady states. The same holds for the question whether temporary or permanent state intervention is necessary to avoid this steady state. The stability properties of the illiterateness steady state as well as the existence and the properties of other steady states depend on the form of the attitude function, $\varphi(h_t)$. To illustrate the different possible outcomes, we use for the rest of this section a simple attitude function, namely

$$
\varphi\left(h_t\right) = \frac{1}{\delta}\left(h_t - 1\right). \tag{8}
$$

Using this attitude function, we get the following

Proposition 4 *In addition to the illiterteness steady state, the system exhibits the following steady states:*

i) If $\beta < \frac{1}{2}$, there exists exactly one interior steady state with formal education.

ii) If $\beta > \frac{1}{2}$, the following holds: Except for non-generic values of the parameters of the model, there exist either two or no interior steady states with formal education.

Proof: See Appendix

Simulations show that for $\beta < \frac{1}{2}$ the interior steady state is globally stable, and hence the illiterateness steady state is unstable. In Figure 2 we represent simulation results for $\alpha = 0.3$, $\beta = 0.4$, $\gamma = 0.5$ and $\delta = 0.04$ for varying initial conditions of human and physical capital.

[Figure 2 here]

Again, each line constitutes the optimal path of human and physical capital depending on the initial conditions h_1 and k_1. As in Figure 1 one can easily see that also with the endogenous formation of attitudes and $\beta < \frac{1}{2}$ the system converges globally to the interior steady state, $h^* = 1.558$ and $k^* = 0.0581$.[6] This

[6]Further simulations with different initial conditions and different parameters were conducted showing the robustness of the results. These simulations are available from the authors upon request.

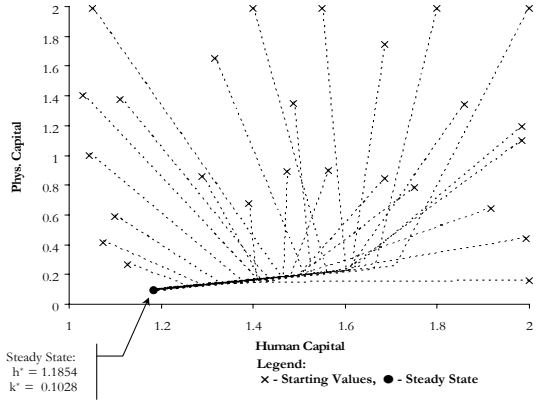

Figure 1: Exogenous education attitude

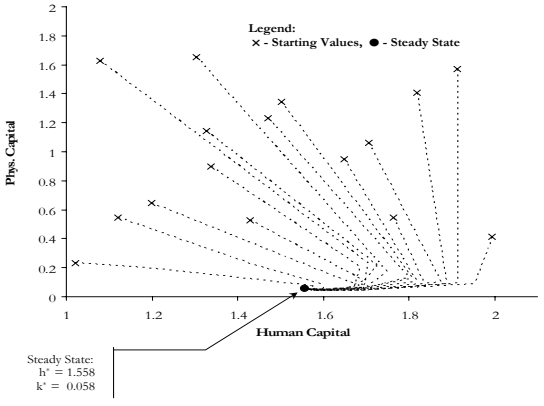

Figure 2: Endogenous education attitude with $\beta < \frac{1}{2}$

suggests that for $\beta < \frac{1}{2}$ the interior steady state with $h^* > 1$ is globally stable, and the illiterateness steady state is unstable. In this case a slight perturbation is enough to overcome the illiterateness trap.

For $\beta > \frac{1}{2}$ the illiterateness steady state is globally stable when no interior steady state exists. In Figure 3 we report the simulation results for $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$ and $\delta = 0.1$.

[Figure 3 here]

With these parameters, no interior steady state exists, and the simulation suggests that the illiterateness steady state with $h^* = 1$ and $k^* = 0.125057$ is globally stable. So in this case a permanent public intervention is necessary to overcome the illiterateness trap.

If two interior steady states exists, one of them and the illiterateness steady state are locally stable. In Figure 4 we report simulation results for $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, $\delta = 0.04$ and different initial conditions.

[Figure 4 here]

One can easily see that depending on the initial level of human and physical capital the system either converges towards the illiterateness steady state $h^* = 1$ and $k^* = 0.125057$ with $w^* = 0.377171$ or to the stable interior steady state $h^* = 1.26750$ and $k^* = 0.076934$ with $w^* = 0.382824$. In this case a temporary public intervention is enough to make the transition from the 'bad' to the 'good' steady state. Note that the lack of disutility of labor implies that the wage is a measure of the welfare of the agents. Hence, agents are indeed worse off in the illiterateness steady state than in the other one.

Proposition 4 refers to economies with a homogeneous population - each member of the first generation is endowed with the same human and physical capital, and hence all their offsprings are. So the possible multiplicity of stable steady states refers to whole economies: Depending on the initial conditions, otherwise identical societies might end up at different steady states (and connected welfare levels). One might wonder whether our model can produce a similar result within an economy: If the initial endowment with human capital is different for otherwise identical members of the first generation, will their descendants end up at different education levels and utility levels? In order to answer this question, we investigate an economy with a heterogeneous population.

3.3 Heterogeneous population

In this section we consider an economy with agents that are identical but for their initial endowment of human capital. So there are two different types of agents, U and O, with initial endowment of human capital of h_1^O and h_t^U. Since the initial human capital endowment of the two groups differ, the human capital of their offsprings might be different, too, leading different savings and physical capital levels.

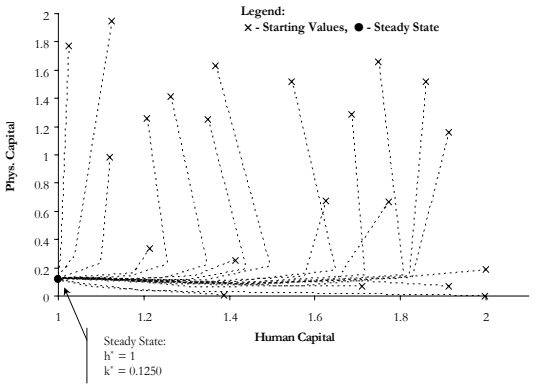

Figure 3: Endogenous education attitude with $\beta > \frac{1}{2}$ and no interior steady state

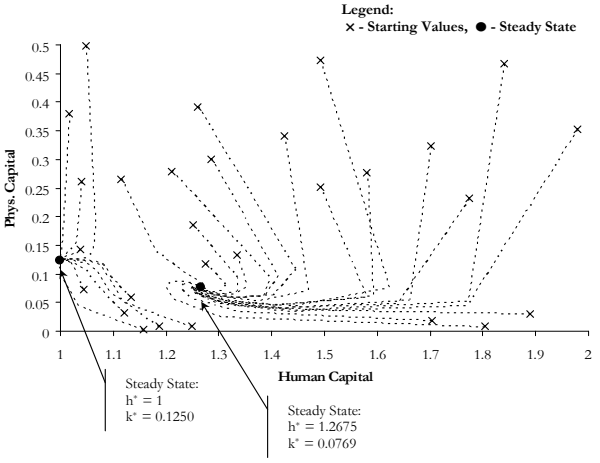

Figure 4: Endogenous education attitude with $\beta > \frac{1}{2}$ and two interior steady states

Denote by s the share of the O types in the population. The average per capita production function of the economy is given by

$$y_t = \left(sk_t^O + (1-s)\left(k_t^U\right)\right)^{\alpha}\left(sh_t^O + (1-s)\left(h_t^U\right)\right)^{1-\alpha}.$$

From this we can derive the effective wage rate per unit of human capital that agents earn:

$$\frac{\partial y_t}{\partial\left(sh_t^O + (1-s)\,h_t^U\right)} = (1-\alpha)\left(\frac{sk_t^O + (1-s)\,k_t^U}{sh_t^O + (1-s)\,h_t^U}\right)^{\alpha}$$

Factor markets are competitive and agents receive the same effective wage rate and interest rate. They differ, however, in the wage that they earn as they differ in the amount of human capital. Wages are given by

$$\begin{aligned} w_t^O &= h_t^O\,(1-\alpha)\left(\frac{sk_t^O + (1-s)\,k_t^U}{sh_t^O + (1-s)\,h_t^U}\right)^{\alpha} \\ w_t^U &= h_t^U\,(1-\alpha)\left(\frac{sk_t^O + (1-s)\,k_t^U}{sh_t^O + (1-s)\,h_t^U}\right)^{\alpha} \end{aligned}$$

Assuming for both types of agents the attitude function (8), the first order conditions for utility maximization are derived by inserting (8) and the wage of the respective type of agent into the FOCs as stated in Lemma 1:

$$\begin{aligned} \frac{\partial U^O}{\partial h_{t+1}^O} &= \frac{\frac{1}{\delta}\left(\widetilde{h}_t^O - 1\right)}{\widetilde{h}_{t+1}^O} - \frac{\frac{1}{\beta}\left(\widetilde{h}_{t+1}^O - 1\right)^{\frac{1}{\beta}-1}}{\widetilde{w}_t^O - \widetilde{k}_{t+1}^O - \left(\widetilde{h}_{t+1}^O - 1\right)^{\frac{1}{\beta}}} = 0 \\ \frac{\partial U^O}{\partial k_{t+1}^O} &= \frac{\gamma}{\widetilde{k}_{t+1}^O} - \frac{1}{\widetilde{w}_t^O - \widetilde{k}_{t+1}^O - \left(\widetilde{h}_{t+1}^O - 1\right)^{\frac{1}{\beta}}} = 0 \\ \frac{\partial U^U}{\partial h_{t+1}^U} &= \frac{\frac{1}{\delta}\left(\widetilde{h}_t^U - 1\right)}{\widetilde{h}_{t+1}^U} - \frac{\frac{1}{\beta}\left(\widetilde{h}_{t+1}^U - 1\right)^{\frac{1}{\beta}-1}}{\widetilde{w}_t^U - \widetilde{k}_{t+1}^U - \left(\widetilde{h}_{t+1}^U - 1\right)^{\frac{1}{\beta}}} = 0 \\ \frac{\partial U^U}{\partial k_{t+1}^U} &= \frac{\gamma}{\widetilde{k}_{t+1}^U} - \frac{1}{\widetilde{w}_t^U - \widetilde{k}_{t+1}^U - \left(\widetilde{h}_{t+1}^U - 1\right)^{\frac{1}{\beta}}} = 0 \end{aligned}$$

Whenever $\widetilde{k}_t^O = \widetilde{k}_t^U$ and $\widetilde{h}_t^O = \widetilde{h}_t^U$ it is easy that these FOCs are equivalent to the one for the homogenous population as stated in Lemma 1. Hence, any steady state of the model with a homogenous population constitutes also a steady state of the heterogenous population model. But for an initially heterogenous population, there may exist in addition steady states where the population remains split in two groups even in the long run. Take for example the model with the following parameter values: $s = 0.5$, $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, and $\delta = 0.04$. With these parameters, one of the steady states is given by $h^{O*} = 1.35307$, $k^{O*} = 0.07008$,

$h^{U*} = 1$, $k^{U*} = 0.10747$, leading to wages of $w^{O*} = 0.43624$ and $w^{U*} = 0.32241$. In this steady state the O-types and their offspring have higher human capital, higher wages, and consequently a higher utility level than the U-types.

Consider the simulations results in Table 2.

[Table 2 here]

In Table 2 we report the results of 26 simulations for different initial conditions of human capital, h_1^O and h_1^U and the same parameter values: $s = 0.5$, $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, and $\delta = 0.04$. For each simulation we give the initial value (Initial Cond.) and the values for h^{O*}, k^{O*}, h^{U*}, k^{U*} as well as w^{O*} and w^{U*} that the system converges to (Sim. Results). The simulation results are sorted, first, by the initial level human capital of type-U and, second, by the absolute difference between the initial values of human capital of type-O and U. One can easily see that depending on the initial conditions the system will either converge towards an egalitarian steady state in which both types have the same human and physical capital (e.g. simulations 10, 11, 13 etc) or to an unegalitarian in which, as mentioned above, type-O converges towards $h^{O*} = 1.35307$, $k^{O*} = 0.07008$, and type-U converges towards $h^{U*} = 1$, $k^{U*} = 0.10747$ (e.g. simulations 1, 2, 3 and 4 etc). Furthermore, the lower the initial level of human capital of type-U and the higher the difference between the initial levels of human capital of type-U and O the more likely it is that differences remain even in the long run.

So depending on the initial conditions and on the parameter values of the model, it is possible that even in the long run the differences remain, irrespective of the fact that factor markets are perfectly competitive and that all agents face the same interest and wage rate. Illiterateness gets inherited from generation to generation, preventing convergence of the two population groups. Comparing the steady state wage of type-O, $w^{O*} = 0.43624$, in the heterogenous case and the steady state wage, $w^* = 0.382824$, in the homogenous case one can see that $w^{O*} > w^*$. The reason for this is twofold. First, the average level of human capital in the homogenous situation is higher implying a lower wage per effciency unit. Secondly, type-O agents have a higher level of human capital in the heterogenous steady state compared to the homogenous situation leading to an additional effect on the wage, w^{O*}. Consequently, O-types are better off in the heterogeneous steady state than in the homogeneous. This suggests that people with higher human capital might resist a special subsidy to overcome the illiterateness trap of the underdogs.

4 The optimal education subsidy

In this section we analyze the efficiency properties of all steady states of the model, and the possibilities to overcome inefficiencies. For tractability reasons, we restrict attention to the homogenous population case. We compare the private investments into human and physical capital with the investments a social planner would make if endowed with the same initial capital levels. It turns out that this analysis can

Table 2: Simulations results for a population of heterogeneous agents with different initial values of human capital and $s = 0.5$, $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, $\delta = 0.04$.

		Type O			Type U		
		Human Capital	Physical Capital	Wage	Human Capital	Physical Capital	Wage
1	Initial Cond.	1.53881	1.0	1.00252	1.00212	1.0	0.65287
	Sim. Results	1.35307	0.07008	0.43624	1.00	0.10747	0.32241
2	Initial Cond.	1.52511	1.0	0.83371	1.02751	1.0	0.66849
	Sim. Results	1.35307	0.07008	0.43624	1.00	0.10747	0.32241
3	Initial Cond.	2.2	1.0	1.32474	1.03208	1.0	0.62663
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
4	Initial Cond.	2.49127	1.0	1.23573	1.03504	1.0	0.61118
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
5	Initial Cond.	1.5	1.0	0.97278	1.05	1.0	0.97278
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
6	Initial Cond.	2.25407	1.0	1.14192	1.05556	1.0	0.63527
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
7	Initial Cond.	1.68072	1.0	1.06966	1.06626	1.0	0.67860
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
8	Initial Cond.	1.97	1.0	1.21502	1.08	1.0	0.66610
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
9	Initial Cond.	2	1.0	1.22991	1.08	1.0	0.66415
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
10	Initial Cond.	1.5	1.0	0.97278	1.08	1.0	0.97278
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
11	Initial Cond.	1.90000	1.0	1.18004	1.08	1.0	0.67076
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
12	Initial Cond.	2.5	1.0	1.22152	1.08728	1.0	0.64116
	Sim. Results	1.35307	0.07008	0.43624	1.0	0.10747	0.32241
13	Initial Cond.	1.96000	1.0	1.21004	1.1	1.0	0.66676
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
14	Initial Cond.	1.55252	1.0	0.84595	1.12882	1.0	0.72365
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
15	Initial Cond.	1.89667	1.0	0.99499	1.15546	1.0	0.71249
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
16	Initial Cond.	1.5	1.0	0.95960	1.20	1.0	0.76768
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
17	Initial Cond.	2.43514	1.0	1.21378	1.22295	1.0	0.71423
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
18	Initial Cond.	1.60495	1.0	1.00936	1.25309	1.0	0.78808
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
19	Initial Cond.	1.62129	1.0	0.87640	1.27515	1.0	0.79875
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
20	Initial Cond.	1.90798	1.0	0.99975	1.30503	1.0	0.79242
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
21	Initial Cond.	1.85826	1.0	1.13101	1.32959	1.0	0.80923
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
22	Initial Cond.	2.33302	1.0	1.36085	1.34022	1.0	0.78175
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
23	Initial Cond.	2.35603	1.0	1.37115	1.34520	1.0	0.78287
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
24	Initial Cond.	2.26506	1.0	1.14633	1.36798	1.0	0.80058
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
25	Initial Cond.	1.73720	1.0	1.06268	1.39741	1.0	0.85482
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282
26	Initial Cond.	1.59003	1.0	0.86260	1.49330	1.0	0.91801
	Sim. Results	1.26750	0.07693	0.38282	1.26750	0.07693	0.38282

Note, simulation results are sorted, first, by the initial level human capital of type-U and, second, by the absolute difference between the initial values of human capital of type-O and U.

be carried out for a general attitude function with the properties as specified in section 2.

The social planner chooses the investment in human and physical capital such that he maximizes the weighted sum of utilities of all generations, subject to the resource constraint of the economy.

$$\begin{aligned}\max_{c_{2,t},c_{3,t+1},h_{t+1}} W &= \sum_{t=1}^{\infty}\omega_t\left[\ln c_{2,t}+\gamma\ln c_{3,t+1}+\varphi(h_t)\ln h_{t+1}\right]\\ \text{s.t. } k_t^{\alpha}h_t^{1-\alpha} &= k_{t+1}+(h_{t+1}-1)^{\frac{1}{\beta}}+c_{2,t}+c_{3,t} \text{ and}\\ h_t &\geq 1,\\ k_t &\geq 0\end{aligned}$$

with ω_t being larger than zero for all t. Denote by $\widehat{k}_{t+1}$ and $\widehat{h}_{t+1}$ the optimal choice of the social planner. The sequence of optimal choices is denoted by $\{\widehat{k}_t,\widehat{h}_t\}_{t=2}^{\infty}$, and $k_1>0$ and $h_1\geq 1$ denote the initial endowments with physical and human capital.

Defining

$$\xi_t=\frac{(1-\alpha)\widehat{k}_{t+1}^{\alpha}\widehat{h}_{t+1}^{-\alpha}}{\widehat{k}_{t+1}^{\alpha}\widehat{h}_{t+1}^{1-\alpha}-\widehat{k}_{t+2}-\left(\widehat{h}_{t+2}-1\right)^{\frac{1}{\beta}}-\widehat{k}_{t+1}\widehat{r}_{t+1}},$$

the socially optimal solution is characterized by the following lemma:

Lemma 5 *If $\widehat{k}_t>0$ it holds that:*

i) The solution of the social planners problem fulfills the first order conditions

$$\begin{aligned}\frac{\partial W}{\partial h_{t+1}}&=\omega_t\left(\frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}}-\frac{\frac{1}{\beta}\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{\widehat{k}_t^{\alpha}\widehat{h}_t^{1-\alpha}-\widehat{k}_{t+1}-\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}-\widehat{k}_t\widehat{r}_t}\right)\\&\quad+\omega_{t+1}\left(\xi_t+\varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right)\\&=0\end{aligned}\tag{9}$$

and

$$\frac{\partial W}{\partial k_{t+1}}=\omega_t\left(\frac{\gamma}{\widehat{k}_{t+1}}-\frac{1}{\widehat{k}_t^{\alpha}\widehat{h}_t^{1-\alpha}-\widehat{k}_{t+1}-\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}-\widehat{k}_t\widehat{r}_t}\right)=0\tag{10}$$

ii) $\widehat{k}_{t+1}>0$, $\widehat{h}_{t+1}>1$, and $\widehat{c}_{2,t}>0$.

Proof: see Appendix.

Comparing Lemma 1 with Lemma 5, one realizes that condition (7) of the private solution coincides with condition (10) of the optimal solution, but condition (9) differs from (6) by the term $\omega_{t+1}\left(\xi_t+\varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right)$. Since $\widehat{c}_{2,t+1}=$

$\widehat{k}_{t+1}^{\alpha}\widehat{h}_{t+1}^{1-\alpha}-\widehat{k}_{t+2}-\left(\widehat{h}_{t+2}-1\right)^{\frac{1}{\beta}}-\widehat{k}_{t+1}\widehat{r}_{t+1}>0, \xi_t>0$ for all t. Similarly, $\varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}>0$. So as long as the social planner cares at least a bit about the future generation, i.e. as long as $\omega_{t+1}>0$, the sequence of private decisions, $\{\widetilde{k}_t,\widetilde{h}_t\}_{t=2}^{\infty}$, differs from the sequence of socially optimal choices, $\{\widehat{k}_t,\widehat{h}_t\}_{t=2}^{\infty}$ - the private solution is not optimal. This result is not surprising, since parents do not care about the welfare of their children, but only about their human capital. This leads to an externality captured by the variable ξ_t. Even if the attitude towards children's education were independent of the own education, an externality would be present. The endogenous attitude towards education implies a second type of externality, which leads to the emergence of $\varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}$ in (9). Both of these externalities are neglected when human capital is privately provided, leading to an underprovision of human capital. Recall that this inefficiency occurs even in the better, interior steady states.

Can this inefficiency be overcome by public expenditures on human capital formation? Think of a situation where the public finances schools and universities. Even if schools and universities are fully financed by the state, parents still have to take care of the children's costs of living, the costs of supplementary education, the costs of teaching material and other things indirectly connected to the human capital formation of children. Hence, parts of the education expenditures are always paid by parents. Furthermore, in such a system of mixed financing a better education of the children requires higher expenditures of parents as well as of the state. Finally, the education level of the children is largely influenced by the parent's willingness to cover the children's costs of living, even in a system where the state finances schools and universities. To model such a situation where human capital formation is partly privately, partly publicly financed, assume that in each period t private education expenditures are subsidized by the state at a rate s_t. To finance this subsidy, wage income is taxed at a rate τ_t. The balanced budget condition for the state for period t is given by:

$$s_t\left(h_{t+1}-1\right)^{\frac{1}{\beta}}=\tau_t w_t. \tag{11}$$

We assume that an individual agent takes the tax rate and the subsidy scheme as given when he maximizes his utility. This implies that he does not take into account the balanced budget condition of the state. This assumption seems plausible for a large economy with many agents. With this simplification, the decision problem of a representative agent working in period t can be written as:

$$\begin{aligned}
\max_{c_{2,t},c_{3,t+1},h_{t+1}} U(c_{2,t},c_{3,t+1},h_{t+1}) &= \ln c_{2,t}+\gamma\ln c_{3,t+1}+\varphi(h_t)\ln h_{t+1}\\
\text{s.t. } (1-\tau_t)w_t &= c_{2,t}+\frac{c_{3,t+1}}{r_{t+1}}+(1-s_t)\left(h_{t+1}-1\right)^{\frac{1}{\beta}} \text{ and}\\
h_{t+1} &\geq 1,\\
c_{2,t} &\geq 0,\\
c_{3,t+1} &\geq 0.
\end{aligned}$$

Denote by $\overline{k}_{t+1}$ and $\overline{h}_{t+1}$ the utility maximizing choice of the agent working in period t, when the human capital formation is subsidized. The sequence of utility

maximizing choices is denoted by $\{\overline{k}_t, \overline{h}_t\}_{t=2}^{\infty}$, and $k_1 > 0$ and $h_1 \geq 1$ denote the initial endowments of the economy with physical and human capital. Using the budget constraint to insert for $c_{2,t}$ the first order conditions are:

$$\frac{\partial U}{\partial h_{t+1}} = \frac{\varphi(\overline{h}_t)}{\overline{h}_{t+1}} - \frac{(1-s_t)\frac{1}{\beta}\left(\overline{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{(1-\tau_t)\overline{w}_t - \overline{k}_{t+1} - (1-s_t)\left(\overline{h}_{t+1}-1\right)^{\frac{1}{\beta}}} = 0 \tag{12}$$

$$\frac{\partial U}{\partial k_{t+1}} = \frac{\gamma}{\overline{k}_{t+1}} - \frac{1}{(1-\tau_t)\overline{w}_t - \overline{k}_{t+1} - (1-s_t)\left(\overline{h}_{t+1}-1\right)^{\frac{1}{\beta}}} = 0. \tag{13}$$

Applying the same reasoning as in the proof of lemma 1 it is easy to see that the first order conditions characterize the solution.

Is it possible to find a sequence of subsidy schemes $\{s_t, \tau_t\}_{t=1}^{\infty}$ such that the sequence of socially optimal choices is induced? For given initial endowment with physical and human capital such a sequence would have to induce a sequence of individual choices $\{\overline{k}_{t+1}, \overline{h}_{t+1}\}_{t=1}^{\infty}$ such hat $\overline{k}_{t+1} = \widehat{k}_{t+1}$ and $\overline{h}_{t+1} = \widehat{h}_{t+1}$ for all periods. Furthermore, the sequence of schemes would have to respect the balanced budget condition (11) in all periods.

The following proposition shows that there exists indeed a sequence of subsidy schemes that induces an optimal outcome.

Proposition 6 *For $\widehat{k}_t > 0$ it holds that:*

i) The sequence of subsidy schemes $\{s_t, \tau_t\}_{t=1}^{\infty}$ defined by

$$s_t = \frac{\beta}{(\gamma+1)} \frac{\left(\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}\right)}{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}} \frac{\omega_{t+1}}{\omega_t} \left(\xi_t + \varphi'(\widehat{h}_{t+1}) \ln \widehat{h}_{t+2}\right) \tag{14}$$

and

$$\tau_t = s_t \frac{(\widehat{h}_{t+1}-1)^{\frac{1}{\beta}}}{\widehat{w}_t} \tag{15}$$

induces a sequence of choices $\{\overline{k}_{t+1}, \overline{h}_{t+1}\}_{t=1}^{\infty}$ such that $\overline{k}_{t+1} = \widehat{k}_{t+1}$ and $\overline{h}_{t+1} = \widehat{h}_{t+1}$ in all t.

ii) $\{s_t, \tau_t\}_{t=1}^{\infty}$ respects the balanced budget condition in all periods.

iii) For all t, $0 < s_t < 1$.

Proof: see Appendix

The above proposition shows that an appropriate subsidy scheme can ensure efficiency. The optimal subsidy rate is always strictly larger than zero, so a permanent subsidy is necessary to achieve efficiency. The optimal rate in period t, however, depends on the optimal values of human and physical capital in periods t, $t+1$, and $t+2$. Since nothing guarantees that these optimal human and physical capital values are constant over time, s_t might vary over time accordingly.

5 Conclusions

We have shown that the private allocation of resources leads to inefficient human capital formation. If parent's attitude towards education of the children depends on their own education, the economy might get trapped in an illiterateness steady state where a low education level of the parents leads to negligence of the children's education, reproducing the low education level in the next generation. To overcome such a steady state, temporary or permanent state intervention is necessary, depending on the stabilty properties of the illiterateness steady state. Because of the intergenerational transmission of education attitudes the population of an economy might also be split in the long run into different education groups, even if the agents are identical in all respects but for their initial endowment with human capital.

When the economy is not trapped in such an illiterateness steady state, the purely private financing of the education system also leads to inefficiencies. These inefficiencies can be overcome by a permanent public support for the education of children. This conclusion requires some qualifications. First, a similar result would occur if the parents' attitude toward the education of their children were independent of their own education. Second, if the economy is not in a steady state, the efficient tax and subsidy rates might change from period to period. For political reasons as well as for lack of information, it may be difficult to make these necessary adjustments. Third, the optimal subsidy rate depends on the weight the social planner puts on the different generations. Hence, there is room for intergenerational conflicts. Finally, our model is based on the assumption that labor supply is fixed. Hence, the taxation of wage income does not create any excess burden on the labor market. If labor supply is elastic and if a non-distortive tax is not available, a trade-off exists between the inefficiency created by the tax system and the inefficiency due to the externalities in the human capital formation.

Notwithstanding these qualifications, it can be concluded that the broadening of the model of human behavior to allow for more complex intergenerational relations leads to inefficiencies that have been neglected so far. The analysis thus gives further support for government intervention to support an optimal investment into the education of our children in order to achieve a maximum amount of welfare.

References

[1] Alexander, R. (1987), *The Biology of Moral Systems*, New York: Aldine de Gruyter.

[2] Arrondel, L. and A Masson (2001), *Family Transfers Involving Three Generations,* Scandinavian Journal of Economics, 103(3), 415-443.

[3] Bisin, A. and T. Verdier (2001), *The Economics of Cultural Transmission and the Dynamics of Preferences*, Journal of Economic Theory, 97, 298-319.

[4] Becker, G. and N. Tomes (1976), *Child endowments and the quantity and quality of children*, Journal of Political Economy, 84(4), 143-162.

[5] Becker, G. and N. Tomes (1986), *Human capital and the rise and fall of families*, Journal of Labor Economics, 4(3), 1-39.

[6] Bernheim, D., A. Shleifer, and L. Summers (1985), *The Strategic Bequest Motive*, Journal of Political Economy, vol. 93(6), 1045-1076.

[7] Bernheim, D., A. Shleifer, and L. Summers (1986), *The Strategic Bequest Motive*, Journal of Labor Economics, vol. 4(3), 151-182.

[8] Boyd, R. and P. Richerson (1985), *Culture and the evolutionary process*, Chicago: University of Chicago Press.

[9] Brooks, R. (1998), *Staying or Leaving? A literature review of factors affecting the take-up of post-16 options*, Slough: National Foundation for Educational Research.

[10] Brooks, R. (2003), *Young people's higher education choices: the role of family and friends*, British Journal of Sociology of Education, 24(3), pp. 283-297.

[11] Cavalli-Sforza, L. and M. Feldman (1973), *Cultural versus biological inheritance from parent to children*, American Journal of Human Genetics, 25,618-637.

[12] Champarnaud, L., V. Ginsburgh, and P. Michel (2005), *Can Public Arts Education Crowd Out Arts Subsidization?*, mimeo.

[13] Dufwenberg, M. and G. Kirchsteiger (2004), *A Theory of Sequential Reciprocity*, Games and Economic Behavior, 47(2), 268-298.

[14] Eckstein, Z. and Zilcha, I. (1994), *The effects of compulsory schooling on growth, income distribution and welfare*, Journal of Public Economics, vol. 54(3), 339-359.

[15] Engelmann D. and Fischbacher U. (2002), *Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game*, University of Zurich, Working Paper No. 132.

[16] Ehrlich P. (2000),*Human natures: Genes, cultures, and the human prospect*. Washington (D. C.): Island Press.

[17] Galor, O. and Tsiddon, D. (1997), *The Distribution of Human Capital and Economic Growth*, Journal of Economic Growth, vol. 2(1), 93-124.

[18] Galor, O. and Tsiddon, D. (1997), *Technological Progress, Mobility, and Economic Growth*, American Economic Review, vol. 87(3), 363-82.

[19] Kapteyn, A., T. Wansbeek and J.Buyze (1980), *The dynamics of preference formation*, Journal of Economic Behavior and Organization, 1, 123-157.

[20] Kohli, M. and Künemund H. (2003), *Intergenerational Transfers in the Family: What Motivates Giving?*, published in: Bengtson V. and Lowenstein A. (eds.) (2003), *Global Aging and Challenges to Families*, New York: Aldine de Gruyter, 123-142.

[21] Mauldin, T., Y. Mimura, and M. Lino (2001), *Parental Expenditures on Children's Education*, Journal of Family and Economic Issues, 22(3), 221-241.

[22] Mauss. M. (1950/1990), *The gift: The form and reason for exchange in archaic society*. London: Routledge.

[23] Nowak, M and K. Sigmund (1998), *The Dynamics of Indirect Reciprocity*, Journal of Theoretical Biology, 194, 561-574.

[24] Payne, J. (2002), *Attitudes to Education and Choices at Age 16: A Brief Research Review*, Report to the DfES Advisory Panel on research Issues for the 14-16 Age Group.

[25] Plug, E. and Vijverberg, W. (2003), *Schooling, Family Background, and Adoption: Is It Nature or Is It Nurture?*, Journal of Political Economy, 111(3), 611-641.

[26] Rabin, M. (1993), *Incorporating Fairness into Game Theory and Economics,* The American Economic Review, 83(5), 1281-1302.

[27] Sacerdote, B. (2002), *The Nature and Nurture of Economic Outcomes*, The American Economic Review, 92(2), 344-348.

6 Appendix

6.1 Proof of Lemma 1

Note first that $\widetilde{h}_{t+1}$ and $\widetilde{k}_{t+1}$ have to be finite for all finite values of $(\widetilde{k}_t, \widetilde{h}_t)$. The utility function is strictly quasiconcave, implying a unique solution, which might be either interior (in which case the first order conditions hold) or at the lower bounds. By (3) $\widetilde{w}_t > 0$ whenever $\widetilde{k}_t > 0$. Furthermore, $\widetilde{w}_t - \widetilde{k}_{t+1} = \widetilde{c}_{2,t} + \left(\widetilde{h}_{t+1} - 1\right)^{\frac{1}{\beta}} \geq \widetilde{c}_{2,t} > 0$ due to the INADA condition of the utility function with respect to the consumption levels. This implies that $\frac{\partial U}{\partial k_{t+1}} = \infty$ at $k_{t+1} = 0$. This requires that the condition (7) as well as ii) must hold.

As for the solution for the human capital, note first that for $\widetilde{h}_t = 1$, $\frac{\partial U}{\partial h_{t+1}} = 0$ at $h_{t+1} = 1$. This gives iii) and that condition (6) holds in this case.

If $\widetilde{h}_t > 1$, $\frac{\partial U}{\partial h_{t+1}} = \infty$ at $h_{t+1} = 1$, implying $\widetilde{h}_{t+1} > 1$. This gives iv) and that condition (6) holds also for $\widetilde{h}_t > 1$, which completes the proof. ■

6.2 Proof of Proposition 2

We first introduce the following dynamic system, denoted as upper bound economy and by superscript b, which will be useful for the proof:

$$h_{t+1}^b = (1-\alpha)^\beta (k_t^b)^{\alpha\beta} (h_t^b)^{(1-\alpha)\beta} + 1 \tag{16}$$

$$k_{t+1}^b = (1-\alpha)(k_t^b)^\alpha (h_t^b)^{(1-\alpha)} \tag{17}$$

The proof now proceeds in three steps. In the first step, we will show that for the same initial conditions for human and physical capital the path of the upper bound economy provides an upper bound for the path of the economy we analyze. In the second step, we will show that the upper bound economy exhibits a globally stable steady state, to which the system converges from any initial conditions. In the third step we will use this steady state to finalize the proof.

Step 1: If the indirect reciprocity economy and the upper bound economy start at the same initial conditions $k_1^b = \widetilde{k}_1 > 0$ and $h_1^b = \widetilde{h}_1$, it holds that:

$$\begin{aligned} k_t^b &\geq \widetilde{k}_t \text{ for all } t > 1 \\ h_t^b &\geq \widetilde{h}_t \text{ for all } t > 1. \end{aligned}$$

The proof is made by induction. For the same initial conditions $k_1^b = \widetilde{k}_1$ and $h_1^b = \widetilde{h}_1$, the definition of the upper bound economy, (3), and (4) give

$$k_2^b = (1-\alpha)(k_1^b)^\alpha (h_1^b)^{(1-\alpha)} = (1-\alpha)(\widetilde{k}_1)^\alpha (\widetilde{h}_1)^{(1-\alpha)} = \widetilde{w}_1 \geq \widetilde{k}_2$$

and

$$\begin{aligned} h_2^b &= (1-\alpha)^\beta (k_1^b)^{\alpha\beta} (h_1^b)^{(1-\alpha)\beta} + 1 \\ &= (1-\alpha)^\beta (\widetilde{k}_1)^{\alpha\beta} (\widetilde{h}_t^b)^{(1-\alpha)\beta} + 1 \\ &= (\widetilde{w}_1)^\beta + 1 \geq (\widetilde{e}_1)^\beta + 1 = \widetilde{h}_2. \end{aligned}$$

So $k_2^b \geq \widetilde{k}_2$ and $h_2^b \geq \widetilde{h}_2$. It is obvious that h_{t+1}^b and k_{t+1}^b are monotonically increasing in h_t^b and k_t^b . This implies that $k_{t+1}^b \geq \widetilde{k}_{t+1}$ and $h_{t+1}^b \geq \widetilde{h}_{t+1}$ whenever $k_t^b \geq \widetilde{k}_t$ and $h_t^b \geq \widetilde{h}_t$, which completes the proof of Step 1.

Step 2: For any initial condition $k_1^b > 0$ [7] the upper bound economy converges to a unique stable state k^{b*}, h^{b*} with $k^{b*} > 0$ and $h^{b*} > 1$.

To show this, note first that $k_1^b > 0$ implies that $k_t^b > 0$ and $h_t^b > 1$ for all $t > 1$. From the definition of the upper bound economy we get

$$h_{t+1}^b = \left(k_{t+1}^b\right)^\beta + 1,$$

implying that

$$h_t^b = \left(k_t^b\right)^\beta + 1.$$

[7] Recall that we restrict our analysis to the nontrivial case of $\widetilde{k}_1 > 0$, which of course implies that $k_1^b > 0$.

Hence, equation of motion of the upper bound economy is characterized by:

$$k_{t+1}^b = (1-\alpha)\,(k_t^b)^\alpha \left(\left(k_t^b\right)^\beta + 1 \right)^{(1-\alpha)}. \tag{18}$$

Differentiating we get

$$\frac{\partial k_{t+1}^b}{\partial k_t^b} = \left[(1-\alpha)\,(k_t^b)^{(\alpha-1)} \left((k_t^b)^\beta + 1 \right)^{(-\alpha)} \right] \left[(\alpha\,(1-\beta)+\beta)\,(k_t^b)^\beta + \alpha 1 \right] > 0$$

and

$$\begin{aligned} \frac{\partial^2 k_{t+1}^b}{\partial k_t^b \partial k_t^b} &= \left[-(1-\alpha)^2\,(k_t^b)^{\alpha-2} \left((k_t^b)^\beta + 1 \right)^{(-1-\alpha)} \right] \\ &\quad \left[\alpha\,(1-\beta) + \beta\,(1-\beta)\,(k_t^b)^{2\beta} + (1-\beta)\,(2\alpha+\beta)\,(k_t^b)^\beta + \alpha \right] < 0 \end{aligned}$$

since

$$\begin{gathered} -(1-\alpha)^2\,(k_t^b)^{\alpha-2} \left((k_t^b)^\beta + 1 \right)^{(-1-\alpha)} < 0 \\ \alpha\,(1-\beta) + \beta\,(1-\beta)\,(k_t^b)^{2\beta} > 0 \\ (1-\beta)\,(2\alpha+\beta)\,(k_t^b)^\beta > 0 \\ \alpha > 0. \end{gathered}$$

Hence, the equation of motion (18) is strictly monotone and concave in k_t. This and the fact that the system has a steady state at $k^b = 0$ implies that there is at most one other steady state with $k^b > 0$.

To investigate the possibility of steady states with $k^{b*} > 0$, we set $k_t^b = k_{t+1}^b = k^{b*}$ in (18) and get:

$$k^{b*} = (1-\alpha)\,(k^{b*})^\alpha [(k^{b*})^\beta + 1]^{(1-\alpha)},$$

implying:

$$(k^{b*})^{\frac{1}{(1-\alpha)}} = (1-\alpha)\,(k^{b*})^{\frac{\alpha}{(1-\alpha)}} [(k^{b*})^\beta + 1].$$

Dividing by $(k^{b*})^{\frac{1}{(1-\alpha)}}$ leads to

$$\begin{gathered} 1 = (1-\alpha)\,(k^{b*})^{(-1)} [(k^{b*})^\beta + 1] \\ 1 = (1-\alpha)\,[(k^{b*})^{(\beta-1)} + (k^{b*})^{(-1)}]. \end{gathered} \tag{19}$$

The right hand side of equation (19) is continuos and strictly monotonically decreasing in k^{b*}. Furthermore,

$$\begin{aligned} \lim_{k^{b*}\to 0} (1-\alpha)\,[(k^{b*})^{(\beta-1)} + (k^{b*})^{(-1)}] &= \infty \\ \lim_{k^{b*}\to \infty} (1-\alpha)\,[(k^{b*})^{(\beta-1)} + (k^{b*})^{(-1)}] &= 0. \end{aligned}$$

Hence, there exists a unique $k^{b*} > 0$ fulfilling (19) characterizing the second steady state of the upper bound economy. The steady state value of h is given by

$$h^{b*} = (k^{b*})^\beta + 1 > 1$$

Recall that the equation of motion (19) is strictly monotone and concave. Hence, $k^b_{t+1} > k^b_t$ whenever in $k^b_t < k^{b*}$ and $k^b_{t+1} < k^b_t$ whenever in $k^b_t > k^{b*}$. Therefore, the upper bound economy converges to the steady state k^{b*}, h^{b*} whenever $k^b_1 > 0$.

Step 3: For any initial conditions $\widetilde{k}_1 \geq 0$ and $\widetilde{h}_1 \geq 0$ there exists a t^m such that:

$$\begin{aligned} \widetilde{h}_t &< h^{b*} + 1 \text{ whenever } t > t^m \\ \widetilde{k}_t &< k^{b*} + 1 \text{ whenever } t > t^m \\ \widetilde{w}_t &< (1-\alpha)(k^{b*}+1)^{\alpha}(h^{b*}+1)^{1-\alpha} \text{ whenever } t > t^m \end{aligned}$$

Step 3) follows immediately from Step 1), Step 2), and the wage equation 3). ■

6.3 Proof of Proposition 3

Education attitude is exogenously given at $\overline{\varphi}$. Taking this into account and inserting 3 and 2 into 6 and 7 gives

$$(1-\alpha)\,\widetilde{k}_t^{\alpha}\widetilde{h}_t^{1-\alpha} = \left(\left(\frac{(1+\gamma)}{\beta\overline{\varphi}}\right)\frac{\widetilde{h}_{t+1}}{\left(\widetilde{h}_{t+1}-1\right)}+1\right)\left(\widetilde{h}_{t+1}-1\right)^{\frac{1}{\beta}} \tag{20}$$

$$(1-\alpha)\,\widetilde{k}_t^{\alpha}\widetilde{h}_t^{1-\alpha} = \left(\frac{(1+\gamma)}{\gamma}+\frac{\beta\overline{\varphi}}{\gamma}\frac{\left(\widetilde{h}_{t+1}-1\right)}{\widetilde{h}_{t+1}}\right)\widetilde{k}_{t+1}. \tag{21}$$

Substituting 21 into 20 for k_t and rearranging terms gives

$$(1-\alpha)\left(\frac{\gamma}{\beta\overline{\varphi}}\right)^{\alpha} = \frac{1}{\widetilde{h}_t\left(\widetilde{h}_t-1\right)^{\left(\frac{1}{\beta}-1\right)\alpha}}\left(\frac{(1+\gamma)}{\beta\overline{\varphi}}\right)\widetilde{h}_{t+1}\left(\widetilde{h}_{t+1}-1\right)^{\frac{1}{\beta}-1} \tag{22}$$

$$+\frac{1}{\widetilde{h}_t\left(\widetilde{h}_t-1\right)^{\left(\frac{1}{\beta}-1\right)\alpha}}\left(\widetilde{h}_{t+1}-1\right)^{\frac{1}{\beta}}. \tag{23}$$

Since we investigate the steady state, we ignore the indices. Rearranging terms one gets

$$(1-\alpha)\left(\frac{\gamma}{\beta\overline{\varphi}}\right)^{\alpha} = \left(\frac{(1+\gamma)}{\beta\overline{\varphi}}\right)(h^*-1)^{\left(\frac{1}{\beta}-1\right)(1-\alpha)} + \frac{1}{h^*}(h^*-1)^{\frac{1}{\beta}-\left(\frac{1}{\beta}-1\right)\alpha}. \tag{24}$$

Obviously, the left hand side of 24 is positive. For the right hand side, notice that $\frac{1}{\beta}-\left(\frac{1}{\beta}-1\right)\alpha > 1$ and $\left(\frac{1}{\beta}-1\right)(1-\alpha) > 0$. This implies that the right hand side is strictly increasing in h^*, that it is zero for $h^* = 1$, and that it goes to infinity for h^* going to infinity. Hence there exists exactly one value h^* that fulfills 24, and this value is strictly larger than 1. ■

6.4 Proof of Proposition 4

Inserting the attitude function, 3 and 2 into 6 and 7 gives

$$(1-\alpha)\,\widetilde{k}_t^{\alpha}\widetilde{h}_t^{1-\alpha} = \left(\left(\frac{(1+\gamma)}{\beta\frac{1}{\delta}\left(\widetilde{h}_t-1\right)}\right)\frac{\widetilde{h}_{t+1}}{\left(\widetilde{h}_{t+1}-1\right)}+1\right)\left(\widetilde{h}_{t+1}-1\right)^{\frac{1}{\beta}} \quad (25)$$

$$(1-\alpha)\,\widetilde{k}_t^{\alpha}\widetilde{h}_t^{1-\alpha} = \left(\frac{(1+\gamma)}{\gamma}+\frac{\beta\frac{1}{\delta}\left(\widetilde{h}_t-1\right)}{\gamma}\frac{\left(\widetilde{h}_{t+1}-1\right)}{\widetilde{h}_{t+1}}\right)\widetilde{k}_{t+1}. \quad (26)$$

Substituting 25 into 26 for k_t, ignoring the indices, and rearranging terms leads to the following condition for an interior steady state:

$$(1-\alpha)\left(\frac{\delta\gamma}{\beta}\right)^{\alpha} = \left(\frac{\delta\,(1+\gamma)}{\beta}\right)(h^*-1)^{\left(\frac{1}{\beta}-2\right)(1-\alpha)}+\frac{1}{h^*}\,(h^*-1)^{\frac{1}{\beta}-\left(\frac{1}{\beta}-2\right)\alpha}. \quad (27)$$

Define

$$\begin{aligned} lhs &: = (1-\alpha)\left(\frac{\delta\gamma}{\beta}\right)^{\alpha} \\ rhs(h^*) &: = \left(\frac{\delta\,(1+\gamma)}{\beta}\right)(h^*-1)^{\left(\frac{1}{\beta}-2\right)(1-\alpha)}+\frac{1}{h^*}\,(h^*-1)^{\frac{1}{\beta}-\left(\frac{1}{\beta}-2\right)\alpha} \\ a(h^*) &: = \left(\frac{\delta\,(1+\gamma)}{\beta}\right)(h^*-1)^{\left(\frac{1}{\beta}-2\right)(1-\alpha)} \\ b(h^*) &: = \frac{1}{h^*}\,(h^*-1)^{\frac{1}{\beta}-\left(\frac{1}{\beta}-2\right)\alpha} \end{aligned}$$

Proof of i) lhs strictly positive. If $\beta<\frac{1}{2}$, $\left(\frac{1}{\beta}-2\right)(1-\alpha)>0$ and $\frac{1}{\beta}-\left(\frac{1}{\beta}-2\right)\alpha>2$. This implies that $\frac{\partial a}{\partial h^*}>0$ and $\frac{\partial b}{\partial h^*}>0$ - rhs is strictly increasing in h^*. Furthermore, $rhs(h^*=1)=0$, and $\lim_{h^*\to\infty}a(h^*)=\infty$. Therefore there exists exactly one $h^*>1$ such that $lhs=rhs(h^*)$.

Proof of ii) Again, lhs strictly positive. If $\frac{1}{2}<\beta<1$, $\left(\frac{1}{\beta}-2\right)(1-\alpha)<0$ and $1<\frac{1}{\beta}-\left(\frac{1}{\beta}-2\right)\alpha$. This implies that $\lim_{h^*\to1}a(h^*)=\infty$, $\lim_{h^*\to\infty}a(h^*)=0$, $\lim_{h^*\to\infty}b(h^*)=\infty$ and $b(h^*=1)=0$. This gives $\lim_{h^*\to1}rhs(h^*)=\lim_{h^*\to\infty}rhs(h^*)=\infty$ and finite values of rhs for all other values of h^*.

Next we show that $rhs(h^*)$ has a unique local extremum in the interior. Because of $\lim_{h^*\to1}rhs(h^*)=\lim_{h^*\to\infty}rhs(h^*)=\infty$, a unique interior local extremum must be a unique local minimum of $rhs(h^*)$. Uniqueness of the local minimum implies that $\frac{\partial rhs(h^*)}{\partial h^*}<0$ for all values of h^* below this minimum and $\frac{\partial rhs(h^*)}{\partial h^*}>0$ for all values of h^* above this minimum. In the interior, any local extremum is characterized by the condition

$$\frac{\partial rhs(h^*)}{\partial h^*}=0$$

leading to

$$\begin{aligned} & \frac{(h^*-1)}{h^*} - \left(\frac{1}{\beta} - 2\right)(1-\alpha)\left(\frac{\delta(1+\gamma)}{\beta}\right)\frac{h^*}{(h^*-1)^2} \qquad (28) \\ = \; & \frac{1}{\beta} - \left(\frac{1}{\beta} - 2\right)\alpha. \end{aligned}$$

For the left hand side of equation 28, we have

$$\lim_{h\to 1}\left(\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2}\right) = \infty,$$

with $z = -\left(\frac{1}{\beta} - 2\right)(1-\alpha)\left(\frac{\delta(1+\gamma)}{\beta}\right) > 0$, and

$$\lim_{h\to\infty}\left(\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2}\right) = 1.$$

This implies that there is at least one local extremum in the interior. To check uniqueness, we will show $\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2}$ is strictly decreasing in in h^* for the relevant values of h^*. The first derivative of the left hand side is given by

$$\frac{\partial\left(\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2}\right)}{\partial h^*} = \frac{-zh^{*2}(h^*+1) + (h^*-1)^3}{h^{*2}(h^*-1)^3} \qquad (29)$$

To see that this derivative is strictly negative for the relevant values of h^*, note first that $\frac{1}{\beta} - \left(\frac{1}{\beta} - 2\right)\alpha > 1$. All solutions to equation 28 must satisfy the condition

$$\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2} > 1$$

which is equivalent to

$$zh^{*2} > h^*(h^*-1)^2 - (h^*-1)^3.$$

Inserting into 29 implies that

$$\begin{aligned} \frac{\partial\left(\frac{(h^*-1)}{h^*} + z\frac{h^*}{(h^*-1)^2}\right)}{\partial h^*} &< \frac{-\left(h^*(h^*-1)^2 - (h^*-1)^3\right)(h^*+1) + (h^*-1)^3}{h^{*2}(h^*-1)^3} \\ &= \frac{-2}{h^{*2}(h^*-1)} < 0 \end{aligned}$$

So the left hand side of 28 is strictly decreasing in the relevant area, and hence equation 28 has a unique solution. This implies that $rhs(h^*)$ has a unique local minimum in the interior whenever $\beta > \frac{1}{2}$, and that $\frac{\partial rhs(h^*)}{\partial h^*} < 0$ for all values of h^* below this minimum and $\frac{\partial rhs(h^*)}{\partial h^*} > 0$ for all values of h^* above this minimum. Furthermore, recall that $\lim_{h^*\to 1} rhs(h^*) = \lim_{h^*\to\infty} rhs(h^*) = \infty$. So for generic parameter values there are two possibilities: Either there exist two different h^*

such that $lhs = rhs(h^*)$. In this case there are two interior steady states. On the other hand, it is possible that $lhs < rhs(h^*)$ for all $h^* > 1$ which implies that there is no interior steady state.

By example we show that both possibilities are indeed feasible. Take first the case $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, $\delta = 0.01$ and then the case. $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$, $\delta = 0.04$. In the first case condition 27 can be written as:

$$\begin{aligned} 0 &= (1-0.3)\left(\frac{(0.1)(0.5)}{0.7}\right)^{0.3} - \left(\frac{(0.1)(1+0.5)}{0.7}\right)(h^*-1)^{\left(\frac{1}{0.7}-2\right)(1-0.3)} \\ &\quad -\frac{1}{h^*}(h^*-1)^{\frac{1}{0.7}-\left(\frac{1}{0.7}-2\right)0.3} = F(h^*). \end{aligned}$$

and in the second case it can be written as:

$$\begin{aligned} 0 &= (1-0.3)\left(\frac{(0.04)(0.5)}{0.7}\right)^{0.3} - \left(\frac{(0.04)(1+0.5)}{0.7}\right)(h^*-1)^{\left(\frac{1}{0.7}-2\right)(1-0.3)} \\ &\quad -\frac{1}{h^*}(h^*-1)^{\frac{1}{0.7}-\left(\frac{1}{0.7}-2\right)0.3} = G(h^*). \end{aligned}$$

When trying to solve $F(h^*) = 0$ for h^* one finds no solution, whereas solving $G(h^*) = 0$ gives exactly two solutions: $h_1^* = 1.09496$ and $h_2^* = 1.2675$.

These results are illustrated by Figure 5.

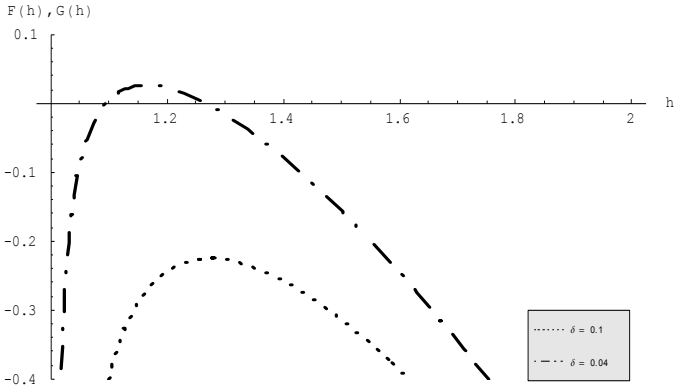

Figure 5: Endogenous education attitude with $\beta = 0.7$, $\delta = 0.1$ and $\delta = 0.04$

As one can easily see in the first case there is no h^* such that $F(h^*) = 0$, whereas in the second case there are two h^* such that $G(h^*) = 0$: $h_1^* = 1.09496$ and $h_2^* = 1.2675$.

6.5 Proof of Lemma 5

Note first that $\widehat{h}_{t+1}$ and $\widehat{k}_{t+1}$ have to be finite for all finite values of $(\widehat{k}_t, \widehat{h}_t)$. The solution might be either interior (in which case the first order conditions hold) or at the lower bounds. Since $\widehat{k}_t > 0$, production takes place in period t. Combining this fact with the INADA condition of the individual utility functions implies that $\widehat{k}_t^{\alpha}\widehat{h}_t^{1-\alpha} - \widehat{k}_{t+1} - \left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}} - \widehat{k}_t\widehat{r}_t = c_{2,t} > 0$. Therefore, $\frac{\partial W}{\partial k_{t+1}} = \infty$ at $k_{t+1} = 0$. Hence, $\widehat{k}_{t+1} > 0$, and condition (10) holds.

As for the solution for the human capital, note again that since $\widehat{k}_t > 0$, production takes place in period t. Again, combining this fact with the Inada condition of the individual utility functions implies that $\widehat{k}_{t+1}^{\alpha}\widehat{h}_{t+1}^{1-\alpha} - \widehat{k}_{t+2} - \left(\widehat{h}_{t+2} - 1\right)^{\frac{1}{\beta}} - \widehat{k}_{t+1}r_{t+1} = c_{2,t+1} > 0$. Hence $\frac{\partial W}{\partial h_{t+1}} = \infty$ at $h_{t+1} = 1$, implying $\widehat{h}_{t+1} > 1$. This gives condition (9), which completes the proof. ■

6.6 Proof of Proposition 6

i) By combining (12), (13) and the balanced budget condition of the government (11) one gets:

$$\frac{\varphi(\overline{h}_t)}{\overline{h}_{t+1}} - (1 - s_t)\frac{(\gamma+1)}{\beta}\frac{\left(\overline{h}_{t+1} - 1\right)^{\frac{1}{\beta}-1}}{\left(\overline{w}_t - \left(\overline{h}_{t+1} - 1\right)^{\frac{1}{\beta}}\right)} = 0. \tag{30}$$

Furthermore combining (9) and (10) gives:

$$\frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}} - \frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}}} + \frac{\omega_{t+1}}{\omega_t}\left(\xi_t + \varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right) = 0. \tag{31}$$

In order to establish the optimal subsidy rate which ensures that $\overline{h}_{t+1} = \widehat{h}_{t+1}$ we set (30) equal to (31), furthermore set $\overline{h}_{t+1} = \widehat{h}_{t+1}$ and solve for s_t.

$$\begin{aligned} & \frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}} - (1 - s_t)\frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}}} \\ = \; & \frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}} - \frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1} - 1\right)^{\frac{1}{\beta}}} + \frac{\omega_{t+1}}{\omega_t}\left(\xi_t + \varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right). \end{aligned}$$

This can be written as:

$$\frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}} - \frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}} + s_t\frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}}$$

$$= \frac{\varphi(\widehat{h}_t)}{\widehat{h}_{t+1}} - \frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}} + \frac{\omega_{t+1}}{\omega_t}\left(\xi_t + \varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right).$$

From which it follows:

$$s_t\frac{(\gamma+1)}{\beta}\frac{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}{\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}} = \frac{\omega_{t+1}}{\omega_t}\left(\xi_t + \varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right).$$

Solving for s_t gives:

$$s_t = \frac{\beta}{(\gamma+1)}\frac{\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}}}{\left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}-1}}\frac{\omega_{t+1}}{\omega_t}\left(\xi_t + \varphi'(\widehat{h}_{t+1})\ln\widehat{h}_{t+2}\right).$$

ii) It is obvious that (15) implies that the balanced budget condition is fulfilled whenever $\overline{h}_{t+1} = \widehat{h}_{t+1}$.

iii) Since $\xi_t > 0$, $\varphi'(\widehat{h}_{t+1}) > 0$, $\widehat{h}_{t+2} > 1$, and $\widehat{w}_t - \left(\widehat{h}_{t+1}-1\right)^{\frac{1}{\beta}} = c_{2,t} + k_{t+1} > 0$, the optimal subsidy rate $s_t > 0$.

On the other hand, if $s_t = 1$ the price parents have to pay for the human capital of the children would be zero. Therefore, demand for education would be infinite, which is of course not feasible. Hence, the optimal subsidy rate must fulfill $0 < s_t < 1$. ■

Procedural Concerns and Reciprocity

Procedural Concerns and Reciprocity*

Alexander Sebald†

Abstract

Different to other scientific disciplines traditional economic theory has remained remarkably silent about procedural aspects of strategic interactions. Much to the contrast, among psychologists there is by now a broad consensus that not only expected outcomes shape human behavior, but also procedures that are used to take decisions. It is argued that procedural concerns are especially pervasive in the resolution of conflicts. In our paper we show that procedural concerns are in fact an inherent feature of the interaction of reciprocal agents. More precisely, using Dufwenberg and Kirchsteiger (2004)'s theory of sequential reciprocity we demonstrate that procedural choices determine the responsibility that people have for outcomes. The responsibility for outcomes in turn influences peoples' evaluations of intentions and, hence, subsequent reactions. Two applications are discussed to highlight the impact and importance of procedural concerns in strategic interactions.

Keywords: Psychological Games, Procedural Concerns, Reciprocity
JEL Classification: D01, C70

Introduction

Imagine a group of three friends. One of them has a free ticket for the local concert of their favorite music band. Unfortunately, however, he cannot go himself, as he has an exam the following day. As his friends love the band as much as he does, he would like to give the ticket to one of them instead. He is indifferent as to whom of the two to give it. He knows, however, that if one of them feels unkindly treated, he will get into a quarrel. It is easy to see that this situation bears much resemblance to the '*So long, Sucker'* game analyzed e.g. by Nalebuff and Shubik

*I am very grateful to Martin Dufwenberg, Georg Kirchsteiger, Pierpaolo Battigalli, Estelle Cantillon, Paolo Casini and the seminar participants at ECARES/ULB and Maastricht University for helpful comments.

†Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands, and ECARES, Université Libre de Bruxelles. Sebald is also member of ECORE, the recently created association between CORE and ECARES. E-mail: a.sebald@algec.unimaas.nl

(1988). A player (A), i.e. the *ticket holder*, is driven to choose an *unlucky* player, i.e. the friend that does not receive the ticket, out of two players (B) and (C). Subsequently the *unlucky* player is allowed to choose an action which is either kind, i.e. not quarreling, or unkind, i.e. quarreling, towards player (A). As in the '*So long, Sucker*' game, it seems also here, at first sight, that the *ticket holder* is trapped: By choosing who gets the ticket he inevitably has to be unkind to one of his friends, creating the risk of trouble. At a second glance, however, when asked how this conflict could be resolved, one is intuitively driven to suggest that he should flip a coin to take the decision as in this way he avoids being unkind to either of them.

This example and our intuition of how to resolve the conflict effectively highlight two essential aspects of any human interaction. First, very often there are numerous ways in which decisions can be taken. On the one hand, the friend holding the ticket could decide to take the decision himself as to whom to give it, but, on the other hand, he could also let *chance* decide by flipping a coin. Secondly, one can easily see that decisions are inherently associated with *procedures* which characterize the way in which they are taken. The *ticket holder*, in our example, first has to decide how he wants to take the decision before he can effectively take it.

Among psychologists there is by now a broad consensus that not only expected outcomes shape human behavior, but also *procedures* that are used to take decisions [e.g. Thibaut and Walker (1975), Lind and Tyler (1988), Collie et al. (2002), Anderson and Otto (2003) and Blader and Tyler (2003)]. It is argued that *procedural concerns* are especially pervasive in the resolution of conflicts. Prominent examples of conflict resolutions are to be found in the areas of workplace relations and the public acceptability of policies and laws. First, psychologists have found evidence that behavioral reactions to promotion decisions, bonus allocations, dismissals etc. strongly depend on the perceived fairness of selection *procedures* [e.g. Lemons and Jones (2001), Konovsky (2000), Bies and Tyler (1993), Lind at al. (2000) and Roberts and Markel (2001)]. Second, it has been shown that public compliance with policies and laws strongly depends on the perceived fairness of their enforcement *procedures* [e.g. Tyler (1990), Wenzel (2002), Murphy (2004), De Cremer and van Knippenberg (2003) and Tyler (2003)].

Psychologists explain the impact of *procedures* on human interactions with the help of attribution theory [e.g. Heider (1958), Kelley (1967), Kelley (1973), Ross and Fletcher (1985)]. Attribution theory rests on the assumption that people need to infer causes and assign responsibilities for why outcomes occur. It is argued that especially when outcomes are unfavorable and perceptions of intention are strong, there is a tendency to assign responsibility for outcomes to people. The assignment of responsibility and blame in turn has been shown to affect the occurrence and intensity of anger and aggression [Blount (1995)]. In other words, people care about others' intentions and reciprocate kind with kind and unkind with unkind behavior. As *procedures* explicitly influence the control that people have over final outcomes, they obviously also influence the evaluation of responsibilities and intentions. To exemplify, imagine a workplace situation in which a principal wants

to promote one out of two agents. If he chooses to take the decision on who is to be promoted intransperantly behind closed doors, agents are driven to attach a high degree of responsibility for the outcome to the principal. His choice is interpreted as intentional, which fosters perceptions of favoritism. If, by contrast, the principal uses a transparent *procedure* which credibly shows that the decision is based on an unbiased criterion, i.e. a criterion which 'a priori' ensures that both agents have the same *chance* to be promoted, the principal is not blamed for the final outcome.

In line with attribution theory Blount (1995) experimentally showed that the responder behavior in ultimatum games is very sensitive to the way, i.e. *procedure*, in which a proposal is made. In her experiments proposals in the ultimatum game were either made by a proposer actively having a stake in the final outcome of the game, by a neutral third party not having any monetary stake in the final outcome or by *chance*. She observed that the same proposal triggered significantly lower rejection rates in case a neutral third party or *chance* had chosen the proposal compared to situations in which the proposal was made by a stakeholder. According to attribution theory lower rejection rates in case of neutrality of the proposer or explicit randomizations hint at the fact that responders attach a lower degree of responsibility and intentionality for outcomes to other stakeholders as they do not have any influence over proposals. In other words, the responders' willingness to punish other stakeholders seems to decrease the lower the others' influence over the final division of the pie.

Notwithstanding this experimental evidence and the fact that e.g. workplace relations play an eminent role in the economic literature, economists have remained remarkably silent so far about the impact of *procedures* on human behavior in strategic interactions. Only three recent economic papers have started to address the issue of *procedural choices* in strategic interactions [Bolton et al. (2005), Trautmann (2006), Krawczyk (2007)]. In contrast to attribution theory, however, they all extend models of distributional concerns to account for the impact of *procedural choices* on strategic behavior. Bolton et al. (2005) only present a sketch of a possible model based on the model of inequity aversion by Bolton and Ockenfels (2000). Trautmann, on the other hand, manipulates Fehr and Schmidt (1999)'s model of inequality aversion suggesting that agents' utilities depend on 'expected outcome differences' 'ex ante' as well as 'ex post' to any outcome realization. In the context of our introductory example this means that even after the flipping of a coin the *ticket holder*'s utility depends on the 'ex ante' expected outcome difference. The expected outcome differential for his friends is lowest when flipping a coin. Hence, an inequality avers *ticket holder* would prefer flipping a coin to any other *procedure* because it ensures a zero expected outcome differential. Although Trautmann's functional form is able to accommodate the experimental finding that rejection rates in random ultimatum games are lower than in the standard ultimatum games, it can only be applied to single decision situations. It cannot be applied to more complicated strategic interactions as the calculation of expected payoffs needs expectations about the other player's play.

In contrast, our paper follows the psychologists' view. As a main result, us-

ing psychological game theory we show that *procedural concerns* are an inherent feature of the interaction of reciprocal agents. We first formally define the concepts of *procedural game* and *procedure* and, secondly, use the 'theory of sequential reciprocity' by Dufwenberg and Kirchsteiger (2004) to highlight the impact of *procedural choices* on the interaction of reciprocal agents. As will be shown, *procedural choices* determine the attribution of responsibilities and the evaluation of intentions. Responsibilities and intentions, in turn, determine the degree of any subsequent reciprocation. In brief, *procedures* are associated with explicit probability distributions defined over pure actions. In our concert-ticket example the two pure actions of friend (A) obviously are: i) giving the ticket to friend (B) and ii) giving the ticket to friend (C). The flipping of a coin assigns the probability $\frac{1}{2}$ to both of them. The more skewed this probability distribution is towards a certain pure action, the stronger the impression that the decision maker is intentionally aiming at this outcome. At the extreme this means, if friend (A) takes the decision directly, i.e. without explicitly randomizing, to give the ticket to friend (B), the *unlucky* friend (C) assigns full responsibility and intentionality to the decision of friend (A). In this situation player (C)'s kindness perceptions are obviously shaped by the fact that player (A) has directly chosen player (B) without giving him any 'credible' *chance* to also get the ticket.

Dufwenberg and Kirchsteiger (2004)'s class of sequential games does not allow for different *procedural choices*. More precisely, it only allows for one type of *procedures*: *procedures* that imply full responsibility and intentionality. To the contrary of this, in our class of *procedural games* we allow for different *procedural choices* which then allows to analyze the impact of *procedural choices* on strategic interactions. To exemplify, when player (A) in our introductory example decides to take his decision by flipping a coin instead of taking the decision himself both his pure actions, i) and ii), are 'ex ante' equally probable. The outcome is pure *chance* and, hence, no responsibility and intentionality is associated with it. As a consequence, reciprocal agents react differently to the same outcomes, i.e. choice of pure actions, depending on the *procedure* which has led to them.

To highlight this impact of *procedural choices* on the strategic interactions of reciprocal agents we analyze two applications in the final section of this paper. More precisely, we allow for different *procedures* in the '*So long, Sucker'* game analyzed by Nalebuff and Shubik (1988) as well as Dufwenberg and Kirchsteiger (2004) and the *Sequential Prisoners Dilemma* also analysed by Dufwenberg and Kirchsteiger (2004). Comparing our results to their equilibrium predictions shows that the interaction of reciprocal agents is very sensitive to the availability of different *procedures*.

The organization of the paper is as follows: In the next section we formally define *procedures* and characterize a *procedural game* in which agents choose for *procedures* rather than actions and strategies. In the second section we point at the impact of *procedures* on the behavior of reciprocal agents. More precisely, we formally define reciprocity in the context of our *procedural game* and in this way explain the impact of *procedural choices* on the strategic interaction of reciprocal agents. We furthermore show that the concept of *sequential reciprocity equilibria*

(SRE) defined by Dufwenberg and Kirchsteiger (2004) can also be applied to our class of *procedural games* in which agents choose for *procedural strategies*. Finally, as said above, two applications are discussed to highlight the impact and importance of *procedural concerns* in strategic interactions.

Procedures

In this section we proceed in two steps. First, we intuitively sketch our argument with the help of two examples. In a second step we *i*) formally define the concept of *procedures* and *ii*) fully characterize our class of *procedural games* in which agents do not choose actions and strategies, as usually assumed in game theory, but *procedures*. This class of multi-stage games in which agents choose *procedures* is thenceforth used in the subsequent sections to analyze the impact of *procedural choices* on the strategic interaction of reciprocal agents.

As a starting point consider games Γ_1 and Γ_2 in Figure 1 and 2:

[Figure 1 and 2 here]

The sole difference between games Γ_1 and Γ_2 is that in Γ_2 player 1 can choose (M) on top of his pure actions (L) and (R). Player 1's pure action (M), however, is nothing else than choosing an explicit randomization device, (0), assigning probabilities α_2 and $(1-\alpha_2)$ to his pure actions (L) and (R) respectively. 'Flipping a coin' or 'throwing a dice' constitute explicit randomization devices, for example. 'Flipping a coin' assigns the probability $\frac{1}{2}$ to both pure actions (L) and (R). 'Throwing a dice', on the other hand, leads to $\alpha_2 = \frac{5}{6}$ and $(1-\alpha_2) = \frac{1}{6}$, if, for example, (L) is chosen, whenever numbers 1 to 5 come up, and (R) is chosen, if 6 appears. Obviously, 'flipping a coin' and 'throwing a dice' are but two *credible ways* in which a decision can be taken. In reality one usually disposes of many different *ways*. Nevertheless the two examples suffice to show how different *ways*, or in our words explicit randomization devices, are associated with differing explicit probability distributions with which an action is indirectly chosen by *chance*.

But not only choices like (M) can be characterized as choices for explicit randomization devices. Taking the thought about the *credible ways* and the differing explicit probability distributions to the extreme shows that also pure actions like (L) and (R) can equally be defined as choices for explicit randomization mechanisms. Imagine, for example, that player 1 in Γ_1 and Γ_2 chooses for his pure actions (L). This is equivalent to saying that player 1 chooses for *chance* to take the decision between (L) and (R) assigning probability 1 to his pure action (L). Hence, although (L) represents a pure action, it can nevertheless be reinterpreted in a way in which the decision is indirectly taken by *chance* randomizing with a degenerated probability distribution over the set $\{(L),(R)\}$.

This shows that in our two examples, Γ_1 and Γ_2, any choice for a pure actions, i.e. (L) and (R), and any choice for an explicit randomization mechanism, i.e. (M), can likewise be reinterpreted as a choice for an explicit randomization device through which the actual decision is subsequently taken by *chance*. Consider, for

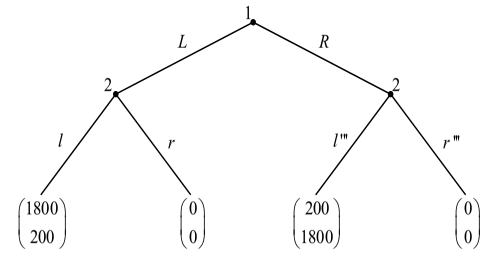

Figure 1: Game Γ_1

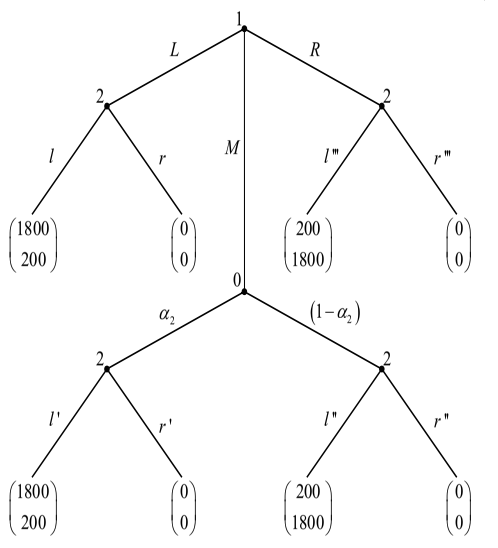

Figure 2: Game Γ_2

example, game Γ_3 in Figure 3, which is a restatement of game Γ_2 in the spirit of this intuition:

[Figure 3 here]

As one can see, in Γ_3 we reformulate all strategic choices of game Γ_2 into choices for explicit randomization mechanisms, i.e. *chance* or player 0, through which decisions are subsequently taken. In game Γ_2 player 1 can decide between (L), (M) and (R), and player 2 can decide between (l) and (r), (l') and (r'), (l'') and (r'') or (l''') and (r''') depending on player 1's choice. Equivalently, in game Γ_3 player 1, for example, has to decide between the explicit randomization devices $\omega(h_1^0)$, $\omega'(h_1^0)$ and $\omega''(h_1^0)$ in the initial history h_1^0. First, by choosing $\omega(h_1^0)$ he can decide to let *chance* take the decision between (L) and (R) assigning probability 1 to (L). Second, by choosing $\omega'(h_1^0)$ he can decide to let *chance* take the decision between (L) and (R) assigning probability α_2 to (L) and $(1-\alpha_2)$ to (R). Finally, by choosing $\omega''(h_1^0)$ he can decide to let *chance* take the decision between (L) and (R) assigning probability 1 to (R). In all these three cases player 1 only determines how *chance* subsequently takes the decision, rather then taking the decision himself. Hence, notwithstanding the formal equivalence between games Γ_2 and Γ_3, an interpretive difference exists. Choosing for an explicit randomization mechanism implies that players do not take decisions themselves. They merely determine how decisions are taken by *chance*. In other words, players decide about the *procedures* which are used to take decisions. The example in Figure 3, thus, uncovers that strategic decision making is not only about choosing actions but also about *how* actions are chosen. For this reason we call game Γ_3 a *procedural game*.

This brings us to a more formal definition of our class of *procedural games*. Formally, let the set of players be $\mathcal{N} = \{0, 1, ..., N\}$ where 0 denotes the uninterested player *chance*. Denote as $\mathcal{H}$, with the empty sequence $\varnothing \in \mathcal{H}$, the finite set of histories, h, and $\mathcal{X}$ the finite set of decision nodes x, such that h^x is the sequence of decisions on the path to the decision node x. The player function, $\mathcal{C}$, assigns to each nonterminal history $h^x \in \mathcal{H}$ a member $i \in \mathcal{N}$ who moves after that history h^x. Therefore, let h_i^x be the history h on the path to the decision node x which is controlled by player $i \in \mathcal{N}$ and $\mathcal{H}_i$ the set of all histories after which player i has to move throughout the game. At each history, h_i^x, after which player $i \in \mathcal{N} \setminus \{0\}$ has to move, he disposes of a nonempty finite set of pure actions $\mathcal{A}(h_i^x)$ and a finite set of explicit randomization devices, $\Omega(h_i^x)$, through which he can choose an action from $\mathcal{A}(h_i^x)$. As already suggested in example Γ_3 players in our *procedural games* do not choose actions $a \in \mathcal{A}(h_i^x)$ directly, but choose explicit randomization mechanisms, denoted $\omega(h_i^x) \in \Omega(h_i^x)$, through which a decision is indirectly taken by *chance*. The choice for a specific explicit randomization device, $\omega(h_i^x)$, in history h_i^x by player $i \in \mathcal{N} \setminus \{0\}$ leads to a specific decision node $v \in \mathcal{X}$ defined by h_0^v in which *chance* takes the actual decision using the explicit probability distribution $\rho(\omega(h_i^x))$ associated with $\omega(h_i^x)$ defined on $\mathcal{A}(h_0^v)$, with $\mathcal{A}(h_0^v) = \mathcal{A}(h_i^x)$. Hence, the choice for a pure action a (e.g. (L) in Γ_2), for example, translates in our *procedural game* into a choice for an explicit randomization

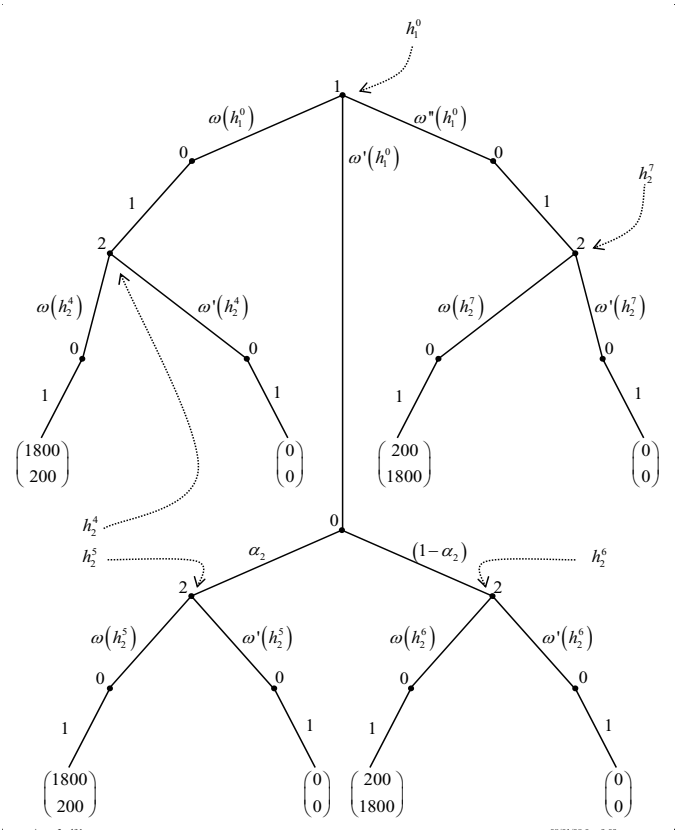

Figure 3: Game Γ_3

mechanisms, $\omega(h_i^x)$, that is associated with a degenerated probability distribution $\rho(\omega(h_i^x))$ which assigns probability 1 to the pure action a in the set of possible actions $\mathcal{A}(h_0^v) = \mathcal{A}(h_i^x)$. The choice for an explicit randomization (e.g. (M) in Γ_2), on the other hand, is a choice for an explicit randomization mechanism, $\omega'(h_i^x)$, that is associated with a non-degenerated probability distribution $\rho(\omega'(h_i^x))$ defined on $\mathcal{A}(h_0^v) = \mathcal{A}(h_i^x)$. As said before, the set of player i's degenerated as well as non-degenerated explicit randomization mechanisms in any history h_i^x is $\Omega(h_i^x)$. The associated set of explicit probability distributions is furthermore denoted as $\mathcal{P}(h_i^x)$, where $\mathcal{P}(h_i^x) = \{\rho(\omega(h_i^x)) \mid \omega(h_i^x) \in \Omega(h_i^x)\}$. It can easily be seen that the minimum number of explicit randomization mechanisms that a player can decide between in any history h_i^x in our *procedural game* equals the number of pure actions that he has in the traditional extensive form representation.

As said before, by choosing for randomization devices players do not take decisions directly but only determine how *chance* subsequently takes them. Intuitively, as players only decide on how the decisions are subsequently taken, they only decide on the *procedure*, which is used to take a decision.

This brings us to a formal definition of *procedures*:

Definition 1 *A procedure, $\omega(h_i^x) \in \Omega(h_i^x)$, for player $i \in \mathcal{N} \setminus \{0\}$ in history $h_i^x \in \mathcal{H}_i$ is a tuple:*

$$\langle \rho(\omega(h_i^x)), \mathcal{A}(h_0^v) \rangle,$$

where:

1. *$\rho(\omega(h_i^x))$ is the explicit probability distribution associated with $\omega(h_i^x)$ defined on $\mathcal{A}(h_0^v)$*
2. *$\mathcal{A}(h_0^v) = \mathcal{A}(h_i^x)$, and*
3. *h_0^v directly succeeds h_i^x.*

In example Γ_3 *procedures* are used to choose for pure actions. We do not exclude, however, the possibility of *procedures* that choose between *procedures* and *procedures* that choose between *procedures* that choose between *procedures* etc. Procedures, $\omega(h_i^x) \in \Omega(h_i^x)$, rather have to be understood as reduced *procedures*. At any history h_i^x the explicit probability distribution associated with a reduced *procedure*, $\rho(\omega(h_i^x)) \in \mathcal{P}(h_i^x)$, basically subsumes the probability distributions of procedures of all levels into one explicit distribution defined on $\mathcal{A}(h_i^x)$. It is assumed that all players learn the outcome of a reduced *procedure* directly after its realization.

We denote a collection of *procedures* for any player $i \in \mathcal{N} \setminus \{0\}$ that specifies a *procedure* for each history after which player i moves a *procedural strategy*, ω_i. A *behavioral procedural strategy*, $m_i \in \mathcal{M}_i$, of player i, on the other hand, has to be understood as an implicit randomization at each history $h_i^x \in \mathcal{H}_i$ over the set of possible *procedures* $\Omega(h_i^x)$. Note, *procedural strategies*, $\omega_i \in \Omega_i$, and *behavioral procedural strategies*, $m_i \in \mathcal{M}_i$, in our class of procedural games are respectively the analogue to pure strategies and mixed strategies in the traditional extensive

form representation. We assume throughout that players choose for *behavioral procedural strategies*.

Given a *behavioral procedural strategy*, m_i, for each player $i \in \mathcal{N}\setminus\{0\}$ and the commonly known system of probability distributions, $\mathcal{P} = \cup_{i\in\mathcal{N}\setminus\{0\}}\mathcal{P}_i$, where $\mathcal{P}_i = \cup_{h_i^x\in\mathcal{H}_i}\mathcal{P}\left(h_i^x\right)$, we can compute a probability distribution over endnodes, $z \in \mathcal{Z}$. By assigning payoffs to endnodes, we can derive an expected payoff function, $\pi_i : \mathcal{Z} \to \Re$, for every player $i \in \mathcal{N}\setminus\{0\}$ which depends on what *behavioral procedural profile*, m in $\mathcal{M}$, where $\mathcal{M} = \times_{\mathcal{N}\setminus\{0\}}\mathcal{M}_i$, is played. In what follows we assume that payoffs are material payoffs like money or any other measurable quantity of some good.

Summarizing, a *procedural game* is a tuple:

$$\Gamma = \left\langle \mathcal{N}, \mathcal{M}, \mathcal{P}, (\pi_i : \mathcal{Z} \to \Re)_{\mathcal{N}\setminus\{0\}} \right\rangle. \tag{1}$$

This concludes the definition of *procedures* and the characterization of the class of *procedural games* which is the basis of our subsequent analysis. Starting from two simple examples, i.e. Γ_1 and Γ_2, we have formalized the idea that players choose for *procedures* rather than actions. In the remainder of the paper we use this class of *procedural games* in order to isolate the impact of *procedures* on strategic behavior. More precisely, the following section uses this characterization of *procedural games* to analyze the impact of *procedural choices* on the interaction of reciprocal agents.

Procedural choices and reciprocity

It is easy to see that if agents are only interested in their own expected material payoff, they would always behave the same in histories representing starting points of identical subgames. Looking again at game Γ_3 in Figure 3, for example, this means that players would react the same in histories h_2^4 or h_2^5. However, experimental evidence contradicts this. For example, in ultimatum games rejection rates for the same proposal significantly decrease if proposals are made by a random draw [Blount (1995) and Bolton et al. (2005)]. In other words responders' behaviors in ultimatum games significantly depend on *how* a certain proposal has come about. Psychologists have termed this dependence *procedural fairness* or *procedural concerns* and explain the observed behavior with the help of attribution theory. According to attribution theory agents behave reciprocally and evaluate the (un)kindness of themselves and others taking into consideration their as well as the others' possible influence on (expected) outcomes. The less influence people have over outcomes at the time of their decision the less they are held responsible for it. Therefore, in order to demonstrate how *procedural concerns* can theoretically be reconciled with economic theory, we broaden the behavioral presumption in this section by assuming that agents are reciprocal. This means we formally define reciprocity in the context of our *procedural game* and show how it can explain the aforementioned evidence on *procedural concerns*.

Generally speaking, reciprocity means that agents do not only care about their own material payoff but also about the intentions of others [e.g. Rabin (1993),

Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006)]. They act kindly or unkindly depending on whether others are kind or unkind to them. Before we can more formally characterize the motivation of reciprocal agents and precisely define kindness and perceived kindness, however, it is necessary to highlight four theoretical peculiarities: kindness and perceived kindness of any player towards/from any other player *i*) cannot be measured directly, *ii*) might change after different histories of a game, *iii*) should be unaffected by inefficient *procedural strategies* and *iv*) realizations of the moves of *chance*.

i) Kindness and perceived kindness cannot be measured directly as they depend on each player's *procedural strategies*, beliefs about the others' *procedural strategies* and beliefs about the others' beliefs. Therefore, to model kindness we assume that every player holds a belief over the *behavioral procedural strategies* as well as a belief over the other players' beliefs. In the spirit of Dufwenberg and Kirchsteiger (2004) we model beliefs as *behavioral procedural strategies,* $m_i \in \mathcal{M}_i, \forall i \in \mathcal{N} \setminus \{0\}$. However, in order to avoid confusion we introduce a separate notation for beliefs. Let $\mathcal{B}_{ij} = \mathcal{M}_j, \forall i, j \in \mathcal{N} \setminus \{0\}$ be the set of possible beliefs of player i about the *behavioral procedural strategy* of player j (i.e. first-order belief). Furthermore let $\mathcal{C}_{ijq} = \mathcal{B}_{jq} = \mathcal{M}_q, \forall i, j, q \in \mathcal{N} \setminus \{0\}$ be the set of possible beliefs of player i about the belief of player j about the *behavioral procedural strategy* of player $q \neq j$ (i.e. second-order belief). Obviously, players do not have beliefs about the moves of the player *chance*. They do know, however, the explicit probability distributions associated with them. Therefore, let $(a)_{h^x}$ denote the collection of all passed realizations of moves of *chance* on the path up to history h^x.

ii) Players are assumed to have initial first- and second-order beliefs about the other players. As the game unravels these beliefs might change, however. In order to capture this it is important to keep track of how each player's behavior, beliefs, kindness and kindness perceptions differ across histories. We do this by updating *behavioral procedural strategies* as well as first- and second-order beliefs at each history that players control. In the spirit of Dufwenberg and Kirchsteiger (2004) we therefore formally define an (updated) *behavioral procedural strategy* as:

Definition 2 *Let* $m_i \in \mathcal{M}_i$ *and* $h_i^x \in \mathcal{H}_i$*, let* $m_i(h_i^x) \in \mathcal{M}_i$ *be the (updated) behavioral procedural strategy that prescribes the same procedural choices as* m_i *except for the procedural choices of player i on the path to* h_i^x *which are made with probability 1.*

In correspondence with the collection of passed realizations of the moves of chance, $(a)_{h_i^x}$, the collection of passed *procedural choices* of player i on the path to h_i^x is denoted $(\omega_i)_{h_i^x}$. Hence, the updated behavioral *procedural strategy* $m_i(h_i^x)$ is identical to $(\omega_i)_{h_i^x}$ on the path to history h_i^x and identical to the initial *behavioral procedural strategy*, m_i, in all other histories. To exemplify consider again game Γ_3 in Figure 3. Let player 2's initial *behavioral procedural strategy* m_2 be an implicit randomization over his set of pure *procedures* at each history that he controls. Player 2 moves after history h_2^5, which means that the implicit randomization prescribed by his initial *behavioral procedural strategy* over his pure *procedural choices*, $\omega(h_2^5)$ and $\omega'(h_2^5)$, leads to some realization. Following this his updated

behavioral procedural strategy becomes such that the implicit randomization at h_2^5 is substituted by its realization, but all other *procedural choices* at histories not reached remain the same. The updating of beliefs is assumed to work in an analogous fashion. Let, for example, player 2's initial belief about player 1's *behavioral procedural strategy* be $b_{21} = (\omega(h_1^0))$. If later on he finds himself in history h_2^5 in game Γ_3, his updated belief about player 1's *behavioral procedural strategy* becomes $b_{21}(h_2^5) = (\omega'(h_1^0))$, where $b_{21}(h_2^5)$ is player 2's updated first-order belief in history h_2^5 about player 1's *behavioral procedural strategy*. This shows that, parallel to the definition of $m_i(h_i^x)$, the updated first order belief $b_{ij}(h_i^x)$ is identical to the passed procedural choices of player j on the path to h_i^x, $(\omega_j)_{h_i^x}$, and identical to the initial belief, b_{ij}, in all other histories.

A remark on mixed strategies and *procedures*. The concept of *psychological games* was first introduced by Geanakoplos et al. (1989). In their seminal work Geanakoplos et al. (1989) only allow for initial beliefs to enter utility functions. Dufwenberg and Kirchsteiger (2004) and more recently Battigalli and Dufwenberg (2007) have shown, however, that in modeling, for example, reciprocity in a sequential setting unreasonable conclusions might be drawn if utility functions only depend on initial beliefs.[1] They show that it is necessary to keep track of how beliefs change as play unravels. Two areas in which the updating of beliefs needs some further explanation are mixed strategies and beliefs in mixed strategies. Dufwenberg and Kirchsteiger (2004) allow for mixed strategies and also allow players to hold beliefs in mixed strategies. Mixed strategies in their setting should be interpreted in terms of frequencies with which pure choices are made in a 'population'. This interpretation then explains why players that possibly hold mixed beliefs about the action of some other player update their beliefs (as soon as they learn his choice) as if he had chosen his actions with probability 1, i.e intentionally. *Procedures*, in comparison to that, might assign probabilities to pure actions in equivalence to mixed strategies. As they are observable, however, players do not update their beliefs after learning their outcome. If a player, for example, uses the flip of a coin to take a decision, this is observed by other players. This observability and the fact that probabilities connected to *procedures* are common knowledge implies that *procedural choices* represent perfect signals about intentions. Consequently, player's beliefs are updated taking into account the degree with which specific outcomes are intentionally aimed at. Therefore, in contrast to Dufwenberg and Kirchsteiger (2004), in our setting players update their beliefs according to the observed *procedural choices* that players make.

iii) For the same reason as in Dufwenberg and Kirchsteiger (2004) we restrict our attention to the set of *efficient procedural strategies*, $\mathcal{E}_i$. The set of *efficient procedural strategies*, $\mathcal{E}_i$, is defined as:

[1]For a more detailed discussion of this issue refer to Battigalli and Dufwenberg (2007) and Dufwenberg and Kirchsteiger (2004).

$$\begin{aligned}\mathcal{E}_i \;=\; & \{m_i \in \mathcal{M}_i \mid \text{there exists no } m_i' \in \mathcal{M}_i \text{ such that for all } h_i^x \in \mathcal{H}_i, \\ & (m_j)_{j\neq i} \in \Pi_{j\neq i}\mathcal{M}_j, q \in \mathcal{N}\setminus\{0\} \text{ it holds that} \\ & \pi_q\left(m_i'\left(h_i^x\right),\left(m_j\left(h_i^x\right)\right)_{j\neq i}\right) \geq \pi_q\left(m_i\left(h_i^x\right),\left(m_j\left(h_i^x\right)\right)_{j\neq i}\right) \\ & \text{with strict inequality for some } \left(h,\left(m_j\left(h_i^x\right)\right)_{j\neq i},q\right)\Big\}.\end{aligned}$$

Strategic choices are inefficient if there exists at least one other choice which conditional on any history of play and subsequent choices by the others provides no lower material payoff for any player, and a higher expected material payoff for some player for some history of play and subsequent choices by the others. In other words any *behavioral procedural strategy* is inefficient if it involves 'wasteful play' following some history, $h_i^x \in \mathcal{H}_i$. As also pointed out by Dufwenberg and Kirchsteiger (2004), it is unreasonable to let kindness and perceived kindness be influenced by strategies or, in our context, *procedural strategies* that imply 'wasteful play'. More precisely, the fact that 'wasteful play' is possible should be irrelevant for drawing conclusions regarding the kindness of the others' 'efficient' choices.[2]

iv) As said above, kindness and perceived kindness should also be unaffected by the realizations of the move of *chance*. Intuitively this captures the idea that people are not held responsible for situations over which they had no control. Or, to put it positively, people are held responsible for situations in as much as they were/are able to influence them. To give an example, if the *ticket holder* in our introductory situation chose to flip a coin to allocate the concert ticket to one of his friends, the friends's kindness perceptions of the *ticket holder*'s choice would depend on his *procedural choice* even after the realization of the move of *chance*. He would not be held responsible for the realization itself as he was not able to influence it after he had taken the decision to flip a coin. Similarly, 'ex ante' the *ticket holder*'s kindness perception of his own choice is also based only on what he is able to influence, i.e. he does not hold himself responsible for the realization of the flip of the coin but only for his *procedural choice*. To capture this idea we define the *decision context* of a person i in any history h_i^x. In every history h_i^x the *decision context* comprises, first, all passed *procedural choices* on the path to history h_i^x, $(\omega)_{h_i^x}$, with $(\omega)_{h_i^x} = \left\{(\omega_i)_{h_i^x},...,(\omega_N)_{h_i^x}\right\}$. Remember, the knowledge of all passed *procedural choices* on the path to history h_i^x is included in the updated procedural strategies $m_i\left(h_i^x\right)$ and the updated first order beliefs $b_{ij}\left(h_i^x\right)$. Second, the *decision context* includes the realizations of the moves of *chance* on the path up to history h_i^x, $(a)_{h_i^x}$, and, third, the remaining explicit probability distributions, $(\mathcal{P})_{\neg h_i^x}$, where $\neg h_i^x$ indicates all histories beside the histories on the path up to h_i^x. Hence, formally speaking:

[2]For a more detailed discussion of this issue refer to Dufwenberg and Kirchsteiger (2004).

Definition 3 *The decision context in any history h_i^x is a tuple:*

$$\left\langle (\omega)_{h_i^x}, (a)_{h_i^x}, (\mathcal{P})_{\neg h_i^x} \right\rangle.$$

This means it is the collection of i) all passed procedural choices of all players on the path to h_i^x, $(\omega)_{h_i^x}$, ii) all passed realizations of the moves of chance on the path up to h_i^x, $(a)_{h_i^x}$, and iii) the unreached explicit probability distributions, $(\mathcal{P})_{\neg h_i^x}$.

Intuitively speaking the *decision context* can be understood as the 'informational background' which players use to evaluate their own kindness towards others and, hence, to take their decisions. It is also the 'informational background' which is used by other players in later stages to evaluate the kindness of passed choices by others. More precisely, the *decision context* helps to decide in how far others were consciously aiming at a certain decision, i.e. pure action, or whether it was by *chance* that it was chosen.

We can now capture the idea that players strive to be kind if treated kindly and are unkind if treated unkindly by assuming that every player $i \in \mathcal{N} \setminus \{0\}$ chooses a *behavioral procedural strategy*, m_i, that maximizes his utility defined as:

$$U_i = \pi_i + \sum_{j \neq i} Y_{ij} \cdot (\kappa_{ij} \cdot \lambda_{iji}), \tag{2}$$

where $i, j \in \mathcal{N} \setminus \{0\}$, κ_{ij} is the believed kindness of player i to player j and λ_{iji} is player i's belief about the kindness of player j towards himself.

More precisely, player i's utility is the sum of N terms. The first term π_i represents player i's self interest. It is his expected material payoff in any history h_i^x after which he moves. It obviously depends on his own *behavioral procedural strategy*, $m_i(h_i^x)$, his belief about the others' *behavioral procedural strategies*, $b_{ij}(h_i^x), \forall j \neq i$, all past outcomes/realizations of procedures $(a)_{h_i^x}$ until history h_i^x, and, finally, on the explicit probability distributions in all histories that have not been reached yet during the course of the game, $(\mathcal{P})_{\neg h^x}$. Hence:

$$\pi_i = \pi_i \left(m_i(h_i^x), (b_{ij}(h_i^x))_{j \neq i}, (a)_{h_i^x}, (\mathcal{P})_{\neg h^x} \right).$$

It can easily be seen that, as we allow for explicit randomizations in our class of *procedural games* our definition of expected material payoffs differs from the definition by Dufwenberg and Kirchsteiger (2004). It takes the player i's *decision context* in history h_i^x into account.

The following $N-1$ terms, $\sum_{j \neq i} Y_{ij} \cdot (\kappa_{ij} \cdot \lambda_{iji})$, in equation (2), on the other hand, represent player i's reciprocity payoff with respect to each other player $j \neq i$. The factor Y_{ij} is a non-negative reciprocity parameter which describes player i's sensitivity to the (un)kindness of player j. The higher Y_{ij} the more sensitive to reciprocity player i is. Finally the factors κ_{ij} and λ_{iji} capture respectively the kindness of player i to any other player j and player i's perceived kindness of player j towards him. Intuitively, kindness κ_{ij} is positive or negative depending on whether i is kind or unkind to j and perceived kindness λ_{iji} is positive (negative) if

player i beliefs player j to be kind (unkind) to him. Notice, reciprocity is captured by the factorial specification of the kindness parameters, κ_{ij} and λ_{iji}. It drives players to match perceived kindness (positive λ_{iji}) with kindness (positive κ_{ij}) and perceived unkindness (negative λ_{iji}) with unkindness (negative κ_{ij}).

This brings us to the formal definition of kindness, κ_{ij}:

Definition 4 *The kindness of player i to another player $j \neq i$ at any history $h_i^x \in \mathcal{H}$ is given by the function $\kappa_{ij} : \mathcal{M}_i \times \Pi_{j \neq i} \mathcal{B}_{ij} \to \Re$ defined as:*

$$\kappa_{ij} = \pi_j \left(m_i \left(h_i^x \right), \left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right) - \pi_j^{e_i} \left(\left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right).$$

The kindness of player i towards player j in history h_i^x is defined as the difference between the expected material payoff of player j , π_j, that player i intends to give j and the average expected material payoff, $\pi_j^{e_i} \left(\left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right)$, defined as:

$$\begin{aligned} & \pi_j^{e_i} \left(\left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right) \\ = \; & \frac{1}{2} \left[\max \left\{ \pi_j \left(m_i \left(h_i^x \right), \left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right) \mid m_i \left(h_i^x \right) \in \mathcal{M}_i \right\} \right. \\ & \left. + \min \left\{ \pi_j \left(m_i \left(h_i^x \right), \left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}, \left(a \right)_{h_i^x}, \left(\mathcal{P} \right)_{\neg h^x} \right) \mid m_i \left(h_i^x \right) \in \mathcal{E}_i \right\} \right]. \end{aligned}$$

Think of $\pi_j^{e_i}$ as a norm for i describing the 'equitable' payoff for player j when i's beliefs about the other players' behavior are summarized by $\left(b_{ij} \left(h_i^x \right) \right)_{j \neq i}$, the past realization on the path to h_i^x are $\left(a \right)_{h_i^x}$ and the unreached explicit probability distributions are given by $\left(\mathcal{P} \right)_{\neg h^x}$. Thus, when $\pi_j^{e_i} = \pi_j$ then player i's kindness towards player j is zero. Intuitively the above definition means that player i is kinder the more he expects to give player j relative to the average that he could give him given his beliefs about the other players play. To exemplify consider, for example, history h_2^5 of game Γ_3. The *behavioral procedural strategy* of player 2, $m_2 \left(h_2^5 \right)$, as well as his first-order belief over the profile of player 1, $b_{21} \left(h_2^5 \right)$, and the past realized move of nature, $\left(a \right)_{h_2^5} = \{(L)\}$, define history h_2^5. Furthermore, player 2's *behavioral procedural strategy* together with his first-order belief and the remaining probability distributions, $\left(\mathcal{P} \right)_{\neg h^x}$, on the other hand, define what player 2 is willing to give to player 1 in expected terms as well as what he could give him. Assume, for example, that player 2's *behavioral procedural strategy* in h_2^5 is $m_2 \left(h_2^5 \right) = \left(\omega \left(h_2^4 \right), \omega \left(h_2^5 \right), \omega \left(h_2^6 \right), \omega \left(h_2^7 \right) \right)$. It can easily be seen that player 2 intends to give player 1 $\pi_1 \left(h_2^5 \right) = 1800$, i.e. according to $m_2 \left(h_2^5 \right)$ he will choose $\omega \left(h_2^5 \right)$ after his history h_2^5. On the other hand, the average of the maximum and minimum which he could give to player 1 is $\pi_1^{e_2} \left(h_2^5 \right) = \frac{1}{2} \left(1800 \right) + \frac{1}{2} \left(0 \right) = 900$. Hence, player 2's kindness towards player 1 in h_2^5 is:

$$\begin{aligned} \kappa_{21} \left(h_2^5 \right) &= \pi_1 \left(h_2^5 \right) - \pi_1^{e_2} \left(h_2^5 \right) = 1800 - 900 \\ &= 900. \end{aligned}$$

The above definition of kindness is a necessary adaptation from Dufwenberg and Kirchsteiger (2004) in the context of our *procedural game*. It includes the *decision context* on which players base their decisions.

The definition of perceived kindness, λ_{iji}, also requires a change though. As said above, in the evaluation of intentions agents take into account in how far others were/are actually responsible for the unraveled play. Hence, it would be unreasonable to assume that player 2 in game Γ_3 perceived the kindness of player 1 in histories h_2^5 and h_2^6 differently. It is simply by *chance* that either of the two histories are reached. In order to capture this we assume that players always evaluate the other players' kindness on the basis of the *decision context* in which the others have taken their last *procedural choice*. Remember, a *decision context* characterizes the 'informational base' on which a decision is taken. As players know all past *procedural choices* as well as the realizations of moves of *chance* along the path up to h_i^x, they obviously not only know their own current *decision context*, but they can also deduce all past *decision contexts* which were the basis of the other players' last *procedural choices*. Denote the history in which any player $j \neq i$ has made his last procedural choice along the path up to h_i^x as $h_i^x\left(h_j^l\right)$. When player i evaluates the kindness of player j's *procedural choice* in history h_i^x, he, hence, uses player j's *decision context* in $h_i^x\left(h_j^l\right)$:

$$\left\langle (\omega)_{h_i^x\left(h_j^l\right)}, (a)_{h_i^x\left(h_j^l\right)}, (\mathcal{P})_{\neg h_i^x\left(h_j^l\right)} \right\rangle,$$

where $(\omega)_{h_i^x\left(h_j^l\right)}$ defines all past *procedural choices on the path to* h_i^x up to history h_j^l, $(a)_{h_i^x\left(h_j^l\right)}$ defines all past realizations of moves of *chance* on the path to history h_i^x up to history h_j^l and $(\mathcal{P})_{\neg h_i^x\left(h_j^l\right)}$ indicates all remaining explicit randomizations in h_j^l. Evaluating player j's kindness only on the basis of the *decision context* in which he has made his last *procedural choice* on the path up to history h_i^x ensures that player j is held solely responsible for the decisions that he has explicitly taken himself. To exemplify, in both histories h_2^5 and h_2^6 player 2 evaluates player 1's kindness on the basis of player 1's *decision context* at the history, h_1^0:

$$\left\langle (\omega)_{h_2^5\left(h_1^0\right)}, (a)_{h_2^5\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^5\left(h_1^0\right)} \right\rangle = \left\langle (\omega)_{h_2^6\left(h_1^0\right)}, (a)_{h_2^6\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^6\left(h_1^0\right)} \right\rangle,$$

in which player 1 had to take his last *procedural decision*, i.e. $h_j^l = h_1^0$. In other words, in histories h_2^5 and h_2^6 player 2 does not take the realization of the move of *chance* after history h_0^2 into account when evaluating the kindness of player 1. The realization of the move of *chance* after h_0^2 is by *chance* and hence not the responsibility of player 1.

Given this let perceived kindness be defined as:

Definition 5 *Player i's beliefs about how kind player $j \neq i$ is to i at history $h_i^x \in \mathcal{H}$ is given by the function $\lambda_{iji} : \mathcal{B}_{ij} \times \Pi_{i \neq j} \mathcal{C}_{iji} \to \Re$ defined as:*

$$\begin{aligned} \lambda_{iji} \;=\; & \pi_i\left(b_{ij}\left(h_i^x\right), \left(c_{ijq}\left(h_i^x\right)\right)_{q \neq j}, (a)_{h_i^x\left(h_j^l\right)}, (\mathcal{P})_{\neg h_i^x\left(h_j^l\right)}\right) \\ & -\pi_i^{ej}\left(\left(c_{iji}\left(h_i^x\right)\right)_{i \neq j}, (a)_{h_i^x\left(h_j^l\right)}, (\mathcal{P})_{\neg h_i^x\left(h_j^l\right)}\right), \end{aligned}$$

where $h_i^x\left(h_j^l\right)$ is the last history after which player j has moved on the path to h_i^x.

As one can see, similar to the definition of kindness also perceived kindness is defined as the difference between what player i beliefs to receive in expected material payoff relative to the average that he could have gotten. To exemplify, assume now again that players find themselves in history h_2^5 of game Γ_3. We have seen above that, given player 2's updated *behavioral procedural strategy*, his first-order belief and the past realizations of the moves of *chance* up to history h_2^5, player 2's kindness towards player 1 is 900 in h_2^5. In addition to player 2's updated first-order belief $b_{21}\left(h_2^5\right) = \left(\omega'\left(h_1^0\right)\right)$, let now player 2's updated second order belief be $c_{212}\left(h_2^5\right) = \left(\omega\left(h_2^4\right), \omega\left(h_2^5\right), \omega\left(h_2^6\right), \omega\left(h_2^7\right)\right)$. The kindness that player 2 perceives from player 1 is then given by:

$$\begin{aligned}
\lambda_{212}\left(h_2^5\right) &= \pi_2\left(b_{21}\left(h_2^5\right), c_{212}\left(h_2^5\right), (a)_{h_2^5\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^5\left(h_1^0\right)}\right) \\
&\quad -\pi_2^{e_1}\left(c_{212}\left(h_2^5\right), (a)_{h_2^5\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^5\left(h_1^0\right)}\right) \\
&= \left(\frac{1}{2}(1800) + \frac{1}{2}(200)\right) - \frac{1}{2}((1800) + (200)) \\
&= 0.
\end{aligned}$$

This means, player 2 has the impression in history h_2^5 that player 1 intends to give him $\pi_2\left(h_2^5\right) = 1000$. As 1000 is also the 'equitable' payoff that player 1 could have given to him, player 2 judges player 1's kindness to be 0. Now consider history h_2^4, on the other hand, which is the starting point of an identical subgame. Player 2's perceived kindness of player 1's *behavioral procedural strategy* given his updated beliefs, $b_{21}\left(h_2^4\right) = \left(\omega_1\left(h_1^0\right)\right)$ and $c_{212}\left(h_2^4\right) = \left(\omega\left(h_2^4\right), \omega\left(h_2^5\right), \omega\left(h_2^6\right), \omega\left(h_2^7\right)\right)$ is:

$$\begin{aligned}
\lambda_{212}\left(h_2^4\right) &= \pi_2\left(b_{21}\left(h_2^4\right), c_{212}\left(h_2^4\right), (a)_{h_2^4\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^4\left(h_1^0\right)}\right) \\
&\quad -\pi_2^{e_1}\left(c_{212}\left(h_2^4\right), (a)_{h_2^4\left(h_1^0\right)}, (\mathcal{P})_{\neg h_2^4\left(h_1^0\right)}\right) \\
&= (200) - \frac{1}{2}((1800) + (200)) \\
&= -800.
\end{aligned}$$

Hence, although h_2^4 and h_2^5 are starting points of identical subgames, players perceives the situations totally different, i.e. perceived kindness of 0 in h_2^5 vs. perceived kindness of -800 in h_2^4. It follows that as both histories are perceived differently, optimal reactions in one history might not be optimal in the other even though the subsequent situation seems to be the same. This exemplifies that reciprocal agents do care about the way a certain situation has come about or, in other words, reciprocity inherently leads to *procedural concerns*.

This completes the description of the reciprocal preferences in the context of our *procedural game*. Putting together the *procedural game*, Γ, as defined in (1) and the vector of utilities, $(U_i)_{i\in\mathcal{N}\setminus\{0\}}$, as defined in (2) we get a tuple

$$\Gamma^p = \left\langle \Gamma, (U_i)_{i\in\mathcal{N}\setminus\{0\}} \right\rangle. \tag{3}$$

We refer to Γ^p as a *procedural game with reciprocity preferences.* Note, as the 'psychological game with reciprocity preferences' defined by Dufwenberg and Kirchsteiger (2004) Γ^p is not a 'traditional game'. In line with Dufwenberg and Kirchsteiger (2004), utility functions, U_i, are defined on richer domains including subjective beliefs. Different to them, however, and also different to 'traditional games' agents in our setting choose for *procedures,* as defined in Definition (1), rather than actions and strategies.

As a solution concept for our class of *procedural games with reciprocity preferences* we propose the *sequential reciprocity equilibrium* (SRE) defined by Dufwenberg and Kirchsteiger (2004). This means, each player in each history chooses his optimal *procedure* given his beliefs. The players' initial first and second order beliefs are required to be correct, and following each history of play the beliefs are updated as explained above.

Let $\mathcal{M}_i\left(h_i^x, m\right)$ be the non-empty set of *behavioral procedural strategies* that prescribe, for each player $i \in \mathcal{N}\setminus\{0\}$, the same choices as the strategy $m_i\left(h_i^x\right)$ for all histories other than h_i^x. Given this, the *sequential reciprocity equilibrium* (SRE) in the context of our *procedural game with reciprocity preferences* is defined as:

Definition 6 *The profile $m^* = \left(m_i^*\right)_{i\in\mathcal{N}\setminus\{0\}}$ is a sequential reciprocity equilibrium (SRE) if for all $i \in \mathcal{N}\setminus\{0\}$ and for each history $h_i^x \in \mathcal{H}$ it holds that*

1. $m_i^*\left(h_i^x\right) \in$

$$\arg\max_{m_i\in\mathcal{M}_i\left(h_i^x,m\right)} U_i\left(m_i\left(h_i^x\right), \left(b_{ij}\left(h_i^x\right), \left(c_{ijq}\left(h_i^x\right)\right)_{q\neq j}\right)_{j\neq i}, (a)_{h_i^x}, (\mathcal{P})_{\neg h_i^x}\right),$$

2. $b_{ij} = m_j^*$ *for all* $j \neq i$,

3. $c_{ijq} = m_q^*$ *for all* $j \neq i$, $q \neq j$.

Condition 1 assures that a SRE is a strategy profile such that at history h_i^x player i makes choices which maximize his utility given his beliefs and given that he follows his equilibrium strategy at other histories. At the initial stage, conditions (2) and (3) guarantee that the initial beliefs are correct. At any subsequent history, condition (1) requires that beliefs assign probability one to the sequence of choices that define that history, but are otherwise as the initial beliefs.

Concluding, in this section we have formally defined the motivation of reciprocal agents in the context of our *procedural game* and have given a glimpse of the impact of *procedural choices* on the strategic interaction of reciprocal agents. In the following section we will more fully analyze the impact of *procedural choices* by applying the concept of the *sequential reciprocity equilibrium* to two examples.

Applications

The first application is the '*Sequential Prisoners Dilemma'* also analyzed by Dufwenberg and Kirchsteiger (2004). The second is the '*So Long, Sucker'* game in

the spirit of Nalebuff and Shubik (1988) and Dufwenberg and Kirchsteiger (2004). Note, a full description of the strategic interaction and all possible equilibria that might arise in these two situations is beyond the scope of this paper. We, therefore, limit the analysis to the characterization of only one equilibrium to demonstrate the impact and importance of *procedural concerns*. Results and intuitions are presented in this section, mathematical proofs are relegated to the Appendix.

Example 1: *Sequential Prisoners Dilemma*

Consider the *Sequential Prisoners Dilemma* in Figure 4: [1]

[Figure 4 here]

As can easily be seen, game Γ_4 is an adaptation of the sequential prisoners dilemma analyzed by Dufwenberg and Kirchsteiger (2004). The difference is that in Γ_4 player 1 cannot only choose to cooperate (c) and defect (d), but can also explicitly randomize by choosing *procedure* (r). One sequential reciprocity equilibrium is:

Result 7 *If player 1's and 2's sensitivity to reciprocity, Y_1 and Y_2, is such that*

$$0 < Y_1 < \frac{1}{2}$$

and

$$Y_2 > \frac{1}{4\alpha_2 - 3}$$

and player 1's procedure $r\left(h_1^0\right)$ is associated with an explicit probability distribution such that $1 > \alpha_2 > \frac{3}{4}$, then the SRE is given by player 1 choosing $r\left(h_1^0\right)$ in history h_1^0 and player 2 choosing $c\left(h_2^4\right), c\left(h_2^5\right), c\left(h_2^6\right)$ and $d\left(h_2^7\right)$ in histories h_2^4, h_2^5, h_2^6 and h_2^7 respectively.[3]

Proof: see Appendix.

The intuition is the following. If α_2 is such that $1 > \alpha_2 > \frac{3}{4}$, player 2 perceives player 1's *procedural choice* as kind. If, in addition, his sensitivity to reciprocity Y_2 is high enough, i.e. $Y_2 > \frac{1}{4\alpha_2 - 3}$, then he reciprocates player 1's kindness by choosing (c) in history h_2^6. At the same time player 2 punishes player 1 in equilibrium at history h_2^7 which is the starting point of a payoff equivalent subgame. The difference between histories h_2^6 and h_2^7 is that the explicit probability α_2 is such that player 2 perceives player 1's choice of (r) as kind. He does not attribute enough responsibility for the outcome, i.e. history h_2^6, to player 1 to make it worth while to punish him. Furthermore, since Y_1 is relatively small, player 1 is mainly interested by money and his expected monetary payoff is highest by playing (r) given that player 2 does not play (d) following player 1's choice of (r).

[3] For simplicity we denote the sensitivity of reciprocity as Y_i in example 1. In example 2 we stick to Y_{ij} as defined in equation (2) to avoid confusion.

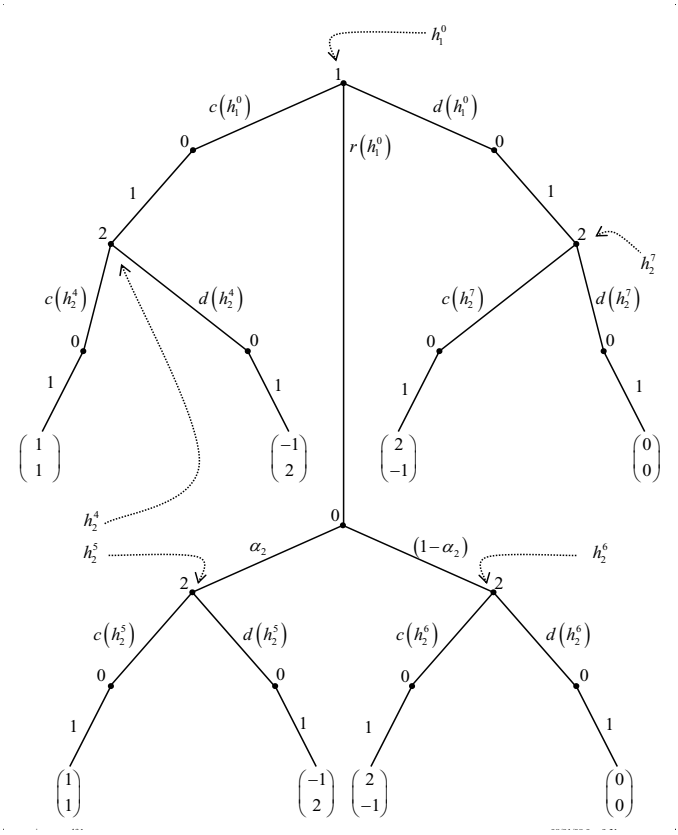

Figure 4: Game Γ_4

Dufwenberg and Kirchsteiger (2004), on the other hand, showed in the context of their setting that if player 2's sensitivity to reciprocity is strong enough, he cooperates if player 1 cooperates and defects if player 1 defects. Furthermore, they showed that if player 1's sensitivity to reciprocity is low and player 2's sensitivity is high, cooperation is player 1's equilibrium behavior for both monetary and reciprocity reasons.

Comparing the results shows that their equilibrium predictions are very sensitive to the availability of other procedures to take the same decision. As player 1 can also use procedure (r) to take his decision cooperation is no longer his optimal action given that player 2 is very sensitive to reciprocity. He chooses (r) because this makes player 2 to cooperate even in history h_2^6 which is identical to the subgame starting in h_2^7. Hence, *procedural choices* influence the kindness and perceived kindness of players and therefore influence the interaction of reciprocal agents.

Example 2: *The 'So Long, Sucker' Game*

In the following we will apply the concept of the *sequential reciprocity equilibrium* to the example, Γ_5, in Figure 5.[1] Example Γ_5 is an adaptation of the '*So Long, Sucker'* game also analyzed by Nalebuff and Shubik (1988) and Dufwenberg and Kirchsteiger (2004). With $\varepsilon = 0$, Γ_5 is a strategic situation in which player 1 has to decide on whom of two other players to give a zero payoff. Following his decision, the player who was unfavorably treated is called upon to decide whether player 1 should get 3 or whether both the others should equally get a payoff of 1. Intuitively it looks as if player 1 is 'a priori' worst off, as whoever he chooses will feel badly treated, and hence take revenge on player 1 by giving him the lowest possible monetary payoff.

[Figure 5 here]

However, if all players are solely motivated by purely selfish monetary concerns, this outcome is not guaranteed, as players 2 and 3 are indifferent between all their choices given that $\varepsilon = 0$. In order to allow for the possibility of revenge, Nalebuff and Shubik (1988) depart from the usual selfishness assumption, and assume that the players have lexicographically ordered objectives. This means that each player primarily maximizes his monetary payoff, but in case some choices yield exactly the same monetary payoff ties are broken so as to allow a player to take revenge. Dufwenberg and Kirchsteiger (2004), on the other hand, show that if agents behave reciprocally this outcome is also guaranteed for $\varepsilon \geq 0$. More precisely, they show that for any $\varepsilon \geq 0$ there exist sensitivities to reciprocity $Y_{21} > 0$ and $Y_{31} > 0$ for which taking revenge on player 1 is the best alternative for player 2 and 3. As said above, if players 2 and 3 are willing to take revenge even if it is costly, it seems that player 1 is trapped, as whatever he does, his action is perceived unkind by the player who has to take the subsequent decision.

As in the *Sequential Prisoners Dilemma* also in this application these results crucially depend on the fact that players 2 and 3 attribute full intentionality to

player 1. In other words, Nalebuff and Shubik (1988)'s and Dufwenberg and Kirchsteiger (2004)'s result is contingent on the unavailability of other *procedures* for player 1 to resolve the conflict between him and the other players. Consider game Γ_6 in Figure 6:

[Figure 6 here]

As can easily be seen, the only difference between games Γ_5 and Γ_6 lies in the fact that in the latter player 1 cannot only take his decision directly but can also e.g. flip a coin, i.e. choose $\omega_1'(h_1^0)$, to take it. Hence, he has an additional *procedure* which he can use to take his decision. It can be shown that with the help of this *procedure* player 1 can avoid the conflict with the others. More precisely:

Result 8 *If player 1, 2 and 3 have a sensitivity to reciprocity of*

$$Y_{12} = Y_{13} \geq 0,$$

$$Y_{21} \geq \frac{\varepsilon}{\varepsilon + 1}$$

and

$$Y_{31} \geq \frac{\varepsilon}{\varepsilon + 1},$$

then the only equilibrium is given by players 2 and 3 playing

$$\left(\omega_2'\left(h_2^4\right), \omega_2\left(h_2^5\right)\right)$$

and

$$\left(\omega_3'\left(h_3^6\right), \omega_2\left(h_3^7\right)\right)$$

respectively and player 1 choosing $\omega_1'(h_1^0)$.

Proof: see Appendix.

This means, if players 2 and 3 are enough sensitive to reciprocity, they will punish player 1, if he chooses one of them directly, and will be kind to him, if he chooses to take the decision by e.g. 'flipping a coin'. Knowing this, player 1 will choose to flip a coin, given that his sensitivity to reciprocity is equal for players 2 and 3, as this gives him a higher monetary as well as reciprocity payoff. In other words, by choosing e.g. to flip a coin, player 1 can get out of his 'trap'. Players 2 and 3 respectively perceive player 1's *procedural strategy* $\omega_1(h_1^0), \omega_1''(h_1^0)$ as unkind and $\omega_1'(h_1^0)$ as kind. If player 1 chooses e.g. to flip a coin, they do not attribute the outcome of the randomization to player 1, as he is only responsible for choosing the *procedure* but not for the outcome itself. Player 1, on the other hand, chooses $\omega_1'(h_1^0)$ for monetary as well as reciprocity reasons.

This highlights ones more how *procedural choices* influence the strategic interaction of reciprocal agents.

Conclusion

As we have seen, any decision in human interactions is inherently associated with a *procedure* which characterizes the way in which the decision is taken. This means it is impossible to take a decision without deciding on *how* to take it. It is widely accepted in other scientific disciplines and it has been shown experimentally that people react differently to identical outcomes depending on the *procedures* which have led to them. Hence, people are concerned about the way in which decisions are taken. Nevertheless economic theory has so far neglected the impact of *procedural choices* on human interaction. It has ignored *procedural concerns* as traditional economic theory is based on consequentialist preferences. However, if preferences are solely outcome oriented, it can hardly be explained why people should react differently to 'outcomewise' identical situations which only differ in the *procedures* which have led to them.

Only in recent years theories of reciprocity have contested the consequentialist view in economic theory by assuming that agents also receive a psychological payoff which, broadly speaking, depends on the agents' perceived intentions of others. As said before, when people behave reciprocally they evaluate the intentions of others and reciprocate kind with kind and unkind with unkind behavior. The evaluation of intentions is implicitly connected to the assignment of responsibilities for outcomes. The assignment of responsibilities, in turn, is related to the amount of control that people have over outcomes. It has been shown in our paper that *procedural choices* influence the control that people have over outcomes and, hence, influence the attribution of responsibilities and the evaluation of intentions. Dufwenberg and Kirchsteiger (2004)'s theory of sequential reciprocity captures situations in which agents have full control over outcomes and, hence, are held fully responsible for all consequences of their actions. In contrast to this, in our class of *procedural games* agents can choose between different *procedures*, which differ in the probabilities that they assign to outcomes. Given this we show, in line with attribution theory, that the less influence people have on outcomes the less responsibility and intentionality is attributed to them.

By defining a class of *procedural games* we have been able to distinguishing between *procedures* which are used to take decisions and the decisions themselves. Furthermore, assuming reciprocal agents and defining the *decision context* as the 'informational background' which any decision is based upon, we have demonstrated that *procedural concerns* are actually an inherent feature of any interaction of reciprocal agents.

List of References

1. Anderson, R. A. and Otto, A. L., 2003. Perceptions of fairness in the justice system: A cross-cultural comparison. Social Behavior and Personality. 31, 557-564.

2. Battigalli, P. and Dufwenberg, M., 2007. Dynamic Psychological Games. forthcoming in Journal of Economic Theory.

3. Bies, R. J. and Tyler, T. R., 1993. The "litigation mentality" in organizations: A test of alternative psychological explanations. Organization Science. 4, 352-366.

4. Blader, S. L. and Tyler, T. R., 2003. A four-component model of procedural justice: Defining the meaning of a "fair" process. Personality and Social Psychology Bulletin. 29, 747-758.

5. Blount, S., 1995. When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. Organizational Behavior and Human Decision Processes. 63, 131-144.

6. Bolton, G. and Ockenfels, A., 2000. ERC: A Theory of Equity, Reciprocity, and Competition. American Economic Review. 90(1), 166-193.

7. Bolton, G., Brandts, J. and Ockenfels, A., 2005. Fair Procedures: Evidence from Games Involving Lotteries. Economic Journal. 115(506), 1054-1076.

8. Collie, T., Bradley, G. and Sparks, B. A., 2002. Fair process revisited: Differential effects of interactional and procedural justice in the presence of social comparison information. Journal of Experimental Social Psychology. 38, 545-555.

9. Dufwenberg, M. and Kirchsteiger, G., 2004. A Theory of Sequential Reciprocity. Games and Economic Behavior. 47(2), 268-298.

10. De Cremer, D. and van Knippenberg, D., 2003. Cooperation with leaders in social dilemmas: On the effects of procedural fairness and outcome favorability in structural cooperation. Organizational Behavior and Human Decision Processes. 91, 1-11.

11. Falk, A. and Fischbacher, U., 2006. A Theory of Reciprocity. Games and Economic Behavior. 54(2), 293-315.

12. Fehr, E. and Schmidt, K., 1999. A Theory Of Fairness, Competition, And Cooperation. The Quarterly Journal of Economics. 114(3), 817-868.

13. Geanakoplos, J., Pearce, D. and Stacchetti, E., 1989. Psychological Games and Sequential Rationality. Games and Economic Behavior. 1, 60-79.

14. Heider, F., 1958. The Psychology of Interpersonal Relations. New York: John Wiley & Sons.

15. Kelley, H. H., 1967. Attribution in social psychology. Nebraska Symposium on Motivation, 15, 192-238.

16. Kelley, H. H., 1973. The processes of causal attribution. American Psychologist. 28, 107-128.

17. Konovsky, M. A. (2000), Understanding Procedural Justice and Its Impact on Business Organizations, Journal of Management, 26(3), 489-511.

18. Krawczyk, M., 2007. A model of procedural and distributive fairness. Mimeo. University of Amsterdam.

19. Lemons, M. A. and Jones, C.A. 2001. Procedural justice in promotion decisions: using perceptions of fairness to build employee commitment. Journal of Managerial Psychology. 16(4), 268-281.

20. Lind, E. A. and Tyler, T. R., 1988. The social psychology of procedural justice. New York, NY, US: Plenum Press.

21. Lind, E. A., Greenberg, J., Scott, K. S. and Welchans, T. D., 2000. The winding road from employee to complainant: Situational and psychological determinants of wrongful termination claims. Administrative Science Quarterly, 45, 557-590.

22. Murphy, K., 2004. The role of trust in nurturing compliance: A study of accused tax avoiders. Law and Human Behavior. 28, 187-209.

23. Nalebuff, B. and Shubik, M., 1988. Revenge And Rational Play. Papers 138. Princeton. Woodrow Wilson School - Public and International Affairs.

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

25. Roberts, K. and Markel, K. S., 2001. Claiming in the name of fairness: Organizational justice and the decision to file for workplace injury compensation. Journal of Occupational Health Psychology. 6, 332-347.

26. Ross, M. and Fletcher, G. J. O., 1985. Attribution and Social Perception. in G. Lindsey & E. Aronson (eds.). The Handbook of Social Psychology. 2, 73-114.

27. Thibaut J. and Walker, L., 1975. Procedural Justice. Erlbaum. Hillsdale, NJ.

28. Trautmann, S. T., 2006. A Fehr-Schmidt Model for Process Fairness. Working Paper. CREED. University of Amsterdam.

29. Tyler, T. R., 1990. Why people obey the law. New Haven, CT, US: Yale University Press.

30. Tyler, T. R., 2003. Procedural justice, legitimacy, and the effective rule of law. in Tonry, M. (ed.), Crime and justice: A review of research. 30, 283-358.

31. Wenzel, M., 2002. The impact of outcome orientation and justice concerns on tax compliance: The role of taxpayers' identity. Journal of Applied Psychology. 87, 629-645.

Appendix

Proof to result (7):

In this proof we show under what conditions the behavior as defined in Result (7) is the equilibrium behavior. Note, as defined in Definition (6) we assume that players' beliefs are correct. Given this, we analyze under what conditions they can be sustained in equilibrium. It can easily be seen that if player 2's second order belief about player 1's belief is $(c(h_2^4), c(h_2^5), c(h_2^6), d(h_2^7))$, then player 2's believed equitable payoff is $\pi_2^{e_1} = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2}$. Hence, player 2's perceived kindness if player 1 chooses $c(h_1^0)$ is

$$\lambda_{212}\left(h_2^4\right) = 1 - \frac{1}{2} = \frac{1}{2},$$

where 1 is player 2's expected monetary payoff and $\frac{1}{2}$ his equitable payoff given his second order belief.

Secondly, if player 1 plays $r(h_1^0)$ player 2's perceived kindness of player 1's *procedural choice* is

$$\begin{aligned} \lambda_{212}\left(h_2^5\right) &= \lambda_{212}\left(h_2^6\right) \\ &= \alpha_2(1) + (1-\alpha_2)(-1) - \frac{1}{2} \\ &= 2\alpha_2 - \frac{3}{2}, \end{aligned} \tag{4}$$

and, thirdly, if player 1 plays $d(h_1^0)$, it is

$$\lambda_{212}\left(h_2^7\right) = 0 - \frac{1}{2} = -\frac{1}{2}.$$

From equation (4) it can directly be seen that player 2 perceives player 1's *procedural choice* $r(h_1^0)$ as kind or unkind depending on α_2. If $\alpha_2 > \frac{3}{4}$ then player 1's choice of $r(h_1^0)$ is perceived as kind. Therefore,

Remark 9 *If α_2 is such that*

$$1 > \alpha_2 > \frac{3}{4},$$

then player 2 perceives player 1's procedural choice $r(h_1^0)$ as kind.

Henceforth we assume that player 1's *procedure* $r(h_1^0)$ is associated with an explicit probability distribution $\alpha_2 > \frac{3}{4}$.

We said before that player 1's first order belief is $(c(h_2^4), c(h_2^5), c(h_2^6), d(h_2^7))$. Furthermore, we said that in equilibrium this belief has to be correct. Hence, under what condition do we expect player 2 to choose $c(h_2^4)$ following player 1's choice of $c(h_1^0)$? By playing $c(h_2^4)$ player 2 receives the following utility

$$u_2\left(c\left(h_2^4\right)\right) = 1 + Y_2(1)\left(\frac{1}{2}\right),$$

where $\kappa_{21}\left(c\left(h_2^4\right)\right) = 1 - \frac{1}{2}\left((1) + (-1)\right) = 1$ is player 2's kindness to player 1 by playing $c\left(h_2^4\right)$. On the other hand, by playing $d\left(h_2^4\right)$ player 2's utility is

$$u_2\left(d\left(h_2^4\right)\right) = 2 + Y_2\left(-1\right)\left(\frac{1}{2}\right),$$

where $\kappa_{21}\left(d\left(h_2^4\right)\right) = -1 - \frac{1}{2}\left((1) + (-1)\right) = -1$. Hence player 2 plays $c\left(h_2^4\right)$ in history h_2^4 if

$$1 + Y_2\left(1\right)\left(\frac{1}{2}\right) \geq 2 + Y_2\left(-1\right)\left(\frac{1}{2}\right).$$

This reduces to

$$Y_2 \geq 1.$$

This shows that if player 1 plays $c\left(h_1^0\right)$, player 2 plays $c\left(h_2^4\right)$ if $Y_2 \geq 1$.

Remark 10 *If player 1 plays $c\left(h_1^0\right)$, player 2 plays $c\left(h_2^4\right)$ if $Y_2 \geq 1$.*

Going back to player 1's first order belief, under what conditions do we expect player 2 to choose $d\left(h_2^7\right)$ following player 1's choice of $d\left(h_1^0\right)$?

In history h_2^7 it is easy to see that player 2's monetary and reciprocity payoff induce him to chooses $d\left(h_2^7\right)$ for all $Y_2 \geq 0$. Hence, if player 1 plays $d\left(h_1^0\right)$, player 2 plays $d\left(h_2^7\right)$ if $Y_2 \geq 0$.

Remark 11 *If player 1 plays $d\left(h_1^0\right)$, player 2 plays $d\left(h_2^4\right)$ if $Y_2 \geq 0$.*

Finally, under what conditions do we expect player 2 to choose $c\left(h_2^5\right)$ in h_2^5 and $c\left(h_2^6\right)$ in h_2^6 following player 1's choice of $r\left(h_1^0\right)$?

Assume that player 1 has chosen $r\left(h_1^0\right)$. Doing the analogous calculations as above for player 2's behavior in history h_2^5 one can see that player 2 plays $c\left(h_2^5\right)$ in h_2^5 if

$$1 + Y_2\left(1\right)\left(2\alpha_2 - \frac{3}{2}\right) \geq 2 + Y_2\left(-1\right)\left(2\alpha_2 - \frac{3}{2}\right),$$

where the lhs is $u_2\left(c\left(h_2^5\right)\right)$ and the rhs is $u_2\left(d\left(h_2^5\right)\right)$. The above reduces to

$$Y_2 \geq \frac{1}{4\alpha_2 - 3}.$$

Note, as $\alpha_2 > \frac{3}{4}$ we know that $Y_2 \geq \frac{1}{4\alpha_2 - 3} > 1$. This shows that any player 2 with $Y_2 \geq \frac{1}{4\alpha_2 - 3}$ would play $c\left(h_2^4\right)$ in history h_2^4 and $c\left(h_2^5\right)$ in history h_2^5. Finally, in history h_2^6 the analogous calculations as in h_2^5 and h_2^4 are

$$-1 + Y_2\left(1\right)\left(2\alpha_2 - \frac{3}{2}\right) \geq 0 + Y_2\left(-1\right)\left(2\alpha_2 - \frac{3}{2}\right),$$

where the lhs is $u_2\left(c\left(h_2^6\right)\right)$ and the rhs is $u_2\left(d\left(h_2^6\right)\right)$. The above also reduces to

$$Y_2 \geq \frac{1}{4\alpha_2 - 3}.$$

Hence, also here it holds that if $Y_2 \geq \frac{1}{4\alpha_2 - 3}$ player 2 plays $c\left(h_2^6\right)$ in history h_2^6.

Remark 12 *If player 1 plays* $r\left(h_1^0\right)$*, player 2 plays* $c\left(h_2^5\right)$ *in* h_2^5 *and* $c\left(h_2^6\right)$ *in* h_2^6 *if* $Y_2 \geq \frac{1}{4\alpha_2-3}$.

Concluding, as we have seen above, if $Y_2 \geq \frac{1}{4\alpha_2-3}$, it holds that player 2' equilibrium behavior is characterized by $c\left(h_2^4\right), c\left(h_2^5\right), c\left(h_2^6\right)$ and $d\left(h_2^7\right)$ in histories h_2^4, h_2^5, h_2^6 and h_2^7 respectively.

Let us now turn to player 1. Player 1's perceived kindness of player 2's equilibrium *procedural strategy* is

$$\begin{aligned}\lambda_{121}\left(h_1^0\right) &= (q+2q'-q'\alpha_2)-(1-q'\alpha_2-q)\\ &= 2q+2q'-1,\end{aligned}$$

where q and q' are player 1's second order beliefs associated with his *procedures* $c\left(h_1^0\right)$ and $r\left(h_1^0\right)$. His kindness to player 2, on the other hand, is

$$\kappa_{12}\left(c\left(h_1^0\right)\right) = 1-\frac{1}{2} = \frac{1}{2},$$

by playing $c\left(h_1^0\right)$,

$$\begin{aligned}\kappa_{12}\left(r\left(h_1^0\right)\right) &= \alpha_2-(1-\alpha)-\frac{1}{2}\\ &= 2\alpha_2-\frac{3}{2},\end{aligned}$$

by playing $r\left(h_1^0\right)$ and

$$\kappa_{12}\left(c\left(h_1^0\right)\right) = 0-\frac{1}{2} = -\frac{1}{2},$$

by playing $d\left(h_1^0\right)$.

Putting the pieces together one can see that player 1 chooses $r\left(h_1^0\right)$ in equilibrium if for $q'=1$ and $q=0$ two conditions hold: *i)* $u_1\left(r\left(h_1^0\right)\right) \geq u_1\left(c\left(h_1^0\right)\right)$ and *ii)* $u_1\left(r\left(h_1^0\right)\right) \geq u_1\left(d\left(h_1^0\right)\right)$. The first condition boils down to

$$(2-\alpha_2)+Y_1\left(2\alpha_2-\frac{3}{2}\right) \geq (1)+Y_1\left(\frac{1}{2}\right),$$

which reduces to

$$Y_1 \leq \frac{1}{2}.$$

The second condition furthermore boils down to

$$(2-\alpha_2)+Y_1\left(2\alpha_2-\frac{3}{2}\right) \geq (0)+Y_1\left(-\frac{1}{2}\right),$$

which holds for all $Y_1 \geq 0$. Hence, given player 2's behavior, the equilibrium behavior of player 1 is characterized by $r\left(h_1^0\right)$ if $0 < Y_1 \leq \frac{1}{2}$.

Remark 13 *Given player 2's equilibrium behavior* $\left(c\left(h_2^4\right), c\left(h_2^5\right), c\left(h_2^6\right), d\left(h_2^7\right)\right)$*, player 1 plays* $r\left(h_1^0\right)$ *if* $0 < Y_1 \leq \frac{1}{2}$.

In other words, if player 2's sensitivity to reciprocity is high and player 1's is not too strong, the equilibrium behavior for both players is player 1 choosing the *procedure* $r\left(h_1^0\right)$ and player 2 choosing $\left(c\left(h_2^5\right), c\left(h_2^6\right)\right)$ in response. This concludes the proof of Result (7).■

Proof to result (8):

In analogy to the aforementioned proof, we first show under what conditions

$$\left(\omega_2'\left(h_2^4\right),\omega_2\left(h_2^5\right)\right)$$

and

$$\left(\omega_3'\left(h_3^6\right),\omega_2\left(h_3^7\right)\right)$$

simultaneously represent the equilibrium behavior of players 2 and 3. Then, secondly, we show the conditions for which it is best for player 1 to choose $\omega_1'\left(h_1^0\right)$, given the behavior of players 2 and 3.

If $\left(\omega_2'\left(h_2^4\right),\omega_2\left(h_2^5\right)\right)$ and $\left(\omega_3'\left(h_3^6\right),\omega_2\left(h_3^7\right)\right)$ are player 2's and 3's *procedural strategies*, then the most and least that player 1 can give to player 2 and 3 is either 1 or $-\varepsilon$. Hence, it can easily be seen that the perceived kindness of player 2 and 3 in either of the four histories h_2^4, h_2^5, h_3^6, h_2^7 is:

$$\begin{aligned}\lambda_{212}\left(h_2^4\right) &= \lambda_{313}\left(h_3^7\right)\\ &= -\varepsilon-\frac{1}{2}\left(1-\varepsilon\right)\\ &= -\frac{1}{2}\left(1+\varepsilon\right),\end{aligned}$$

$$\begin{aligned}\lambda_{212}\left(h_2^5\right) &= \lambda_{212}\left(h_3^6\right)=\lambda_{313}\left(h_2^5\right)=\lambda_{313}\left(h_3^6\right)\\ &= \frac{1}{2}-\frac{1}{2}\left(1-\varepsilon\right)\\ &= \frac{1}{2}\varepsilon,\end{aligned}$$

and

$$\begin{aligned}\lambda_{212}\left(h_3^7\right) &= \lambda_{313}\left(h_2^4\right)\\ &= 1-\frac{1}{2}\left(1-\varepsilon\right)\\ &= \frac{1}{2}\left(1+\varepsilon\right),\end{aligned}$$

where $\pi_2^{e_1}=\pi_3^{e_1}=\frac{1}{2}\left(1-\varepsilon\right)$. In other words, if player 1 chooses $\omega_1\left(h_1^0\right)$ player 2 perceives this as unkind and player 3 as kind. On the other hand, if player 1 chooses $\omega_1''\left(h_1^0\right)$, player 2 perceives this as kind and player 3 as unkind. Furthermore, if player 1 takes his decision by flipping a coin, i.e. $\omega_1'\left(h_1^0\right)$, then both players do not perceive this as unkind as $\varepsilon\geq 0$.

Remark 14 *Player 2 perceives player 1's procedural choice of $\omega_1\left(h_1^0\right)$ as unkind. Likewise, player 3 perceives player 1's procedural choice $\omega_1''\left(h_1^0\right)$ as unkind. On the other hand, both player do not perceive player 1's choice $\omega_1'\left(h_1^0\right)$ as unkind.*

Consider now all histories in turn. Looking at history h_2^4 after which player 2 has to choose one can see that player 2 can either show a kindness of

$$\begin{aligned}\kappa_{12}\left(\omega_2\left(h_2^4\right)\right) &= 3-\frac{1}{2}(3+1)\\ &= 1\end{aligned}$$

by playing $\omega_2\left(h_2^4\right)$ or he can show a kindness of

$$\begin{aligned}\kappa_{12}\left(\omega_2'\left(h_2^4\right)\right) &= 1-\frac{1}{2}(3+1)\\ &= -1\end{aligned}$$

by playing $\omega_2'\left(h_2^4\right)$. Obviously, player 2's behavior in history h_2^4 in general also creates some (un)kindness towards player 3. In our case, however, 3's monetary payoff is invariant to player 2's choice in h_2^4. Hence, player 2's kindness towards player 3 is 0 in h_2^4. Given this, the utilities from either of player 2's choices are

$$u_2\left(\omega_2\left(h_2^4\right)\right) = (0)+Y_{21}(1)\left(-\frac{1}{2}(1+\varepsilon)\right),$$

and

$$u_2\left(\omega_2'\left(h_2^4\right)\right) = (-\varepsilon)+Y_{21}(-1)\left(-\frac{1}{2}(1+\varepsilon)\right).$$

Again in equilibrium player 2 chooses the latter if $u_2\left(\omega_2'\left(h_2^4\right)\right) \geq u_2\left(\omega_2\left(h_2^4\right)\right)$. This can be written as

$$(-\varepsilon)+Y_{21}(-1)\left(-\frac{1}{2}(1+\varepsilon)\right) \geq (0)+Y_{21}(1)\left(-\frac{1}{2}(1+\varepsilon)\right),$$

which reduces to

$$Y_{21} \geq \frac{\varepsilon}{\varepsilon+1}.$$

This means if $Y_{21} \geq \frac{\varepsilon}{\varepsilon+1}$ then player 2 takes revenge on player 1 by choosing $\omega_2'\left(h_2^4\right)$ in history h_2^4. From the symmetry of the game it necessarily also follows that everything which holds for player 2 in history h_2^4 also holds for player 3 in history h_3^7. In other words if

$$Y_{31} \geq \frac{\varepsilon}{\varepsilon+1},$$

then player 3 takes revenge on player 1 in history h_3^7 by playing $\omega_3'\left(h_3^7\right)$.

Remark 15 *Players 2 and 3 take revenge on player 1 by playing $\omega_2'\left(h_2^4\right)$ in h_2^4 and $\omega_3'\left(h_3^7\right)$ in h_3^7 respectively, if $Y_{21}, Y_{31} \geq \frac{\varepsilon}{\varepsilon+1}$.*

Turning now to histories h_2^5 and h_2^6 one can see that due to the symmetry of the situation both players, 2 and 3, perceive player 1's kindness identically. Therefore, in history h_2^5 player 2's utilities from choosing either of his *procedures* is

$$u_2\left(\omega_2\left(h_2^5\right)\right) = (0)+Y_{21}(1)\left(\frac{1}{2}\varepsilon\right)$$

and

$$u_2\left(\omega_2'\left(h_2^5\right)\right) = (-\varepsilon) + Y_{21}(-1)\left(\frac{1}{2}\varepsilon\right).$$

He chooses $\omega_2(h_2^5)$ rather than $\omega_2'(h_2^5)$ if $u_2(\omega_2(h_2^5)) \geq u_2(\omega_2'(h_2^5))$, i.e.

$$(0) + Y_{21}(1)\left(\frac{1}{2}\varepsilon\right) \geq (-\varepsilon) + Y_{21}(-1)\left(\frac{1}{2}\varepsilon\right),$$

which reduces to

$$Y_{21} \geq -1.$$

Note, this holds for all $Y_{21} \geq 0$. Again, for equal reasons also player 3 chooses $\omega_3(h_3^6)$ rather than $\omega_2'(h_3^6)$ in history h_3^6 if $Y_{21} \geq 0$.

Remark 16 *If player 2's and 3's sensitivity to reciprocity is*

$$Y_{21} \geq 0,$$

and

$$Y_{31} \geq 0,$$

then they respectively choose $\omega_2(h_2^5)$ and $\omega_3(h_3^6)$ in histories h_2^5 and h_3^6 following player 1's choice of $\omega_1'(h_1^0)$.

Concluding, if $Y_{21} \geq \frac{\varepsilon}{\varepsilon+1}$ and $Y_{31} \geq \frac{\varepsilon}{\varepsilon+1}$ then players 2 and 3 play $(\omega_2'(h_2^4), \omega_2(h_2^5))$ and $(\omega_3'(h_3^6), \omega_2(h_3^7))$ in their histories h_2^4, h_2^5 and h_3^6, h_3^7 respectively.

Given this under what conditions is it best for player 1 to choose $\omega_1'(h_1^0)$? Assume for simplicity that player 1's sensitivity to reciprocity is equal towards both, player 2 and 3. In other words, assume that $Y_{12} = Y_{13} = Y$. Denote player 1's second order beliefs about player 2's and 3's beliefs p_2, p_2' and $(1 - p_2 - p_2')$ as well as p_3, p_3' and $(1 - p_3 - p_3')$. More precisely, let p_i and p_i' be player 1's belief about the probabilities that any player $i \in \{2, 3\}$ attaches to player 1's *procedures* $\omega_1(h_1^0)$ and $\omega_1'(h_1^0)$ respectively. Therefore, player 1's perceived kindness from player 2's and 3's procedural strategies is

$$\begin{aligned}\lambda_{121} &= p_2(-1) + p_2'\left(\frac{1}{2}(1) + \frac{1}{2}(0)\right) + (1 - p_2 - p_2')(0) \\ &= p_2'\left(\frac{1}{2}\right) - p_2,\end{aligned}$$

and

$$\begin{aligned}\lambda_{131} &= p_3(0) + p_3'\left(\frac{1}{2}(0) + \frac{1}{2}(1)\right) + (1 - p_3 - p_3')(-1) \\ &= p_3'\left(\frac{1}{2}\right) - (1 - p_3 - p_3')\end{aligned}$$

Player 1's kindness, on the other hand, towards player 2 and 3 is given by $\kappa_{12}(\omega_1(h_1^0)) = \kappa_{13}(\omega_1''(h_1^0)) = -\frac{1}{2}(1 + \varepsilon)$, $\kappa_{13}(\omega_1(h_1^0)) = \kappa_{12}(\omega_1''(h_1^0)) = \frac{1}{2}(1 + \varepsilon)$ and $\kappa_{12}(\omega_1'(h_1^0)) = \kappa_{13}(\omega_1'(h_1^0)) = \frac{1}{2}\varepsilon$.

Hence, given that players 2 and 3 choose $(\omega_2'(h_2^4), \omega_2(h_2^5))$ and $(\omega_3'(h_3^6), \omega_2(h_3^7))$ the utilities from all of player 1's *procedural choices* can be written as

$$\begin{aligned} u_1(\omega_1(h_1^0)) &= 1 + Y\left(-\frac{1}{2}(1+\varepsilon)\right)\left(p_2'\left(\frac{1}{2}\right) - p_2\right) \\ &\quad + Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(p_3'\left(\frac{1}{2}\right) - (1 - p_3 - p_3')\right) \end{aligned}$$

by playing $\omega_1(h_1^0)$,

$$\begin{aligned} u_1(\omega_1'(h_1^0)) &= 3 + Y\left(\frac{1}{2}\varepsilon\right)\left(p_2'\left(\frac{1}{2}\right) - p_2\right) \\ &\quad + Y\left(\frac{1}{2}\varepsilon\right)\left(p_3'\left(\frac{1}{2}\right) - (1 - p_3 - p_3')\right) \end{aligned}$$

by playing $\omega_1'(h_1^0)$ and

$$\begin{aligned} u_1(\omega_1''(h_1^0)) &= 1 + Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(p_2'\left(\frac{1}{2}\right) - p_2\right) \\ &\quad + Y\left(-\frac{1}{2}(1+\varepsilon)\right)\left(p_3'\left(\frac{1}{2}\right) - (1 - p_3 - p_3')\right) \end{aligned}$$

by playing $\omega_1''(h_1^0)$.

Obviously, player 1 plays $\omega_1'(h_1^0)$ if $u_1(\omega_1'(h_1^0)) \geq u_1(\omega_1(h_1^0))$ and $u_1(\omega_1'(h_1^0)) \geq u_1(\omega_1''(h_1^0))$ with $p_2 = p_3 = 0$, $p_2' = p_3' = 1$ and $p_2'' = p_3'' = 0$. The first of the two conditions can be written as

$$\begin{aligned} & 3 + Y\left(\frac{1}{2}\varepsilon\right)\left(\frac{1}{2}\right) + Y\left(\frac{1}{2}\varepsilon\right)\left(\frac{1}{2}\right) \\ \geq\ & 1 - Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(\frac{1}{2}\right) + Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(\frac{1}{2}\right), \end{aligned}$$

which holds for all $Y > 0$. Secondly, it has to hold that

$$\begin{aligned} & 3 + Y\left(\frac{1}{2}\varepsilon\right)\left(\frac{1}{2}\right) + Y\left(\frac{1}{2}\varepsilon\right)\left(\frac{1}{2}\right) \\ \geq\ & 1 + Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(\frac{1}{2}\right) - Y\left(\frac{1}{2}(1+\varepsilon)\right)\left(\frac{1}{2}\right), \end{aligned}$$

which is identical to the above. Hence, whenever $Y = Y_{12} = Y_{13} > 0$ it holds that player 1's best response to player 2's and 3's *procedural strategy* $(\omega_2'(h_2^4), \omega_2(h_2^5))$ and $(\omega_3'(h_3^6), \omega_2(h_3^7))$ is to play $\omega_1'(h_1^0)$.

Remark 17 *Given player 2's and 3's equilibrium play, player 1 chooses procedure* $\omega_1'(h_1^0)$*, if* $Y = Y_{12} = Y_{13} > 0$.

This concludes the proof of Result (8).■

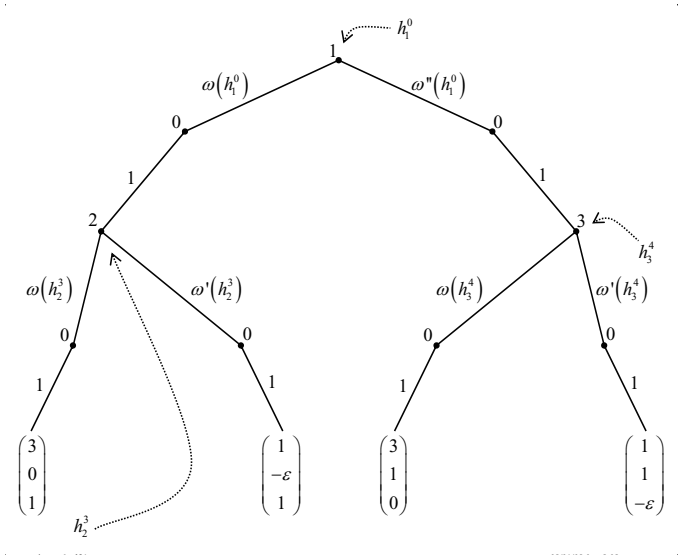

Figure 5: Game Γ_5

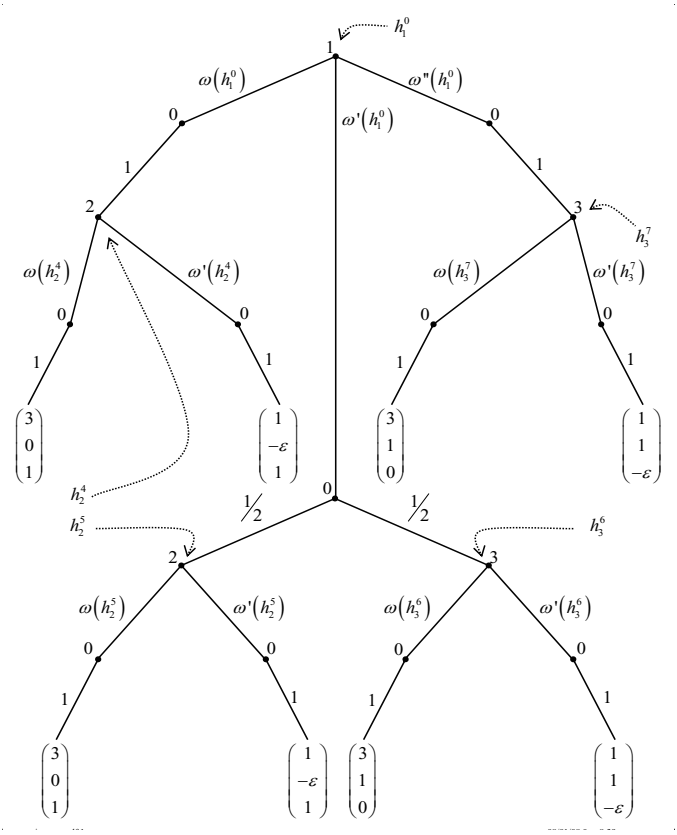

Figure 6: Game Γ_6

Procedural Concerns in Psychological Games

Procedural Concerns in Psychological Games*

Alexander Sebald†

Abstract

One persistent finding in experimental economics is that people react very differently to outcomewise identical situations depending on the procedures which have led to them. In accordance with this, there exists a broad consensus among psychologists that not only expected outcomes shape human behavior, but also the way in which decisions are taken. Economists, on the other hand, have remained remarkably silent about procedural aspects of strategic interactions. This paper provides a game theoretic framework that integrates procedural concerns into economic analysis. Building on Battigalli and Dufwenberg (2007)'s framework of dynamic psychological games, we show how procedural concerns can be conceptualized assuming that agents are (also) motivated by belief-dependent psychological payoffs. Procedural choices influence the causal attribution of responsibilities, the evaluation of intentions and the arousal of emotions. Two applications highlight the impact and importance of procedural concerns in strategic interactions.

Keywords: Psychological Game Theory, Procedural Concerns, Reciprocity, Guilt Aversion
JEL Classification: D01, C70

1 Introduction

One persistent finding in experimental economics is that people react very differently to outcomewise identical situations depending on the *procedures* which have led to them [e.g. Blount (1995), Falk et al. (2000), Charness (2004), Brandts et al. (2006), Charness and Levine (2007)]. For example, Charness and Levine (2007) experimentally analyze workers' reactions to pay decisions by firms following different wage-setting procedures.[1] They find that the process leading to a specific

*I am very grateful to Estelle Cantillon, Paolo Casini, Gary Charness, Paola Conconi, Werner Güth and Georg Kirchsteiger for helpful comments.

†Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands, and ECARES, Université Libre de Bruxelles. Sebald is also member of ECORE, the recently created association between CORE and ECARES. E-mail: a.sebald@algec.unimaas.nl

[1]In their experiment firms have to choose between either a low ($4) or a high ($8) wage. Following the firm's decision, this wage is either decreased or increased by $2 depending on a stochastically determined (i.e. flip of a coin) economic condition which can either be good or bad.

wage affects the workers' effort choice. Given the same wage, workers choose significantly more often low effort when the wage-setting process reveals less-good intentions by firms compared to situations in which intentions are perceived as good. In the same spirit, Brandts et al. (2006) show that selection procedures matter in a three-player game in which one player has to select one of the other players to perform a specific task.[2] In their experiment selected players behave very differently in their subsequent tasks depending on the type of procedure which was used to select them. They suggest that people exhibit *procedural concerns* because selection procedures affect the beliefs that people hold about each others' intentions and expectations which subsequently influence their behaviors.

This paper provides a game theoretic framework that integrates procedural concerns into economic analysis. Building on Battigalli and Dufwenberg (2007)'s *dynamic psychological games*, we show how procedural concerns can be conceptualized assuming that agents are (also) motivated by belief-dependent psychological payoffs. Our paper consist of three building blocks: a class of *procedural games* in which agents choose for procedures rather than for actions as traditionally assumed in game theory, agents with belief-dependent utilities as defined by Battigalli and Dufwenberg (2007) and a solution concept, sequential psychological equilibrium. Using these three building blocks we show how procedural choices influence the causal attribution of responsibilities, the evaluation of intentions and the arousal of emotions.

Among psychologists there exists by now a broad consensus that not only expected outcomes shape human behavior, but also the way in which decisions are taken [e.g. Thibaut and Walker (1975), Lind and Tyler (1988), Collie et al. (2002), Anderson and Otto (2003) and Blader and Tyler (2003)]. Prominent examples of areas in which procedures have been found to play an eminent role are workplace relations and the public acceptability of policies and laws. Psychologists have found evidence that behavioral reactions to promotion decisions, bonus allocations, dismissals etc. strongly depend on the perceived fairness of selection procedures [e.g. Lemons and Jones (2001), Konovsky (2000), Bies and Tyler (1993), Lind et al. (2000) and Roberts and Markel (2001)] and that public compliance with policies and laws strongly depends on the perceived fairness of their enforcement procedures [e.g. Tyler (1990), Wenzel (2002), Murphy (2004), De Cremer and van Knippenberg (2003) and Tyler (2003)].

Psychologists explain the impact of procedures on human interactions with the help of *attribution theory* [e.g. Heider (1958), Kelley (1967), Kelley (1973), Ross and Fletcher (1985)]. Attribution theory assumes that people need to infer causes and assign responsibilities for why outcomes occur. It is argued that especially

After the revelation of the economic condition, the workers have to choose their effort level: low, medium or high. The flip of the coin introduces the possibility to compare to different intentional states that represent two different ways through which the same wage, i.e. \$6, is determined: *i*) good intentions: high wage coupled with bad economic condition and *ii*) less-good intentions: less costly low wage coupled with good economic condition.

[2]Two different treatments are studied which differ with regard to the selection procedure. In both treatments the task of the selected player is to choose between two different payoff allocations determining the payoff of all three players.

when outcomes are unfavorable and perceptions of intentions are strong, there is a tendency to assign responsibility for outcomes to people. The assignment of responsibility and blame in turn has been shown to affect the occurrence and intensity of emotions like disappointment, guilt, anger and aggression [Blount (1995)]. To exemplify, imagine a workplace situation in which a principal wants to promote one out of two agents. If he chooses to take the decision on who is to be promoted intransparently, e.g. behind closed doors, agents are driven to attach a high degree of responsibility for the outcome to the principal. His choice is interpreted as intentional, which fosters perceptions of favoritism. If, by contrast, the principal uses a transparent procedure which credibly shows that the decision is based on an unbiased criterion, i.e. a criterion which a priori ensures that both agents have the same chance to be promoted, the principal is not blamed for the final outcome. Hence, if the agents care about intentions their reaction to the same promotion decision will differ depending on the promotion procedure used by the principal. Hence, according to the psychological literature procedures influence the responsibility that people have for specific outcomes, they mitigate the evaluation of intentions and subsequent behaviors.

Notwithstanding the experimental and psychological evidence and the fact that e.g. workplace relations also play an eminent role in the economic literature, traditional economic theory has remained remarkably silent about the impact of procedures on human behavior. Only three recent economic papers have started to theoretically address the issue of procedural concerns [Bolton et al. (2005), Trautmann (2006), Krawczyk (2007)]. In contrast to the psychologists' view, however, they all extend models of distributional preferences to account for the impact of procedural choices on strategic interactions. Bolton et al. (2005) and Krawczyk (2007)'s models are based on the theory of inequity aversion by Bolton and Ockenfels (2000). Trautmann, on the other hand, builds on Fehr and Schmidt (1999)'s model of social concerns. All three take a similar approach suggesting that the experimental evidence on procedural concerns can be accounted for when agents' utilities depend on expected outcome differences ex ante as well as ex post to any outcome realization.

As indicated in the beginning, Brandts et al. (2006) and Charness and Levine (2007) follow the psychologists' view. They argue that intention-based models, e.g. models of reciprocity and guilt aversion, rather than distributional preferences, explain the experimental evidence on procedural concerns.

Economic theory has widely neglected emotions and intentions as these issues are difficult to reconcile with the traditional presumption of stable consequentialist preferences. Spurred by experimental findings, economists have only recently started to look at the impact of belief-dependent motivations on strategic interactions. Departing from the strictly consequentialist tradition in economics e.g. Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2007) have developed a framework to analyze the strategic interaction of agents with belief-dependent motivations: psychological game theory. Roughly speaking, psychological games are games in which agents are (also) motivated by belief-dependent psychological payoffs capturing their emotional involvement. Emotions depend on beliefs

about intentions [Elster (1998)]. Geanakoplos et al. (1989) concentrate on games in which only agents' initial beliefs matter, whereas Battigalli and Dufwenberg (2007) develop a dynamic framework in which agents update their beliefs about their own and the others' intentions as games unfold. Many types of emotions (e.g. regret, disappointment, guilt, reciprocity) have already been formalized in the context of psychological games. Rabin (1993), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006), for example, analyze the strategic interaction of agents that act reciprocally. Battigalli and Dufwenberg (2006) look at the interaction of agents that feel guilt, i.e. that are guilt averse.

Although all of these models are able to explain observed behaviors in experiments in contradiction to classical assumptions about human conduct [e.g. Charness and Dufwenberg (2006), Charness and Rabin (2002), Fehr and Gaechter (2000)], none of them explores the role of procedural choices in the interaction of *emotional agents*.[3] Therefore, different from the existing literature on psychological games, in this paper we concentrate on the impact of procedural choices on the interaction of agents with belief-dependent motivations. First, we show that procedural concerns can theoretically be conceptualized assuming that agents are (also) incentivized by belief-dependent psychological payoffs. As procedural choices affect the beliefs that people hold and agents utilities are belief-dependent, emotional agents exhibit procedural concerns. Second, we show that the behavioral predictions of the already existing literature on psychological games are sensitive to the availability of different procedures to take the same decision. In the existing literature on psychological games it is implicitly assumed that people can only use procedures that make them fully responsible for the outcomes of their actions. In our procedural games people can choose between different procedures to take the same decision. As will be seen, this leads to different equilibrium predictions compared to the existing literature on psychological games. In another paper [Sebald (2007)] it was already shown how procedural concerns affect the strategic interaction of reciprocal agents. Sebald (2007), thus, is an application of the general framework presented here.

Our work is related to the (experimental) literature on the impact of *institutions* on human interaction [North (1991), Bowles (1998), Bohnet (2006), Bohnet (2007)]. In this literature institutions are commonly defined as humanly devised *rules of the game* that structure political, economic and social interactions. The argument is that institutions create and direct incentives, affect preferences, provide information on processes leading to certain outcomes and allow people to make inferences about others' motivations [Bohnet (2006)]. Bohnet and Zeckhauser (2004) and Hong and Bohnet (2005), for example, experimentally investigate the effect of causal attribution in different institutional settings. They analyze the first-mover behavior in two closely related but different institutional environments, a binary-choice trust and a binary-choice risky dictator game.[4] Participants act differently

[3]Following Elster (1998), throughout the paper we will sometimes refer to agents with belief-dependent psychological payoffs as emotional agents.

[4]In both situations the first-mover has to decide between either an outside option or to let a second-mover decide between two alternative payoff allocations. In the binary-choice trust game the second-mover is another player. In the binary-choice risky dictator game, on the other hand,

in the two settings suggesting that people dislike being betrayed by others more than losing a lottery. This implies that there is an additional psychological factor influencing the strategic interaction related to the causal attribution of responsibilities [Bohnet (2006)]. In line with this, our methodological approach sheds light on the hidden relation connecting the information on procedures entailed in institutions and the process of causal attribution. Our work suggests that the process information entailed in institutions creates the possibility for causal attribution and directs it in such a way that people are only held accountable for what they are actually responsible.

The organization of the paper is as follows: In the next section we formally define procedures and characterize a class of *procedural games* in which agents choose procedures rather than actions and strategies. In section 3 we study the impact of procedures on the behavior of emotional agents. More precisely, we characterize agents with belief-dependent psychological payoffs in the context of our class of procedural games and provide a first example of the impact of procedural choices on their strategic interactions. In section 4 we develop the concept *sequential psychological equilibrium* for our procedural games with psychological incentives. Finally, we discuss two applications that highlight the impact and importance of procedural concerns in strategic interactions of reciprocal and guilt averse agents.

2 Procedures and Procedural Games

In this section we proceed in two steps. First, we intuitively sketch our methodological approach with the help of two examples. In a second step we formally define the concept of procedures and fully characterize our class of procedural games in which agents do not choose actions and strategies, as usually assumed in game theory, but procedures. This class of multi-stage games is used in the subsequent sections to capture and analyze the impact of procedural choices on the strategic interaction of agents with belief-dependent utilities.

As a starting point consider games Γ_1 and Γ_2 in Figure 1 and 2:

[Figure 1 and 2 here]

The sole difference between games Γ_1 and Γ_2 is that in Γ_2 player 1 can choose (M) on top of his pure actions (L) and (R). Player 1's pure action (M), however, is nothing else than choosing an explicit randomization device, (0), assigning probabilities α_2 and $(1-\alpha_2)$ to his pure actions (L) and (R) respectively. *Flipping a coin* constitutes such an explicit randomization device, for example. It assigns the probability $\frac{1}{2}$ to both pure actions (L) and (R). Obviously, flipping a coin is just one way in which a decision can be taken. In reality, one usually disposes of many different *credible ways*. Consider, for example, the workplace situation sketched in the introduction. The principal could take the promotion decisions

the second-mover is *chance* reducing the role of the second player to being a dummy. It is found that first-movers act differently if the responder is the other player compared to the situation in which *chance* acts as the second-mover.

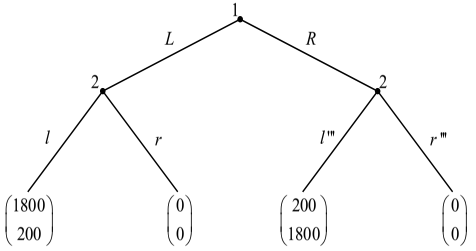

Figure 1: Game Γ_1

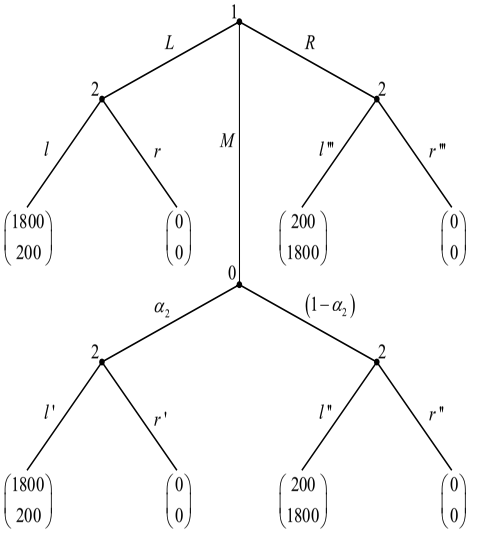

Figure 2: Game Γ_2

by organizing a promotion tournament using an objective evaluation criterion. Given that both agents are identical, i.e. are equally skilled, and this is commonly known, this would induce a commonly known probability distribution over the set of pure actions giving both agents an equal *chance* to be promoted. Note that we distinguish between explicit, i.e. credible, randomizations which are observed by all players and implicit randomizations, i.e. behavioral strategies. The choice (M) of player 1 in game Γ_2 is a pure choice for an explicit randomization device and it differs from player 1 choosing a behavioral strategy in game Γ_1 which implicitly randomizes over his pure actions (L) and (R) without the others observing the random draw.

But not only choices like (M) can be formalized as choices for explicit randomization devices. Taking the thought about the *credible ways* to the extreme shows that also pure actions like (L) and (R) can equally be defined as choices for explicit randomization mechanisms. Imagine, for example, that player 1 in Γ_1 and Γ_2 chooses his pure action (L). This is equivalent to saying that player 1 chooses for *chance* to take the decision between (L) and (R) assigning probability 1 to his pure action (L). Hence, although (L) represents a pure action, it can nevertheless be reinterpreted in a way in which the decision is indirectly taken by *chance* randomizing with a degenerated probability distribution over the set $\{(L), (R)\}$. This shows that in our two examples, Γ_1 and Γ_2, any choice for a pure action, i.e. (L) and (R), and any choice for an explicit randomization mechanism, i.e.(M), can likewise be reinterpreted as a choice for an explicit randomization device through which the actual decision is subsequently taken by *chance*.

Consider, for example, game Γ_3 in Figure 3, which is a restatement of game Γ_2 in the spirit of this intuition:[5]

[Figure 3 here]

As one can see, in Γ_3 we reformulate all strategic choices of game Γ_2 into choices for explicit randomization mechanisms through which decisions are subsequently taken. In game Γ_2 player 1 can decide between (L), (M) and (R). Equivalently, in game Γ_3 he has to decide between the explicit randomization devices ω_{1,h^0}, ω'_{1,h^0} and ω''_{1,h^0} in the initial history h^0. First, by choosing ω_{1,h^0} he decides to let *chance* take the decision between (L) and (R) assigning probability 1 to (L), i.e. $\rho(L) = 1$. Second, by choosing ω'_{1,h^0} he decides to let *chance* take the decision between (L) and (R) assigning probability α_2 to (L), i.e. $\rho(L) = \alpha_2$, and $(1 - \alpha_2)$ to (R), i.e. $\rho(R) = (1 - \alpha_2)$. Finally, by choosing ω''_{1,h^0} he decides to let *chance* take the decision between (L) and (R) assigning probability 1 to (R), i.e. $\rho(R) = 1$. In all these three cases player 1 only determines how *chance* subsequently takes the decision, rather then taking the decision himself.

Hence, despite the equivalence between games Γ_2 and Γ_3, an interpretive difference exists. Choosing for an explicit randomization mechanism implies that players do not take decisions themselves. They merely determine how decisions are taken by *chance*. In other words, players decide about the procedures which

[5]Note, actions that are played by player *chance* with probability 0 are disregarded in Figure 3.

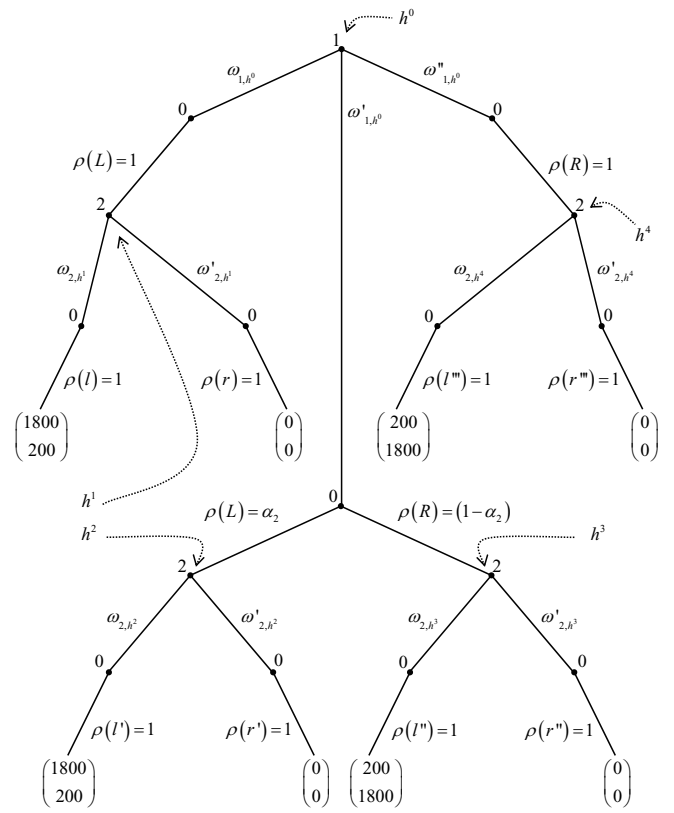

Figure 3: Game Γ_3

are used to take decisions. The example in Figure 3, thus, uncovers that strategic decision making is not only about choosing actions but also about *how* actions are chosen. For this reason we call game Γ_3 a *procedural game*.

This brings us to the formal definition of our class of procedural games. Formally, let the set of players be $\mathcal{N} = \{0, 1, ..., N\}$ where 0 denotes the uninterested player *chance*. Denote as $\mathcal{H}$ the finite set of histories h, with the empty sequence $h^0 \in \mathcal{H}$, and Z the set of end nodes. Histories $h \in \mathcal{H}$ are sequences that describe the choices that players have made on the path to history h. We assume that only one player moves after each non-terminal history. Hence, the set of histories $\mathcal{H} \backslash \{Z\}$ can be partitioned into sets $\mathcal{H}_i$, with $i \in \mathcal{N}$. At each non-terminal history $h \in \mathcal{H}_i$ after which player $i \in \mathcal{N} \backslash \{0\}$ has to move he disposes of a finite set of feasible actions denoted by $\mathcal{A}_i(h)$ and a finite set of explicit randomization devices denoted by $\Omega_i(h)$ through which he can indirectly choose an action from $\mathcal{A}_i(h)$. In fact, as already suggested in example Γ_3, in our procedural games players $i \in \mathcal{N} \backslash \{0\}$ do not choose actions $a_{i,h} \in \mathcal{A}_i(h)$ directly, but choose explicit randomization mechanisms, denoted by $\omega_{i,h} \in \Omega_i(h)$, through which a decision is indirectly taken by *chance*. The choice for a specific explicit randomization device $\omega_{i,h}$ in history h by player $i \in \mathcal{N} \backslash \{0\}$ determines the explicit probability distribution $\rho_{0,h'}$ with which *chance* takes the actual decision in the following history $h' = (h, \omega_{i,h})$. Hence, any history h controlled by a player $i \in \mathcal{N} \backslash \{0\}$ is succeeded by a history h' controlled by player 0. More formally, if player $i \neq 0$ chooses $\omega_{i,h}$ in history h with length x, then player 0 takes a decision $a_{0,h'}$ in history $h' = (h, \omega_{i,h})$ of length $x + 1$ explicitly randomizing with the probability distribution $\rho_{0,h'}$ over the set $\mathcal{A}_0(h') = \mathcal{A}_i(h)$.[6]

To exemplify, the choice for a pure action (e.g. (L) in Γ_2) translates in our procedural game into a choice for an explicit randomization mechanisms $\omega_{i,h}$ that is associated with a degenerated probability distribution $\rho_{0,h'}$ which assigns probability 1 to the pure action $a_{i,h}$ in the set of possible actions $\mathcal{A}_0(h') = \mathcal{A}_i(h)$. The choice for an explicit randomization device like e.g. (M) in Γ_2), on the other hand, is a choice for an explicit randomization mechanism, $\omega'_{i,h}$, that is associated with a non-degenerate probability distribution $\rho'_{0,h'}$ defined on $\mathcal{A}_0(h') = \mathcal{A}_i(h)$.

This means, player *chance* essentially plays a commonly known, i.e. explicit, mixed strategy $\rho_0 = (\rho_{0,h})_{h \in \mathcal{H}_0}$ which specifies for each history $h \in \mathcal{H}_0$ that he controls a behavioral strategy $\rho_{0,h}$ according to which an action $a_{0,h}$ is chosen from $\mathcal{A}_0(h)$. Consequently, one can denote as $\rho_0(s_0|h)$ the probability with which player 0 plays the pure strategy $s_0 = (a_{0,h})_{h \in \mathcal{H}_0}$ conditional on history h.

Intuitively, as players only decide on *how* decisions are taken, they only decide on the procedures, which are used to take them. This brings us to a formal definition of procedures:

Definition 1 *A procedure, $\omega_{i,h} \in \Omega_i(h)$, for player $i \in \mathcal{N} \backslash \{0\}$ in history $h \in \mathcal{H}_i$*

[6]Note that the length of a history corresponds to the number of choices that are contained in that history.

is a tuple:[7]

$$\langle \rho_{0,h'}, \mathcal{A}_0(h') \rangle,$$

where $h' = (h, \omega_{i,h})$ *and* $\rho_{0,h'}$ *is an explicit probability distribution defined on* $\mathcal{A}_0(h') = \mathcal{A}_i(h)$.

For a given set of procedures $\Omega_i(h)$, the associated set of explicit probability distributions is denoted by $\mathcal{P}_i(h) = \{\rho_{0,h'} \mid \omega_{i,h} \in \Omega_i(h)\}$. The minimum number of procedures that a player $i \in \mathcal{N} \setminus \{0\}$ can decide between in any history h controlled by him equals the number of pure actions that he has in the traditional extensive form representation.

We define a *procedural strategy* for player $i \in \mathcal{N} \setminus \{0\}$ as a collection that specifies a procedure for each history $h \in \mathcal{H}_i$ after which player i moves, $\omega_i = (\omega_{i,h})_{h \in \mathcal{H}_i}$, where $\omega_{i,h}$ is the procedure that would be selected by player i if h occurred. It is assumed that all players learn the outcome of a procedure directly after its realization and perfect recall holds.

Let $\Omega_i = \times_{h \in \mathcal{H}_i} \Omega_i(h)$ and $\Omega = \times_{\mathcal{N} \setminus \{0\}} \Omega_i$. Given a procedural strategy, $\omega_i \in \Omega_i$ for each player $i \in \mathcal{N} \setminus \{0\}$ and the commonly known system of probability distributions, $\mathcal{P} = \cup_{i \in \mathcal{N} \setminus \{0\}} \mathcal{P}_i$, where $\mathcal{P}_i = \cup_{h \in \mathcal{H}} \mathcal{P}_i(h)$, we can compute a probability distribution over end nodes. By assigning payoffs to end nodes, we can derive an expected payoff function, $\pi_i : Z \to \Re$, for every player $i \in \mathcal{N} \setminus \{0\}$ which depends on what *procedural profile*, $\omega \in \Omega$ is played. In what follows, we assume that payoffs are material payoffs like money or any other measurable quantity of some good.

Summarizing:

Definition 2 *A procedural game is a tuple:*

$$\Gamma = \left\langle \mathcal{N}, \Omega, (\pi_i : Z \to \Re)_{\mathcal{N} \setminus \{0\}} \right\rangle.$$

This concludes the definition of procedures and the characterization of our class of procedural games which is the basis for our subsequent analysis. Starting from two simple examples, i.e. Γ_1 and Γ_2, we have formalized the idea that players choose for procedures rather than actions. In this way we have separated choices for procedures and actual decisions. In the remainder of the paper we use this class of procedural games in order to isolate the impact of procedural choices on the strategic behavior of agents with belief-dependent utilities.

[7] In example Γ_3 procedures are used to choose pure actions. We do not exclude, however, the possibility that players use procedures to choose between procedures and procedures that choose between procedures that choose between procedures etc. Procedures, $\omega_{i,h} \in \Omega_i(h)$, rather have to be understood as *reduced prædures*. The explicit probability distribution associated with a reduced procedure, $\rho_{0,h'} \in \mathcal{P}_i(h)$, basically subsumes the probability distributions of procedures of all levels into one explicit distribution indirectly defined on $\mathcal{A}_i(h)$.

3 Procedural Games with Psychological Incentives

It is easy to see that if agents are only interested in their own expected material payoff, the set of all subgame perfect equilibria of two identical subgames is the same. Looking again at game Γ_3 in Figure 3, for example, this means that players are expected to react the same in histories h^1 and h^2. However, as already mentioned in the introduction, there exists ample evidence contradicting this traditional behavioral presumption. People very often react differently in outcomewise identical situations depending on the procedures which have led to them. Following the psychologist's view procedural choices affect peoples' beliefs about intentions. Hence, if people are (also) motivated by belief-dependent psychological payoffs, they exhibit procedural concerns. To conceptualize this idea, in this section we define procedural games in which agents have belief-dependent psychological incentives. This will allow us to formally capture the impact of procedural choices on the strategic behavior of emotional agents.

Economists have only recently developed a framework, i.e. psychological game theory, to formally account for behavioral traits such as emotions and intentions [e.g. Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2007)]. Psychological games are games in which agents are (also) motivated by belief-dependent psychological payoffs capturing their emotional involvement. Psychological payoffs arise from the beliefs that agents have about their opponents' strategies and beliefs. Therefore let agents have:

i) beliefs about the strategies of other players,

ii) beliefs about the beliefs of other players,

and

iii) let them update their beliefs as events unfold.

In order to formally capture assumptions *i*)-*iii*), we have to define an epistemic structure (*collectively coherent hierarchies of beliefs*) which describes what people initially belief and how they update their beliefs as play unfolds. This epistemic structure can be characterized in the context of our procedural games by assuming that players hold hierarchies of conditional beliefs over the procedural strategies as well as beliefs of other players $i \in \mathcal{N} \setminus \{0\}$.

As in Battigalli and Dufwenberg (2007) we only summarize the theory of hierarchies of conditional beliefs.[8] We describe, first, a system of conditional first-order-beliefs and then, secondly, show how this extends to higher orders (i.e. second-order beliefs etc). In our class of procedural games denote by Ω_{-i} the set of procedural strategies of the opponents j where $j \in \mathcal{N} \setminus \{0, i\}$. At the beginning of any game, i.e. in the initial history h^0, player i does not know the true procedural strategies of his opponents. He only learns the true strategy $\omega_{-i} \in \Omega_{-i}$

[8] For topological details, proofs and further references see Brandenburger and Dekel (1993) and Battigalli and Siniscalchi (1999).

step-by-step by updating his beliefs as the game unfolds. More formally, player i assigns probabilities to the events in the Borrel sigma algebra $\mathcal{B}$ of Ω_{-i} according to some probability measure. Let $\Delta(\Omega_{-i})$ be the set of all such probability measures. Denote $\mathcal{C} \subseteq \mathcal{B}$ the set of potential conditioning events at which player i can update his beliefs. In other words, $\mathcal{C}$ is the set of potentially observable events. Player i holds probabilistic beliefs about his opponents's procedural strategies conditional on each event $F \in \mathcal{C}$. These probabilistic beliefs are captured in a conditional probability system (*cps*).

From Battigalli and Dufwenberg (2007) consider the following definition:

Definition 3 *A conditional probability system (cps) is a function* $\mu(\cdot|\cdot) : \mathcal{B} \times \mathcal{X} \to [0,1]$ *defined on* $(X, \mathcal{B}, \mathcal{C})$ *such that for all* $E \in \mathcal{B}$ *and* $F', F \in \mathcal{C}$*:*

1. $\mu(\cdot|\cdot) \in \Delta(X)$,
2. $\mu(F|F) = 1$,
3. $E \subseteq F' \subseteq F$ *implies* $\mu(E|F) = \mu(E|F')\,\mu(F'|F)$,

where X *is a set, e.g.* Ω_{-i}*, whose 'true' element* $x \in X$ *is initially unknown and only learned step-by-step as conditioning events, e.g.* $F \in \mathcal{C}$*, are reached.*

Concentrating, first, on beliefs of order 1 means $X = \Omega_{-i}$. The first two conditions of definition 3 ensure that $\mu(\cdot|F)$ is indeed a probability measure (i.e. $\mu(\cdot|F) \in \Delta(X)$) which puts all probability weight on F given that F is observed. Condition 3 ensures that players update their beliefs according to Bayes' rule. The set of all functions μ for which conditions *1-3* hold is denoted by $\Delta^H(X)$. Hence, $\Delta^H(\Omega_{-i})$ is the set of all conditional probability systems of order 1 of player i.

Definition 3 can easily be extended to higher-order-beliefs. In the construction of the first-order *cps* we start from an initial situation in which player i does not know the true procedural strategy of his opponents. He has a conditional first-order-belief over it which is updated as play unfolds. Analog to this, in the construction of a second-order-belief we start from an initial situation in which player i does not know the true procedural strategy and the true conditional first-order-belief of players $-i$. Hence, the relevant set X in definition 3 becomes:

$$X = \Omega_{-i} \times \prod_{j \neq i} \Delta^H(\Omega_{-j}),$$

where $i, j \in \mathcal{N} \setminus \{0\}$ and $\Delta^H(\Omega_{-j})$ is the set of conditional first-order *cps* of player j. The resulting conditional probability system does not only represent player i's belief about the strategy of players $-i$, but also about their first-order-beliefs.

Generalizing this idea, first- and higher-order *cps* are defined recursively as follows. Let:

$$\begin{aligned} X^0_{-i} &= \Omega_{-i}, \text{ where } i \in \mathcal{N} \setminus \{0\}, \\ X^k_{-i} &= X^{k-1}_{-i} \times \prod_{j \neq i} \Delta^H\left(X^{k-1}_{-j}\right), \text{ where } i \in \mathcal{N} \setminus \{0\} \text{ and } k = 1, 2, \ldots . \end{aligned}$$

Then, a *cps* $\mu_i^k \in \Delta^H\left(X_{-i}^{k-1}\right)$ is called a k-order *cps* or simply a k-order belief. For $k > 1$, μ_i^k is a joint *cps* on the opponents' strategies and $(k-1)$-order *cps*', i.e.:

$$\begin{aligned}
\mu_i^1 &\in \Delta^H\left(X_{-i}^0\right) \text{ where } X_{-i}^0 = \Omega_{-i}, \\
\mu_i^2 &\in \Delta^H\left(X_{-i}^1\right) \text{ where } X_{-i}^1 = \Omega_{-i} \times \Delta^H\left(\Omega_{-j}\right), \\
\mu_i^3 &\in \Delta^H\left(X_{-i}^2\right) \text{ where } X_{-i}^2 = \Omega_{-i} \times \Delta^H\left(\Omega_{-j}\right) \times \Delta^H\left(\Omega_{-j} \times \Delta^H\left(\Omega_{-i}\right)\right) \text{ etc. .}
\end{aligned}$$

This brings us to the formal definition of hierarchies of *cps*':[9]

Definition 4 *A hierarchy of cps is a countably infinite sequence of cps':*

$$\mu_i = \left(\mu_i^1, \mu_i^2, ...\right) \in \prod_{k>0} \Delta^H\left(X_{-i}^{k-1}\right).$$

As one can see, each piece of information appears many times in the belief hierarchy of player i. This implies that one can calculate marginal beliefs of higher-order-beliefs. As also Geanakoplos et al. (1989) point out, these marginal beliefs of higher-order-beliefs should coincide with lower-order-beliefs in the belief hierarchy for the hierarchy to be meaningful. In other words beliefs should be *coherent*. We say a hierarchy of *cps*' is coherent if the *cps*' of distinct orders assign the same conditional probabilities to lower-order-events. This means,

$$\mu_i^k\left(\cdot|h\right) = marg_{X_{-i}^{k-1}} \mu_i^{k+1}\left(\cdot|h\right) (k = 1, 2, ...; h \in \mathcal{H}),$$

where $marg_{X_{-i}^{k-1}} \mu_i^{k+1}\left(\cdot|h\right)$ is the event of order $k-1$ in the *cps* of order $k+1$, $\mu_i^{k+1}\left(\cdot|h\right)$. If this condition holds, player i is said to have a coherent conditional belief system. It can be shown that a coherent hierarchy of *cps'* induces a single *cps* ν_i on the cross product of Ω_{-i} and the sets of hierarchies of *cps*' of i's opponents $-i$. Note, however, coherency regarding the own beliefs does not exclude the possibility that the *cps* ν_i puts a positive probability on the opponents *incoherence*. But as players are rational they should not believe that their opponents entertain incoherent beliefs. Hence, in order to rule this out, say that a coherent hierarchy μ_i satisfies belief in coherency of order 1 if the induced *cps* ν_i is such that each $\nu_i\left(\cdot|h\right)$ with $h \in \mathcal{H}$ assigns probability one to the opponents' coherence of order 1. The hierarchy of coherent beliefs μ_i satisfies belief in coherency of order k, if it satisfies belief in coherency of order $k-1$, μ_i is *collectively coherent*, if it satisfies belief in coherency of order k for each positive integer k.[10] We denote the set of *collectively coherent hierarchies of beliefs* of player i by M_i. The set of collectively coherent beliefs of the opponents $-i$ is M_{-i} and $M = \prod_{j \in \mathcal{N} \setminus \{0\}} M_j$.

Finally, as the probability distributions associated with the moves of the player *chance*, i.e. player 0, are commonly known, nobody faces any uncertainty with regard to his *true type*. In other words, players do not learn the true strategy of player 0 over the course of the game, as it is ex ante commonly known. As

[9] See also Battigalli and Dufwenberg (2007), p. 13.
[10] See also Battigalli and Dufwenberg (2007), p.13.

will be seen in the Example presented below, this is crucial in the context of our procedural games. As the mixed strategy, ρ_0, of the player *chance* is commonly known, causal attribution is linked to procedural choices and not to outcomes.

To come to full circle, belief-dependent utilities are utilities that are not only defined on monetary outcomes, but also on collectively coherent hierarchies of beliefs and the commonly known probability distributions associated with the moves of the player *chance*:

Definition 5 *A belief dependent utility u of player i is a function:*

$$u_i : \mathcal{Z} \times \mathcal{P} \times \mathcal{M}_i \times \Pi_{j \neq i} (\Omega_j \times \mathcal{M}_j) \rightarrow \Re.$$

As mentioned in the introduction, strategic interactions with belief-dependent utilities have so far only been analyzed in traditional dynamic decision contexts, i.e. traditional extensive form representations, (e.g. Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Battigalli and Dufwenberg (2006)). Definition 5 represents an adaptation of these earlier approaches to our class of procedural games in which agents choose procedures rather than actions and strategies. In order to get a first impression of the impact of procedural choices on the interaction of emotional agents consider the following example:

Example: Assume that players 1 and 2 in game Γ_3 are reciprocal. This means they react kindly (unkindly) if they perceive the other to be kind (unkind). As we only want to give a first glimpse of the importance of procedures, we concentrate in this example on the perception that player 2 has about the kindness of player 1 in the histories h^1 and h^2. As said before, histories h^1 and h^2 are starting points of identical subgames.

Following Dufwenberg and Kirchsteiger (2004) the perceived kindness of player 2 in h^1 and h^2 can be defined as:

$$\lambda_{212} = \pi_2 \left(\mu_2^1 (\cdot|h^x), \mu_2^2 (\cdot|h^x), \rho_0 \right) - \pi_2^{e_1} \left(\mu_2^1 (\cdot|h^x), \mu_2^2 (\cdot|h^x), \rho_0 \right),$$

where $x \in \{1, 2\}$ and

$$\pi_2^{e_1} (\cdot) = \frac{1}{2} \Big[\max \left\{ \pi_2 \left(\mu_2^1 (\cdot|h^x), \mu_2^2 (\cdot|h^x), \rho_0 \right), \omega_1 \in \Omega_1 \right\} + \min \left\{ \pi_2 \left(\mu_2^1 (\cdot|h^x), \mu_2^2 (\cdot|h^x), \rho_0 \right), \omega_1 \in \Omega_1 \right\} \Big].$$

The perceived kindness λ_{212} is defined as the difference between what player 2 believes player 1 intends to give him, $\pi_2 (\cdot)$ (conditional on history h^x and given player 2's first- and second-order beliefs, μ_2^1 and μ_2^2, and the mixed strategy of the player *chance*, ρ_0) and an equitable payoff, $\pi_2^{e_1}$. Dufwenberg and Kirchsteiger (2004) define the equitable payoff, $\pi_2^{e_1}$, as the average of the minimum and the maximum that player 2 believes player 1 could give him (again conditional on history h^x and given player 2's first- and second-order beliefs, μ_2^1 and μ_2^2, and the mixed strategy of the player *chance*, ρ_0).

Assume, for example, that $\alpha_2 = (1 - \alpha_2) = \frac{1}{2}$ and imagine that player 2 believes that player 1 believes that he plays left in all the histories that he controls, i.e.

histories h^1, h^2, h^3 and h^4. Given this, player 2 has to believe that player 1 intended to give him a material payoff of

$$\pi_2\left(\mu_2^1\left(\cdot|h^1\right),\mu_2^2\left(\cdot|h^1\right),\rho_0\right)=200,$$

if he finds himself in history h^1 after player 1 has chosen procedure ω_{1,h^0}. In contrast to this, if player 2 finds himself in history h^2 he has to believe that player 1 intended to give him:

$$\pi_2\left(\mu_2^1\left(\cdot|h^2\right),\mu_2^2\left(\cdot|h^2\right),\rho_0\right)=\frac{1}{2}\left(200\right)+\frac{1}{2}\left(1800\right)=1000,$$

by choosing procedure ω'_{1,h^0}. The equitable payoff, on the other hand, is given by:

$$\pi_2^{e_1}\left(\cdot\right)=\frac{1}{2}\Big[200+1800\Big],$$

where 200 is the minimum that player 2 believes player 1 could have given him in history h^0 (by playing ω_{1,h^0}) and 1800 is the maximum (by playing ω''_{1,h^0}). Putting the pieces together, player 2's perceived kindness in history h^1 and h^2 are respectively:

$$\lambda_{212}(h^1)=200-\frac{1}{2}\Big[200+1800\Big]=-800,$$
$$\lambda_{212}(h^2)=1000-\frac{1}{2}\Big[200+1800\Big]=0.$$

Hence, although histories h^1 and h^2 are starting points of identical subgames, they are perceived very differently by player 2 due to the different procedural choices which have led to them. It is now easy to see that player 2 who is concerned about the intentions of player 1 might react differently in histories h^1 and h^2 depending on the strength of his reciprocal preferences. This gives a first idea of how procedural choices influence the causal attribution of responsibilities and the strategic interaction of emotional agents. ■

Given our class of procedural games as defined in the previous section and the belief-dependent utilities (Definition 5), we are now ready to define procedural games with psychological incentives:

Definition 6 *A procedural game with psychological incentives is a tuple:*

$$\Gamma_P=\left\langle\Gamma,\left(u_i\right)_{\mathcal{N}\setminus\{0\}}\right\rangle \text{ where } u_i:\mathcal{Z}\times\mathcal{P}\times\mathcal{M}_i\times\Pi_{j\neq i}\left(\Omega_j\times\mathcal{M}_j\right)\rightarrow\Re.$$

Procedural games with psychological incentives are the framework which we use to capture the impact of procedural choices on the interaction of psychologically motivated agents. Before presenting some applications, however, we subsequently adapt Battigalli and Dufwenberg (2007)'s *sequential equilibrium* to our class of procedural games with psychological incentives.

4 Sequential Psychological Equilibria in Procedural Games with Psychological Incentives

Battigalli and Dufwenberg (2007) adapt Kreps and Wilson (1982)'s concept of sequential equilibrium to their class of dynamic psychological games. They do so by characterizing *consistent assessments* that do not only consist of first-, but also of higher-order beliefs and defining sequential equilibria as sequentially rational consistent assessments.

As in Battigalli and Dufwenberg (2007), also our equilibrium concept refers to mixed strategies, i.e. implicit randomizations over sets of procedures. Note, however, that, following Aumann and Brandenburger (1995), we interpret player i's mixed strategy as a conjecture on the part of his opponents as to what player i will do. Hence, denote a *behavioral procedural strategy* of player i as $\sigma_i = (\sigma_{i,h})_{h \in \mathcal{H}_i} \in \Sigma_i$, where Σ_i is the set of all mixed strategies of player i. The behavioral choice $\sigma_{i,h} \in \Sigma_i(h)$ in h has to be understood as an implicit randomization over the set of procedures $\Omega_i(h)$ in history h and interpreted as an array of common conditional first-order-beliefs held by i's opponents.[11] This means that the behavioral procedural strategy σ_i is part of an assessment $((\sigma, \rho_0), (\mu, \rho_0)) = ((\sigma_i, \rho_0), (\mu_i, \rho_0))_{i \in \mathcal{N} \setminus \{0\}}$ of behavioral strategies and hierarchies of conditional beliefs.

Three conditions ensure consistency of assessments in the original characterization by Kreps and Wilson (1982):

1. Beliefs must be derived using Bayes' rule,
2. Beliefs must reflect that players choose their strategies independently,
3. Players with identical information have identical beliefs.

In addition to these conditions, Battigalli and Dufwenberg (2007) add another requirement for consistency:

4. Players hold correct beliefs about each others' beliefs.

Condition 1 holds by the definition of hierarchies of conditional belief systems (Definition 3). In other words, hierarchies of beliefs are defined in such a way that conditional beliefs are consistent with Bayes' rule. In order to formalize conditions 2-4 we first need to define what is meant by *stochastic independence*. Note, the observability of past actions allows us to define stochastic independence of the conditional belief systems in terms of marginal *cps*'. Different to the concept of marginal beliefs used in the previous section, a marginal *cps* now refers to player i's marginal belief on the procedural strategies of a particular player j and it is denoted as $\mu^1_{ij} \in \Delta^H(\Omega_j)$, where $\Delta^H(\Omega_j)$ is the set of marginal *cps* on the procedural strategies of player j. Given this we can define stochastic independence of beliefs as:[12]

[11] See Battigalli and Dufwenberg (2007), p 16.

[12] See also Battigalli and Dufwenberg (2007)'s definition of stochastic independence, p 17, and their definition of sequential equilibrium, p 19.

Definition 7 *A first-order cps* $\mu_i^1 \in \Delta^H(\Omega_{-i})$ *satisfies stochastic independence, if there exists a profile of marginal cps'* $(\mu_{ij}^1)_{j\neq i} \in \prod_{j\neq i} \Delta^H(\Omega_j)$ *such that* $\mu_i^1(\omega_{-i}|h) = \prod_{j\neq i} \mu_{ij}^1(\omega_j|h)$ *for all* $h \in \mathcal{H}_i$. *We denote the set of stochastically independent first-order cps' of a player* i *as* $\Delta_I^H(\Omega_{-i})$.

This brings us to our definition of consistent assessments:

Definition 8 *An assessment* $((\sigma, \rho_0), (\mu, \rho_0))$ *is consistent if:*

1. *The first-order cps of each player satisfies stochastic independence as formalized in Definition (7), i.e.:*

$$\forall i \in \mathcal{N}\setminus\{0\}, \mu_i^1 \in \Delta_I^H(\Omega_{-i}).$$

2. *The marginal first-order cps of any two players about any third player coincide, i.e.:*

$$\forall i \in \mathcal{N}\setminus\{0\}, \forall l \in \mathcal{N}\setminus\{i,j,0\}, \forall h \in \mathcal{H}, \mu_{il}^1(\cdot|h) = \mu_{jl}^1(\cdot|h).$$

3. *Each players higher order beliefs in* μ *assign probability 1 to the lower order beliefs in* μ *itself:*

$$\forall i \in \mathcal{N}\setminus\{0\}, \forall k > 1, \forall h \in \mathcal{H}, \mu_i^k(\cdot|h) = \mu_i^{k-1}(\cdot|h) \times \delta_{\mu_{-i}^{k-1}},$$

where $\delta_{\mu_{-i}^{k-1}}$ *is the probability measure which assigns probability 1 to* μ_{-i}^{k-1}.

Conditions 1 and 2 capture the assumption that beliefs should be the end-product of a transparent reasoning process of intelligent people [Battigalli and Dufwenberg (2007)]. Condition 3, on the other hand, is analog to Geanakoplos et al. (1989)'s condition requiring that players hold common and correct beliefs about each others' beliefs.

After having defined consistent assessments we can formally characterize *sequential psychological equilibria* (henceforth: SPE) by requiring sequential rationality:

Definition 9 *An assessment* $((\sigma, \rho_0), (\mu, \rho_0))$ *is a sequential psychological equilibrium (SPE), if for all* $i \in \mathcal{N}\setminus\{0\}, h \in \mathcal{H}_i$ *it holds:*

$$Supp(\sigma_{i,h}) \subseteq argmax_{\omega_{i,h}\in\Omega(h)} E_{\mu,\rho_0}[u_i|h, \omega_{i,h}],$$

where $E_{\mu,\rho_0}[u_i|h, \omega_{i,h}]$ *is the expected utility of player* i *conditional on history* h, *procedural choice* $\omega_{i,h} \in \Omega(h)$ *and given the system of hierarchies of conditional beliefs* μ *and the commonly known mixed strategy,* ρ_0, *played by player* 0.

Note, the expected utility of any player $i \in \mathcal{N}\setminus\{0\}$ (conditional on history h, *procedural choice* $\omega_{i,h} \in \Omega(h)$, given the system of consistent hierarchies of conditional beliefs μ and the commonly known mixed strategy, ρ_0) can be defined

as:

$$E_{\mu,\rho_0}\left[u_i|h,\omega_{i,h}\right] = \sum_{s_0\in S_0(h)} \rho_0\left(s_0|h\right) \sum_{\omega_{-i}\in\Omega_{-i}(h)} \mu_i^1\left(\omega_{-i}|h\right) \sum_{\omega_i\in\Omega_i\left(h,\omega_{i,h}\right)} \mu_{ji}^1\left(\omega_i|\left(h,\omega_{i,h},\omega_{-i,h}\right)\right) u_i\left(\zeta\left(\omega_i,\omega_{-i},s_0\right),\rho_0,\mu,\omega_{-i}\right),$$

where $\zeta\left(\omega_i,\omega_{-i},s_0\right)\in Z$ denotes the terminal history induced by the procedural strategies ω_i and ω_{-i}, and the strategy s_0 of player 0. Note, this specification is different from the expected utility formula traditionally used. Furthermore, it is also different from the specification used by Battigalli and Dufwenberg (2007) as it encloses the behavioral moves of the player *chance*.

The following proposition shows that there exists at least one sequential psychological equilibrium in any procedural game with psychological incentives and continuous utility functions:

Proposition 1 *If the utility functions are continuous, there exists at least one sequential psychological equilibrium assessment.*

Proof: Consider a procedural game with psychological incentives in which any procedure at any history is played with a strictly positive minimal probability ε. More formally, consider an ε-perturbed game Γ^ε in which players $i\in\mathcal{N}\backslash\{0\}$ dispose of 'constrained' choice sets $\Sigma_i^\varepsilon(h)$ at each history $h\in\mathcal{H}_i$. The 'constrained' choice set $\Sigma_i^\varepsilon(h)$ of player i in history h is defined as:

$$\Sigma_i^\varepsilon(h) := \{\tau_{i,h}\in\Sigma_i(h)|\tau_{i,h}(\omega_{i,h})\geq\varepsilon,\,\forall\omega_{i,h}\in\Omega_i(h)\}.$$

So $\Sigma_i^\varepsilon(h)$ consists of only those elements in $\Sigma_i(h)$ that put a strictly positive probability greater or equal to ε on all elements $\omega_{i,h}\in\Omega_i(h)$, i.e. $\Sigma_i^\varepsilon(h)\subset\Sigma_i(h)$. It follows that in any Γ^ε the set of strictly mixed procedural strategies of players $i\in\mathcal{N}\backslash\{0\}$ is $\Sigma_i^\varepsilon = \times_{h\in\mathcal{H}_i}\Sigma_i{}^\varepsilon(h)$ and the set of all strictly positive behavioral procedural strategy profiles is $\Sigma^\varepsilon := \times_{i\in\mathcal{N}\backslash\{0\}}\Sigma_i^\varepsilon$. Note, for each $\sigma\in\Sigma^\varepsilon$ there exists a unique corresponding profile of hierarchies of *cps*' $\mu=\beta(\sigma)$ such that $((\sigma,\rho_0),(\beta(\sigma),\rho_0))$ is consistent.

Now, define for $\sigma\in\Sigma^\varepsilon$, $\varepsilon>0$, $i\in\mathcal{N}\backslash\{0\}$ and $h\in\mathcal{H}_i$ the local best-response of player i in history h as:

$$BR_{i,h}^\varepsilon(\sigma) := \{\hat{\tau}_{i,h}\in\Sigma_i^\varepsilon(h)|u_i(\sigma_i/\hat{\tau}_{i,h},\sigma_{-i},\rho_0)\geq u_i(\sigma_i/\tau_{i,h},\sigma_{-i},\rho_0),\,\forall\tau_{i,h}\in\Sigma_i^\varepsilon(h)\},$$

where $\sigma_i/\tau_{i,h}$ denotes the behavioral procedural strategy for player i that specifies the strictly positive mixture $\tau_{i,h}$ at history $h\in\mathcal{H}_i$ and σ_i at every other history controlled by player i. In other words, local best-response-correspondences are strictly mixed behavioral choices that put at least a minimum probability ε on each procedure $\omega_{i,h}\in\Omega_i(h)$ given i's choices in all other histories controlled by him and given the behavioral procedural strategy of the opponents. The domain of the local best-response-correspondence is Σ^ε. The set $\Sigma^\varepsilon=\Sigma_1^\varepsilon\times\Sigma_2^\varepsilon...\times\Sigma_N^\varepsilon$ and each Σ_i^ε with $i\in\mathcal{N}\backslash\{0\}$ is defined as $\Sigma_i^\varepsilon=\times_{h\in\mathcal{H}}\Sigma_i^\varepsilon(h)$. As said above,

$\Sigma_i^\varepsilon(h)$ is the set of all behavioral procedural strategies of player i at history h that put at least a strictly positive probability ε on each procedure $\omega_{i,h} \in \Omega_i(h)$. It is non-empty (because $\Omega_i(h)$ is non-empty), compact and convex. Hence, also Σ^ε is non-empty, compact and convex (because the finite Cartesian product of nonempty, convex and compact sets is itself nonempty, convex and compact). Furthermore, $BR_{i,h}^\varepsilon(\sigma)$ is upper-semi-continuous. Note, the local best-response-correspondence $BR_{i,h}^\varepsilon(\sigma)$ is upper-semi-continuous, if for any sequence $(\sigma_i/\hat{\tau}_{i,h}^m, \sigma_{-i}^m) \to (\sigma_i/\hat{\tau}_{i,h}, \sigma_{-i})$ such that $\sigma_i/\hat{\tau}_{i,h}^m \in BR_{i,h}^\varepsilon(\sigma_i/\hat{\tau}_{i,h}^m, \sigma_{-i}^m)$ for all $m \in \{1, 2, ...\}$, we have $\sigma_i/\hat{\tau}_{i,h} \in BR_{i,h}^\varepsilon(\sigma_i/\hat{\tau}_{i,h}, \sigma_{-i})$. To see that this is indeed the case, note that for all m, the $u(\sigma_i/\hat{\tau}_{i,h}^m, \sigma_{-i}^m) \geq u(\sigma_i/\hat{\tau}_{i,h}', \sigma_{-i}^m)$ for all $\sigma_i/\hat{\tau}_{i,h}' \in \Sigma_i^\varepsilon$. Hence, by the continuity of the utility function, we have $u(\sigma_i/\hat{\tau}_{i,h}, \sigma_{-i}) \geq u(\sigma_i/\hat{\tau}_{i,h}', \sigma_{-i})$.

Given the local best-response correspondence $BR_{i,h}^\varepsilon(\sigma)$, the best-response correspondence $BR^\varepsilon(\sigma)$ is defined as:

$$BR^\varepsilon = (\hat{\tau}_{i,h})_{h \in \mathcal{H}_i \wedge i \in \mathcal{N}\{0\}}.$$

This implies that also $BR^\varepsilon : \Sigma^\varepsilon \to \Sigma^\varepsilon$ is upper semi continues, compact and convex and, hence, has a fixed point $\hat{\sigma}^\varepsilon$. As already pointed out by Geanakoplos et al. (1989), the profile $\hat{\sigma}^\varepsilon$ constitutes an equilibrium of the constrained game Γ^ε.

Now, let ε^k be a sequence converging to 0 and $\hat{\sigma}^k$ the corresponding sequence of equilibrium assessments with $\hat{\sigma}^k$ being an equilibrium of Γ^{ε^k}. By the compactness of Σ, $\hat{\sigma}^k$ has an accumulation point σ^* and by the upper-semi-continuity of the local best-response-correspondents, $BR_h^\varepsilon(\sigma)$, $\sigma_{i,h}^*$ assigns positive probability only to those actions that are best responses to $(\sigma^*, \beta(\sigma^*), \rho_0)$ at h. Therefore $((\sigma^*, \rho_0), (\beta(\sigma^*), \rho_0))$ is a sequential equilibrium assessment. This concludes the proof. ■

Concluding, in this section we have formally defined sequential psychological equilibria in the context of our class of procedural games with psychological payoffs. Furthermore we have shown that every procedural game with psychological incentives with continues utility functions has at least one SPE. Using our solution concept we demonstrate in the following section the impact of procedural choices on the interaction of psychologically motivated agents by means of two examples.

5 Applications

In the first application we analyze *a principal-agent relation* in which agents behave reciprocally towards their principal. This application shows the impact that different promotion procedures have on the interaction of psychologically motivated agents. In the second application we analyze the *'So long, Sucker'* game which has also been discussed by Dufwenberg and Kirchsteiger (2004). Different to them, however, we do not assume reciprocal behavior but guilt aversion. A full description of the strategic interaction with all possible sequential psychological equilibria is beyond the scope of this paper. We therefore limit the analysis to the characterization of only one equilibrium per application to demonstrate the impact and importance of procedural choices in the interaction of agents with

belief-dependent utilities. Results and intuitions are presented in this section, lengthy mathematical proofs are relegated to the Appendix.

With the help of these two applications it is demonstrated i) how procedural choices influence the interaction of agents with belief-dependent psychological payoffs and ii) that the equilibrium predictions of the already existing literature using psychological games are sensitive to the availability of different procedures to take the same decision.

5.1 A Principal-Agent Relation

Imagine a principal, p, with two agents, $e1$ and $e2$, that is offered a project, b. He figures that in order to realize b he needs a *project manager*, pm, that is supported in the final phase of the project by an *assistant*, a, within the realm of his normal work. He knows that if both invest *high* effort, h, the project is a success, s, and he gets a payoff of $\pi(h,h) > 0$. If one of them invests *low* effort, l, however, the project will fail, f, and he will get a payoff of $\pi(h,l) = \pi(l,l) < \pi(h,h)$. Let both agents, $e1$ and $e2$, be equally skilled to perform either as *project manager* or *assistant*, implying that both have the same effort costs equal to v in case of *high* effort and 0 otherwise. Note that for simplicity we abstract in this principal-agent example from the usual question regarding the optimal incentive scheme. We take wages as given (e.g. due to a collective labor agreement) in order to single out the impact that the selection procedure has on the effort choices of the reciprocal agents. It is assumed that, in case of success, the principal pays $w(pm|s) > w(a)$ to the *project manager* and $w(a) < \frac{1}{2}((w(pm|s) - v) + (w(a) - v))$ to the *assistant*. On the other hand, in case of failure both get $w(pm|f) = w(a)$. Let efforts be observable, which implies that the *assistant* is aware of the *project managers*'s effort choice when choosing his own effort level, as he only collaborates in the final phase of the project. Furthermore, assume that the profits, $\pi(\cdot)$ minus the wage costs in case of a failure are 0 for the principal and positive if the project is a success.

Remark 1 *From the payment structure to the agents one can already see that, if effort is costly, the assistant has no monetary incentive to perform high effort since his wage will be $w(a)$ independent of the outcome of the project, b.*

The similarity of the two agents complicates the principal's decision on who is to become the *project manager* and who the *assistant*. Let the principal have two types of *procedures* that he can use to take his decision. He can either decide *behind closed doors*, *bcd*, or he can use a small *selection tournament*, *st*. This means his set of procedural strategies is $\Omega_p = \{st, bcd(e1), bcd(e2)\}$. For simplicity let the *selection tournament* be costless and credible to the agents. It is just about *concentration*, *c*, or *no concentration*, *nc*. Let it be commonly known that, if both *concentrate* or both do *not concentrate* during the short *selection tournament*, both are equally likely to become the *project manager*. If one *concentrates* and the other one does not, then the agent who *concentrates* gets the job.[13]

[13]This means that if both perform equally during the *tournament* the principal flips a coin in front of them.

From the outset it is clear that the principal's profit is maximized if both his agents invest *high* effort and he shares part of the profit with the *project manager*. Against the background of Remark (1) it is easy to see, however, that if agents are only concerned about their own monetary payoff, it is impossible for the principal to elicit *high* effort from both agents after his selection decision.

Result 1 *If both agents are only concerned about their own monetary payoff, it is impossible for the principal to elicit high effort, h, from agents e1 and e2 independent of the selection procedure. As a consequence, the 'selection tournament' can never be strictly preferred to a decision 'behind closed doors'.*

Proof: As said in Remark (1), the *assistant* never has a monetary incentive to perform *high* effort (as it is costly) independent of the selection procedure. Obviously this is also known to the *project manager* who conjectures that no matter what he does the project will fail and he will get $w(pm|f) = w(a)$ independent of the selection procedure which the principal has used to take his decision. Hence, his optimal choice is also to always perform *low* effort, l. Given this the principal is indifferent between his two different types of selection procedures. ■

Consider now a situation in which agents $e1$ and $e2$ behave reciprocally towards the principal p. As pointed out before, this means they reciprocate kind with kind and unkind with unkind behavior. This type of behavior can be captured by assuming that each agent $i \in \{e1, e2\}$ maximizes the following utility function:

$$u_i = \pi_i + Y_{ip} \cdot (\kappa_{ip} \cdot \lambda_{ipi}),$$

where $Y_{ip} > 0$ is a positive constant that captures agents i's sensitivity to reciprocity, κ_{ip} is agent i's belief about his kindness towards the principal, λ_{ipi} is the agent i's perceived kindness of the principal towards himself and π_i is agent i's own expected monetary payoff. Note, this conceptualization of reciprocity is analog to the definition by Dufwenberg and Kirchsteiger (2004).

In the Example in section 3 we have already defined perceived kindness. For completeness, however, let us restate it here in the context of our *principal-agent relation*. Agent i's perceived kindness of the principal, λ_{ipi}, in history h^x is defined as:

$$\lambda_{ipi} = \pi_i\left(\mu_i^1(\cdot|h^x), \mu_i^2(\cdot|h^x), \rho_0\right) - \pi_i^{e_p}\left(\mu_i^1(\cdot|h^x), \mu_i^2(\cdot|h^x), \rho_0\right).$$

As before, $\pi_i(\cdot)$ describes what agent i believes the principal intends to give him and $\pi_i^{e_p}(\cdot)$ is the equitable payoff which characterizes agent i's belief about the average that the principal could have given him. More formally:

$$\pi_i^{e_p}(\cdot) = \frac{1}{2}\Big[\max\left\{\pi_i\left(\mu_i^1(\cdot|h^x), \mu_i^2(\cdot|h^x), \rho_0\right), \omega_p \in \{st, bcd(e1), bcd(e2)\}\right\} + \min\left\{\pi_i\left(\mu_i^1(\cdot|h^x), \mu_i^2(\cdot|h^x), \rho_0\right), \omega_p \in \{st, bcd(e1), bcd(e2)\}\right\}\Big].$$

Similarly agent i's kindness towards the principal in history h^x can be described as:

$$\kappa_{ip} = \pi_p\left(\mu_i^1(\cdot|h^x), \omega_i, \rho_0\right) - \pi_p^{e_i}\left(\mu_i^1(\cdot|h^x), \omega_i, \rho_0\right),$$

where,

$$\pi_p^{e_i}(\cdot) = \frac{1}{2}\Big[\max\left\{\pi_i\left(\mu_i^1(\cdot|h^x), \omega_i, \rho_0\right), \omega_i \in \{l,h\}\right\} + \min\left\{\pi_i\left(\mu_i^1(\cdot|h^x), \omega_i, \rho_0\right), \omega_i \in \{l,h\}\right\}\Big].$$

In line with the above, the expected material payoff $\pi_p(\cdot)$ describes what agent i believes the principal will get, given his beliefs, the commonly known 'mixed strategy' of player *chance* and his own choice $\omega_i \in \{l,h\}$. Furthermore, $\pi_p^{e_i}(\cdot)$ is agent i's belief about the the average that he can give to the principal p.

In contrast to Result 1, the question arises whether the profit maximizing principal is also indifferent between his selection procedures, given that the agents behave reciprocally towards him. Note that the *principal-agent relation* is symmetric. This allows us to state the following result in terms of *project manager* and *assistant* rather than the behavior of agents $e1$ and $e2$ in their different possible roles.

Result 2 *If the project manager's sensitivity to reciprocity is:*

$$Y_{pmp} \geq \frac{(w(pm|f) - w(pm|s)) + \upsilon}{\frac{1}{2}\left[\frac{1}{2}\left[w(pm|s) - w(a)\right] - \upsilon\right]\left[\Delta\pi_p + \Delta w(pm)\right]},$$

and the assistant's sensitivity to reciprocity is:

$$Y_a \geq \frac{\upsilon}{\frac{1}{2}\left[\frac{1}{2}\left[w(pm|s) - w(a)\right] - \upsilon\right]\left[\Delta\pi_p + \Delta w(pm)\right]},$$

where $\Delta\pi_p = \pi_p(h,h) - \pi_p(h,l)$ *and* $\Delta w(pm) = w(pm|f) - w(pm|s)$*, then the sequential psychological equilibrium is given by:*

1. *The project manager i) chooses low effort following a decision of the principal taken behind closed doors, ii) chooses concentration and iii) high effort, if the principal uses a selection tournament to take his decision.*

2. *The assistant i) chooses low effort following a decision of the principal taken behind closed doors, ii) chooses concentration, iii) low effort following low effort by the project manager and the selection tournament and iv) high effort, if the principal uses a selection tournament to take his decision and the project manager has chosen high effort as well.*

3. *The principal uses the selection tournament.*

Proof: See appendix

The intuition behind this result is the following: The *assistant* feels unkindly treated, if the principal has taken the decision *behind closed doors.*[14] As effort is costly, he thus chooses *low* effort independent of the effort choice of the *project*

[14]This is also in analogy to the 'promotion' example briefly sketched in the introduction.

manager. In comparison to that, the *assistant* does feel kindly treated if the principal has used a *selection tournament* to choose the *project manager*. Thus, if he is sensitive enough to reciprocity, i.e. condition Y_a in Result (2) holds, then he chooses *high* effort given that the *project manager* has chosen *high* effort following *st*. But if, on the other hand, the *project manager* has chosen *low* effort following the *selection tournament*, the *assistant* knows that it is useless to invest *high* effort and, hence, he optimally chooses l. The *project manager* obviously knows all this. Hence, if he was selected *behind closed doors*, he chooses *low* effort because he knows that the *assistant* will. If he was selected via a *selection tournament*, however, and he knows that the assistant is sufficiently sensitive to reciprocity he will choose *high* effort to reciprocate the kind behavior of the principal. Given this the principal will choose the *selection tournament*, as in this way he maximizes his own profit. This highlights the importance of procedural choices in the interaction of psychologically motivated agents.

In addition, one can also confront Result (2) with the results obtained in the setting of Dufwenberg and Kirchsteiger (2004) who do not allow for different types of procedures. In order to do so, consider the same situation as described above, but without the principal's possibility to perform a *selection tournament*. In other words, the principal can only decide *behind closed doors*.

Result 3 *For all $Y_{ap} \geq 0$ and $Y_{pmp} \geq 0$ the SPE is given by:*

1. *The assistant always chooses low effort either out of pure cost (if $Y_{ap} = 0$) or cost and kindness considerations (if $Y_{ap} > 0$).*

2. *The project manager knows this and, consequently, also chooses low effort independent of Y_{pmp}.*

3. *The principal is indifferent between choosing agent $e1$ or $e2$ as project manager. Hence, any choice of the principal is part of an equilibrium.*

Proof: The *assistant* will always perceive the principal's decision as unkind. Hence he is never inclined to choose *high* effort out of kindness considerations. This is even reinforced by the fact that *high* effort is costly. Consequently the *assistant*'s optimal strategy is to choose l in every history in which he is active. As said above, the *project manager* knows this and figures that what ever he does the project will fail. Hence, his optimal choice is also to invest *low* effort. Given this the principal is indifferent between $bcd(e1)$ and $bcd(e2)$. ■

As one can see, if alternative procedures to take the same decision are neglected different equilibrium predictions result. This is not a mere artifact in this particular example but holds true also in other settings as will also be seen in the next application. Hence, the behavioral predictions that have been made in the hitherto existing literature on psychological games [e.g. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006)] are sensitive to the availability of different procedures to take the same decision.

Concluding, we have seen in this *principal-agent* example how procedural concerns inherently arise if agents are also psychologically motivated. Furthermore,

taking different procedures to take the same decision explicitly into account leads to behavioral predictions that differ from the results with mere consequentialist preferences, as traditionally assumed in economic theory, and they also differ from the results obtained in settings allowing for belief-dependent utilities but neglecting procedural choices. In the next application we will demonstrate the impact of procedural choices when agents are guilt averse [e.g. Charness and Dufwenberg (2006), Battigalli and Dufwenberg (2006)].

5.2 The 'So long, Sucker' Game with Guilt Aversion

Consider the game in Figure 4:[15]

[Figure 4 here]

This *'So long, Sucker' game* is a three-player game in which a player 1 seems to be trapped since he has to be unkind to one of the other players 2 and 3. This setting has already been analyzed by Nalebuff and Shubik (1988) and Dufwenberg and Kirchsteiger (2004) to explain why agents might choose to punish others (in this case player 1), i.e. reciprocate for any perceived unkindness, even if it is costly for themselves.

Different to Nalebuff and Shubik (1988) as well as Dufwenberg and Kirchsteiger (2004) assume that agent 1 is guilt averse. More precisely, assume that player 1 feels guilty, if the other two players get the impression that he did not treat them equally. This can be conceptualized as follows: At any endnode $z \in Z$ player j's inference ($j \in \{2,3\}$) with regard to what player 1 intended to give him by playing the procedural strategy ω_i is:

$$E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_j|\mathcal{H}\left(z\right),\omega_1\right].$$

Obviously, player j also has a belief in z about what player 1 intended to give to the other player q, where $q \neq j \wedge q \neq 1$:

$$E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_q|\mathcal{H}\left(z\right),\omega_1\right].$$

This means player j can infer player 1's intended difference, i.e. player 1's favoritism, between j and the other player q:

$$E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_j|\mathcal{H}\left(z\right),\omega_1\right] - E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_q|\mathcal{H}\left(z\right),\omega_1\right].$$

In line with the above-sketched intuition concerning player 1's guilt feeling and similar to Battigalli and Dufwenberg (2006), we say that player 1 is affected by 'guilt from blame', if players 2 and 3 get a perception of favoritism. His preferences can hence be written as:

$$u_1\left(z,\mu_{-1}^1,\mu_{-1}^2\right) = \pi_1 - \sum_j Y_{1j}\left(\left|E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_j|\mathcal{H}\left(z\right),\omega_1\right] - E_{\mu_j^1,\mu_j^2,p_0}\left[\pi_q|\mathcal{H}\left(z\right),\omega_1\right]\right|\right),$$

[15] Note that actions that are played by player *chance* with probability 0 are disregarded in Figure 4.

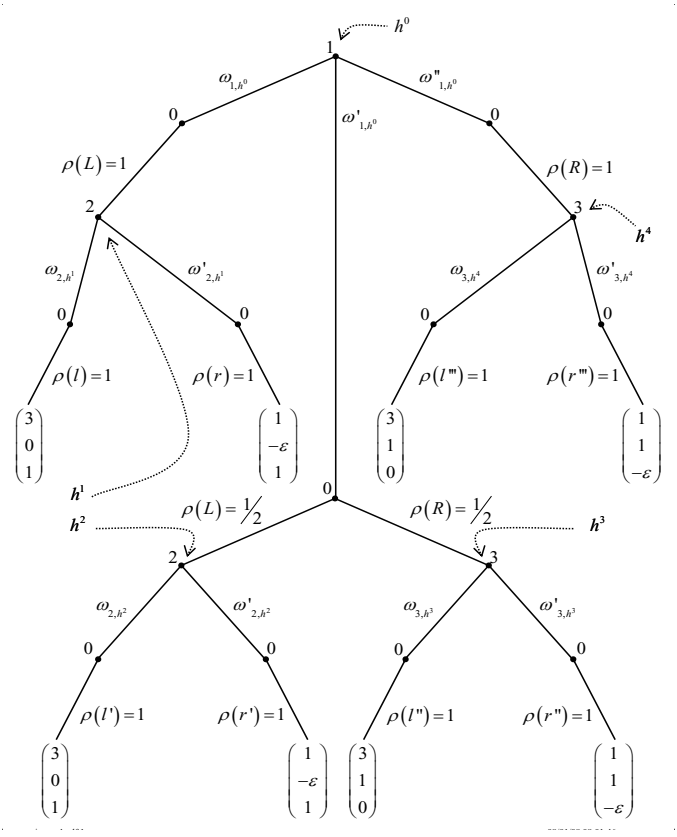

Figure 4: 'So long, Sucker' Game

where Y_{1j} is a positive constant capturing player 1's sensitivity to guilt and μ^1_{-1} and μ^2_{-1} are the other players first- and second-order beliefs. Note, in each history player 1 maximizes his utility u_i conditional on his belief up to the third-order because he takes his belief about the other players' second-order-belief, μ^2_{-1}, into account.[16] For simplicity, assume that players 2 and 3 perceive player 1's favoritism, but this does not have any effect on their utility.[17] In other words, players 2 and 3 are only concerned about their own material welfare.

As a benchmark let us first state how player 1 behaves if all players are only concerned about his own monetary payoff:

Result 4 *For all $\varepsilon \geq 0$, if all players are only interested in their own material payoff, then player* 1 *is indifferent with regard to his procedural choice.*

Proof: By backward induction, players 2 and 3 respectively choose $\{\omega_{2,h^1}, \omega_{2,h^2}\}$ and $\{\omega_{3,h^3}, \omega_{3,h^4}\}$ in the histories that they control. This implies player 1 knows that he gets 3 for sure independent of his own choice. Hence, he is indifferent between ω_{1,h^0}, ω'_{1,h^0} and ω''_{1,h^0}. ■

The situation changes assuming that player 1 is guilt averse as defined above. Given our set up with guilt aversion, it is possible to state the following result:

Result 5 *If $Y_{12} > 0$ and $Y_{13} > 0$, then the only SPE is given by:*

1. *Player 1 chooses ω'_{1,h^0} in history h^0.*
2. *Players 2 and 3 choose respectively $\{\omega_{2,h^1}, \omega_{2,h^2}\}$ and $\{\omega_{3,h^3}, \omega_{3,h^4}\}$ in the histories that they control.*

Proof: In line with player 2's and 3's preferences, let player 1's first-order-belief and player 1's belief about the second-order-belief of players 2 and 3 be $\{\omega_{2,h^1}, \omega_{2,h^2}\}$ and $\{\omega_{3,h^3}, \omega_{3,h^4}\}$. This implies that player 1's belief about players 2's and 3's perception of his intended favoritism is:

i) $(0-1) = -1$ (player 2) and $(1-0) = 1$ (player 3), if he chooses ω_{1,h^0},

ii) $\frac{1}{2}(0-1) + \frac{1}{2}(1-0) = 0$ (for both players), if he chooses ω'_{1,h^0} and

iii) $(1-0) = 1$ (player 2) and $(0-1) = -1$ (player 3), if he chooses ω''_{1,h^0}.

This means, his guilt feeling is minimized by playing ω'_{1,h^0}. Furthermore, his own expected material payoff given his first-order-beliefs is 3 independent of his procedural choice. Therefore, it is easy to see that the rational player 1 that is guilt averse optimally chooses the procedure ω'_{1,h^0} to take his decision between players 2 and 3. In addition, players 2 and 3 choose $\{\omega_{2,h^1}, \omega_{2,h^2}\}$ and $\{\omega_{3,h^3}, \omega_{3,h^4}\}$

[16] For comparison see Battigalli and Dufwenberg (2006)'s definition of 'guilt from blame'.

[17] Note that one could in addition assume that players 2 and 3 are disappointed due to the perceived favoritism. This would, however, only complicate the analysis without changing the results.

in line with player 1's beliefs because of their material concerns. This concludes the proof. ■

Note, also in this example it holds that procedures mitigate the own as well as the others' psychological payoffs.

Remark 2 *As in the previous example, ignoring player 1's possibility to choose a randomization procedure to take his decision, i.e.* ω'_{1,h^0}*, would lead to a different behavioral prediction. He would be indifferent between choosing* ω_{1,h^0} *and* ω''_{1,h^0}*.*

Hence, also here it holds that neglecting different procedures to take the same decision, as it is done in the hitherto existing literature on psychological games, leads to different equilibrium predictions. This highlights again, how procedural concerns can be conceptualized as an inherent part of the interaction of agents with belief-dependent utilities.

All in all, in this section we have used the concept of sequential psychological equilibrium developed in the previous section to formally demonstrate the impact of procedural choices on the strategic interaction of emotional agents. We have seen how procedural choices influence their interactions and how the inclusion of different procedures to take the same decision affects the behavioral predictions of the existing literature on psychological games.

6 Conclusion

Any decision in human interactions is inherently associated with a procedure which characterizes the way in which the decision is taken. This means it is impossible to take a decision without deciding first on *how* to take it. It is widely accepted in other scientific disciplines and it has been shown experimentally that people react differently in outcomewise identical situations depending on the procedures which have led to them. People are concerned about the way in which decisions are taken. Nevertheless traditional economic theory has neglected the impact of procedural choices on human interaction. It has ignored the influence of procedures on human interactions as traditional economic theory is based on consequentialist preferences which are difficult to reconcile with the existing evidence on procedural concerns.

Only in recent years psychological game theory has contested the consequentialist view in economic theory by assuming that agents also sense psychological payoffs which, broadly speaking, depend on agents' beliefs about the other's strategies and beliefs. It has been shown in our paper how procedural concerns can be conceptualized in a game theoretic setting assuming that agents are (also) incentivized by belief-dependent psychological payoffs. According to our approach procedural choices influence the beliefs that people hold with regard to others. In this way they mitigate the causal attribution of responsibilities and the evaluation of intentions.

With the help of two applications we have furthermore demonstrated i) how procedural concerns influence the strategic interaction of agents with belief-dependent utilities and ii) that the equilibrium predictions in the already existing literature

on psychological games are sensitive to the availability of different procedures to take the same decision. The hitherto existing literature on psychological games solely concentrates on situations in which agents are held fully responsible for all consequences of their actions. In contrast to this, in our class of procedural games agents can choose between different procedures. They can influence the process of causal attribution and the evaluation of intentions. Consequently, different equilibrium predictions arise.

Concluding, procedural concerns can play an important role in areas of eminent concern to economists. Hence, they should not be neglected.

7 References

1. Anderson, R. and Otto, A. (2003), *Perceptions of fairness in the justice system: A cross-cultural comparison*, Social Behavior and Personality, 31, 557-564.

2. Aumann, R. and Brandenburger, A. (1995), *Epistemic conditions for Nash equilibrium*, Econometrica, 63, 1161-1180.

3. Battigalli, P. and Siniscalchi, M. (1999), *Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games*, Journal of Economic Theory, 88, 188-230.

4. Battigalli, P. and Dufwenberg, M. (2007), *Dynamic Psychological Games* forthcoming in Journal of Economic Theory.

5. Battigalli, P. and Dufwenberg, M. (2006), *Guilt in Games*, American Economic Review, Papers and Proceedings, forthcoming.

6. Bies, R. and Tyler, T. (1993), *The 'litigation mentality' in organizations: A test of alternative psychological explanations*, Organization Science, 4, 352-366.

7. Blader, S. and Tyler, T. (2003), *A four-component model of procedural justice: Defining the meaning of a 'fair' process*, Personality and Social Psychology Bulletin, 29, 747-758.

8. Blount, S. (1995), *When Social Outcomes Aren't Fair: The Effect of Casual Attributions on Preferences*, Organizational Behavior and Human Decision Processes, 63, 131-144.

9. Bohnet, I. (2006), *How Institutions Affect Behavior: Insights from Economics and Psychology*, in: De Cremer, D., Zeelenberg, M. and Murnighan, K. (eds), *Social Psychology and Economics*, London: Lawrence Erlbaum, 2006, 213-238.

10. Bohnet, I. (2007), *Institutions and Trust: Implications for Preferences, Beliefs and Behavior*, Rationality and Society, 19(1), 99-135.

11. Bohnet, I. and Zeckhauser, R. (2004), *Trust, Risk and Betrayal*, Journal of Economic Behavior and Organization, 55, 467-484.

12. Bolton, G. and Ockenfels, A. (2000), *ERC: A Theory of Equity, Reciprocity, and Competition*, American Economic Review, 90(1), 166-193.

13. Bolton, G. Brandts, J. and Ockenfels, A. (2005), *Fair Procedures: Evidence from Games Involving Lotteries*, Economic Journal, 115(506), 1054-1076.

14. Bowles, S. (1998), *Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions*, Journal of Economic Literature, 36(1), 75-111.

15. Brandenburger and Dekel (1993), *Hierarchies of Beliefs and Common Knowledge*, Journal of Economic Theory, Elsevier, 59(1), 189-198.

16. Brandts, J., Gueth, W. and Stiehler, A. (2006), *I want you!: An experiment studying the selection effect when assigning distributive power*, Journal of Labor Economics, 13, 1-17.

17. Charness, G. and Rabin, M. (2002), *Understanding Social Preferences With Simple Tests*, Quarterly Journal of Economics, 117(3), 817-869.

18. Charness, G. (2004), *Attribution and Reciprocity in an Experimental Labor Market*, Journal of Labor Economics, 22(3), 553-584.

19. Charness, G. and Dufwenberg, M. (2006), *Promises and Partnership*, Econometrica, 74(6), 1579-1601.

20. Charness, G. and Levine, D. (2007), *Intention and Stochastic Outcomes: An Experimental Study*, Economic Journal, 117(522), 1051-1072.

21. Collie, T., Bradley, G. and Sparks, B. (2002), *Fair process revisited: Differential effects of interactional and procedural justice in the presence of social comparison information*, Journal of Experimental Social Psychology, 38, 545-555.

22. De Cremer, D. and van Knippenberg, D. (2003), *Cooperation with leaders in social dilemmas: On the effects of procedural fairness and outcome favorability in structural cooperation*, Organizational Behavior and Human Decision Processes, 91, 1-11.

23. Dufwenberg, M. and Kirchsteiger, G. (2004), *A Theory of Sequential Reciprocity*, Games and Economic Behavior, 47(2), 268-298.

24. Elster, J. (1998), *Emotions and Economic Theory*, Journal of Economic Literature, 36(1), 47-74.

25. Falk, A. and Fischbacher, U. (2006) *A Theory of Reciprocity*, Games and Economic Behavior, 54(2), 293-315.

26. Falk, A., Fehr, E. and Fischbacher, U. (2000), *Testing Theories of Fairness - Intentions Matter*, Working Paper No. 63, University of Zurich, forthcoming in: Games and Economic Behavior.

27. Fehr, E. and Schmidt, K. (1999), *A Theory Of Fairness, Competition, And Cooperation*, The Quarterly Journal of Economics, 114(3), 817-868.

28. Fehr, E. and Gaechter, S. (2000), *Fairness and Retaliation: The Economics of Reciprocity*, Journal of Economic Perspectives, 14(3), 159-181.

29. Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989) *Psychological Games and Sequential Rationality*, Games and Economic Behavior, 1, 60-79.

30. Heider, F. (1958), *The Psychology of Interpersonal Relations*, New York: John Wiley & Sons.

31. Hong, K. and Bohnet, I. (2005), *Status and Distrust: The Relevance of Inequality and Betrayal Aversion*, KSG Working Paper No. RWP04-041, Harvard PON Working Paper.

32. Kelley, H. (1967), *Attribution in social psychology*, Nebraska Symposium on Motivation, 15, 192-238.

33. Kelley, H. (1973), *The processes of causal attribution*, American Psychologist, 28, 107-128.

34. Konovsky, M. (2000), *Understanding Procedural Justice and Its Impact on Business Organizations*, Journal of Management, 26(3), 489-511.

35. Krawczyk, M. (2007), *A model of procedural and distributive fairness*, Mimeo, University of Amsterdam.

36. Kreps, D. and Wilson, R. (1982), *Sequential Equilibria*, Econometrica, 50(4), 863-894.

37. Lemons, M. and Jones, C. (2001), *Procedural justice in promotion decisions: using perceptions of fairness to build employee commitment*, Journal of Managerial Psychology, 16(4), 268-281.

38. Lind, E. and Tyler, T. (1988), *The social psychology of procedural justice*, New York, NY, US: Plenum Press.

39. Lind, E., Greenberg, J., Scott, K. and Welchans, T. (2000), *The winding road from employee to complainant: Situational and psychological determinants of wrongful termination claims*, Administrative Science Quarterly, 45, 557-590.

40. Murphy, K. (2004), *The role of trust in nurturing compliance: A study of accused tax avoiders*, Law and Human Behavior, 28, 187-209.

41. Nalebuff, B. and Shubik, M. (1988) *Revenge And Rational Play*, Papers 138. Princeton. Woodrow Wilson School - Public and International Affairs.

42. North, D. (1991), *Institutions*, Journal of Economic Perspective, 5(1), 97-221.

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

44. Roberts, K. and Markel, K. (2001), *Claiming in the name of fairness: Organizational justice and the decision to file for workplace injury compensation*, Journal of Occupational Health Psychology, 6, 332-347.

45. Ross, M. and Fletcher, G. (1985), *Attribution and Social Perception*, in G. Lindsey & E. Aronson (eds.). The Handbook of Social Psychology, 2, 73-114.

46. Sebald, A. (2007), *Procedural Concerns and Reciprocity*, ECORE Discussion Paper, 2007/54, Bruxelles.

47. Thibaut J. and Walker, L. (1975), *Procedural Justice*, Erlbaum, Hillsdale, NJ.

48. Trautmann, S. (2006), *A Fehr-Schmidt Model for Process Fairness*, Working Paper, CREED, University of Amsterdam.

49. Tyler, T. (1990), *Why people obey the law*, New Haven, CT, US: Yale University Press.

50. Tyler, T. (2003), *Procedural justice, legitimacy, and the effective rule of law*, in Tonry, M. (ed.), Crime and justice: A review of research, 30, 283-358.

51. Wenzel, M. (2002), *The impact of outcome orientation and justice concerns on tax compliance: The role of taxpayers' identity*, Journal of Applied Psychology, 87, 629-645.

8 Appendix

8.1 Proof of Result 2

Note, as the *principal-agent relation* is symmetric we will concentrate on the behavior of *project manager* and *assistant* rather than the behavior of agents $e1$ and $e2$ in the different possible roles. Let us start by looking at the behavior of the *assistant* following the decision by the principal to take the decision *behind closed doors*. Remember, when he has to decide about his effort level, he knows about the *project manager*'s effort level, the principal's procedural choice etc, i.e. he is perfectly informed about the history of the game he is in.

To start with, assume that in any history that the *assistant* can find himself following *bcd* he believes that the principal believes, i.e. the *assistant* 's second-order-belief, that:

1. the *assistant* chooses *low* effort, given that the principal has taken the decision *behind closed doors*,

2. the *project manager* and the *assistant* will choose *concentration* and *high* effort, given that the principal has taken the decision by means of a *selection tournament*.

This means, if it is the *assistant*'s turn and the principal has taken his decision *behind closed doors*, then the *assistant* believes that the principal intends to give him:

$$\pi_a(\cdot) = w\left(a\right). \tag{1}$$

Given this and the *assistant*'s second order belief his perceived kindness of the principal following *bcd* is:

$$\lambda_{apa} = w\left(a\right) - \frac{1}{2}\left(w(a) + \frac{1}{2}\left((w(a) - \upsilon) + (w(pm|s) - \upsilon)\right)\right) < 0, \tag{2}$$

where $\frac{1}{2}\left(w(a) + \frac{1}{2}\left((w(a) - \upsilon) + (w(pm|s) - \upsilon)\right)\right)$ is the *assistant*'s belief about the average that the principal could have given him and $\frac{1}{2}\left((w(a) - \upsilon) + (w(pm|s) - \upsilon)\right)$ is the *assistant*'s expected payoff given that the principal had chosen the *selection tournament*. Furthermore, the *assistant*'s kindness is either:

$$\kappa_{ap}(l) = \left(\pi_p\left(l, l\right) - w\right) - \frac{1}{2}\left(\left(\pi_p\left(l, h\right) - w\right) + \left(\pi_p\left(l, l\right) - w\right)\right) = 0, \tag{3}$$

if he chooses *low* effort or

$$\kappa_{ap}(h) = \left(\pi_p\left(l, h\right) - w\right) - \frac{1}{2}\left(\left(\pi_p\left(l, h\right) - w\right) + \left(\pi_p\left(l, l\right) - w\right)\right) = 0, \tag{4}$$

if he chooses *high* effort, conditional on the *low* effort by the *project manager*. Conditional on the *high* effort by the *project manager* his kindness towards the principal is either:

$$\kappa_{ap}(l) = \left(\pi_p\left(h, l\right) - w\right) - \frac{1}{2}\left(\left(\pi_p\left(h, h\right) - w\right) + \left(\pi_p\left(h, l\right) - w\right)\right) > 0, \tag{5}$$

if he chooses *low* effort or

$$\kappa_{ap}(h) = \left(\pi_p\left(h, h\right) - w\right) - \frac{1}{2}\left(\left(\pi_p\left(h, h\right) - w\right) + \left(\pi_p\left(h, l\right) - w\right)\right) < 0, \tag{6}$$

if he chooses *high* effort. To summarize the perceptions and the optimal behavior of the *assistant*:

1. If the principal chooses to take the decision *behind closed doors* and given the *assistant*'s aforementioned second-order beliefs, the perceived kindness of the *assistant* is negative independent of what the *project manager* does.

2. The *assistant*'s kindness towards the principal can: i) be 0 independent of his own choice, if the *pm* chooses low effort as well, or, ii) positive and negative given that the *project manager* has chosen *high* effort.

3. Hence, first, as effort is costly, he optimally chooses *low* effort, given *low* effort by the *project manager*. Secondly, as effort is costly and perceived kindness is negative, he also optimally chooses *low* effort given *high* effort by the *project manager*. Note, the *assistant*'s optimal behavior is in line with his second order beliefs.

In contrast to this, the *assistant*'s perceived kindness of the principal following the *selection tournament* is:

$$\begin{aligned}\lambda_{apa} &= \frac{1}{2}\left(w\left(a\right)-\upsilon\right)+\frac{1}{2}\left(w(pm|s)-\upsilon\right)\\ &- \frac{1}{2}\left(w(a)+\frac{1}{2}\left((w(a)-\upsilon)+(w(pm|s)-\upsilon)\right)\right)>0, \end{aligned} \tag{7}$$

where $\frac{1}{2}\left(w\left(a\right)-\upsilon\right)+\frac{1}{2}\left(w(pm|s)-\upsilon\right)$ is the *assistant*'s belief about what the principal intended to give him by choosing the *selection tournament*. Note, it can easily be seen that the *assistant*'s kindness towards the principal in the histories in which he is active is the same as under *bcd*, i.e. equations (3),(4),(5) and (6). From equation (7) we already see that the *assistant* perceives the *selection tournament* as kind. Given this the question arises whether and under what conditions this would make him choose *high* effort. Rationality requires that he chooses *high* effort if his utility from choosing *high* effort is bigger or equal to his utility from choosing *low* effort, i.e.:

$$u_a(h) \geq u_a(l), \tag{8}$$

which means

$$(w(a)-\upsilon)+Y_{ap}\left(\kappa_{ap}(h)\lambda_{apa}\right) \geq w(a)+Y_{ap}\left(\kappa_{ap}(l)\lambda_{apa}\right). \tag{9}$$

As $\kappa_{ap}(l)$ and $\kappa_{ap}(h)$ are 0 in histories following *st* and *low* effort by the *pm*, it can easily be seen that equation (9) never holds as $\upsilon > 0$. Hence, the *assistant* always chooses *low* effort under the *selection procedure* given that the *project manager pm* has chosen *low* effort as well. In case the *project manager* has chosen *high* effort, however, the situation changes. Equation (9) can be rewritten as:

$$Y_{ap} \geq \frac{\upsilon}{\lambda_{apa}\left(\kappa_{ap}(h)-\kappa_{ap}(l)\right)}. \tag{10}$$

Plugging in for λ_{apa} and $\kappa_{ap}(\cdot)$ gives:

$$\begin{aligned} Y_{ap} &\geq \frac{\upsilon}{\frac{1}{2}\left[\frac{1}{2}\left[w\left(pm|s\right)-w(a)\right]-\upsilon\right]\left[\pi_p\left(h,h\right)-\pi_p\left(h,l\right)+w(pm|f)-w(pm|s)\right]}\\ &> 0. \end{aligned} \tag{11}$$

This shows, if condition (11) holds, then the *assistant* optimally chooses *high* effort following the *selection tournament* and *high* effort by the *project manager*.

To summarize again, given that the principal uses the *selection tournament* to take his decision the *assistant* chooses *concentration* and:

1. *low* effort if the *project manager* has chosen *low* effort.

2. *high* effort if the *project manager* has chosen *high* effort and condition (11) holds.

3. *low* effort if the *project manager* has chosen *high* effort and condition (11) does not hold.

This brings us to the optimal behavior of the *project manager*. Consider first the *project manager*'s optimal behavior following the principal's choices to take the decision *behind closed doors*. From the above we know that the *project manager* and the principal know that the *assistant* always chooses *low* effort under *bcd*. Given this, the *project manager*'s perceived kindness of the principal's *procedural choice bcd*is:

$$\lambda_{pmppm} = w(pm|f) - \frac{1}{2}\left(w(pm|f) + \frac{1}{2}\left[(w\left(a\right) - \upsilon) + (w(pm|s) - \upsilon)\right]\right). \quad (12)$$

As $w(pm|f) = w(a)$, equation (12) reduces to:

$$\lambda_{pmppm} = w(a) - \frac{1}{2}\left(w(a) + \frac{1}{2}\left[(w\left(a\right) - \upsilon) + (w(pm|s) - \upsilon)\right]\right) < 0, \quad (13)$$

which is identical to equation (2). Hence, as the *assistant*'s optimal behavior is known to the *project manager* and the *project manager* also knows that the principal knows, the *project manager*' perceived kindness of the principal is identical to the assistant's perception following *bcd*. The same holds true for the *project manager*'s kindness. Given the optimal behavior of the *assistant*, the *project manager*'s kindness towards the principal reduces to:

$$\kappa_{pmp}(l) = (\pi_p\left(l, l\right) - w) - \frac{1}{2}\left((\pi_p\left(h, l\right) - w) + (\pi_p\left(l, l\right) - w)\right) = 0, \quad (14)$$

if he chooses *low* effort or

$$\kappa_{pmp}(h) = (\pi_p\left(h, l\right) - w) - \frac{1}{2}\left((\pi_p\left(h, l\right) - w) + (\pi_p\left(l, l\right) - w)\right) = 0, \quad (15)$$

if his effort choice is *high*. Concluding, as effort is costly also the optimal behavior of the *project manager* is *low* effort following the principal's procedural choice of *bcd*. What about the *selection tournament*? Remember, the assistant chooses *concentration* and l given that the *pm* chooses l and h if the *pm* chooses h and condition (11) holds. Hence, the *project manager*'s perceived kindness following the *selection tournament* is:

$$\begin{aligned} \lambda_{pmppm} &= \frac{1}{2}\left[(w\left(a\right) - \upsilon) + (w(pm|s) - \upsilon)\right] \\ &- \frac{1}{2}\left(w(a) + \frac{1}{2}\left[(w\left(a\right) - \upsilon) + (w(pm|s) - \upsilon)\right]\right) > 0. \end{aligned} \quad (16)$$

As can easily be seen, the *project manager* and the *assistant* feel equally treated. Hence, the perceptions about the principals kindness are identical (equations (16)

and (7)). The *project manager* kindness towards the principal, on the other hand, following is:

$$\kappa_{pmp}(l) = (\pi_p(l,l) - w) - \frac{1}{2}((\pi_p(h,h) - w) + (\pi_p(l,l) - w)) < 0, \tag{17}$$

if he chooses *low* effort and

$$\kappa_{pmp}(h) = (\pi_p(h,h) - w) - \frac{1}{2}((\pi_p(h,h) - w) + (\pi_p(l,l) - w)) > 0, \tag{18}$$

if his effort choice is *high*. From this follows that the *project manager* chooses *concentration* and *high* effort following the *selection tournament*, if:

$$u_{pm}(h) \geq u_{pm}(l), \tag{19}$$

which can also be written as

$$(w(pm|s) - \upsilon) + Y_{pmp}(\kappa_{pmp}(h)\lambda_{pmppm}) \geq w(pm|f) + Y_{pmp}(\kappa_{pmp}(l)\lambda_{pmppm}). \tag{20}$$

This reduces to:

$$Y_{pmp} \geq \frac{(w(pm|f) - w(pm|s)) + \upsilon}{\lambda_{pmppm}(\kappa_{pmp}(h) - \kappa_{pmp}(l))}. \tag{21}$$

Plugging in for the perceived kindness, λ_{pmppm}, and kindness, κ_{pmp} gives:

$$Y_{pmp} \geq \frac{(w(pm|f) - w(pm|s)) + \upsilon}{\frac{1}{2}\left[\frac{1}{2}[w(pm|s) - w(a)] - \upsilon\right][\pi_p(h,h) - \pi_p(h,l) + w(pm|f) - w(pm|s)]}.$$

One can easily see that:

$$Y_{ap} \geq Y_{pmp}, \tag{22}$$

as $(w(pm|f) - w(pm|s)) < 0$. Hence, the *project manager* optimally chooses *concentration* and *high* effort following the *selection tournament* already at lower levels of sensitivity to reciprocity compared to the *assistant*. This is due to the fact that he gets a financial reward for bringing *high* effort compared to the *assistant* who only supports him within the realm of his normal work and gets $w(a)$ independent of the success or failure of the project. Summarizing:

1. if the principal chooses to take the decision *behind closed doors*, both players optimally choose *low* effort in line with their beliefs and, in addition,

2. if conditions (11) and (22) hold, both choose *concentration* and *high* effort following the *selection tournament*.

Assume that both conditions (11) and (22) hold. Given this it is easy to see that the profit maximizing *principal* always chooses the *selection tournament* to take his decision, as this gives him a profit of $\pi_p(h,h)$. This concludes the proof. ■

How (too much) self-esteem facilitates contracts with subjective evaluations

(with Markus Walzl)

How (too much) self esteem facilitates contracts with subjective evaluations*

Alexander Sebald† Markus Walzl‡

Abstract

We analyze the impact of aggressive reactions to ego-threatening feedback on principal-agent relationships. More specifically, we show how peoples' desire to protect their self-esteem can explain the existence of contractual relationships in environments with unobservable effort and subjective measures of performance. We concentrate on situations in which neither effort nor output can be measured objectively as these constitute exactly the settings in which disagreements about effort and performance arise.

Keywords: Contracts, Subjective Evaluations, Self-Esteem, Ego-Threats, Feedback

JEL classification: D80, J41

1 Introduction

Self-esteem is one of the oldest and widely studied concepts in social psychology going back to the 1890s. It refers to peoples' self-evaluation or, in other words, the belief they hold about their self-worth. The unbroken attention that self-esteem attracts stems from the fact that people everywhere care about it, try to enhance, maintain and protect it [e.g. Greenwald (1980)]. Anything that gives a boost in self-esteem is almost universally welcome. People feel good when their self-perception is high and rising, and people feel bad when it is low or dropping. Hardly anyone enjoys events that constitute a blow or a loss to their self-esteem [Baumeister (2005)].

In recent years also economists have started to acknowledge the importance of self-esteem in decision making and strategic interactions [e.g. Köszegi (2006), Bénabou & Tirole (2002), Compte & Postlewaite (2004), Ellingsen & Johannesson (2007)]. It is argued that people strive for positive self-perceptions because

*Financial support by METEOR is gratefully acknowledged. The second author is financially supported by NWO.

†Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands, and ECARES, Université Libre de Bruxelles. Sebald is also member of ECORE, the recently created association between CORE and ECARES. E-mail: a.sebald@algec.unimaas.nl

‡Department of Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: m.walzl@algec.unimaas.nl

it entails a consumption, signaling and motivational value. Köszegi (2006), for example, endows individuals with 'ego-utility' and demonstrates the effects on choice between more or less ambitious tasks. In particular, this model explains the phenomenon of overconfidence by individuals who update believes according to Bayes' rule. Bénabou & Tirole (2002) and Compte & Postlewaite (2004), on the other hand, center on self-confidence as motivational value. It is argued that confidence in one's ability and efficacy can help individuals to *e.g.* undertake more ambitious goals. When people have imperfect knowledge about their own ability and/or when effort and ability are complements, then a higher self-confidence enhances peoples' motivation to act [Bénabou & Tirole (2002): 873].

Psychologists, however, have not only identified the implicit impact of self-esteem on information processing and motivation, but also stress the individual's eagerness to actively maintain and protect their positive self-perceptions [Greenwald (1980), Bushman & Baumeister (1998), Baumeister (2005)]. First, people protect their self-esteem by systematically taking credit for successes and denying blame for failures. Second, people have a tendency to uncritically accept positive feedback and eagerly search for flaws/faults in other's criticisms [*e.g.* Baumeister (2005), Greenwald (1980)]. Third and most importantly, psychologists have found that aggression and conflicts tend to result from positive self-images that are challenged or threatened [*e.g.* Baird (1977), Raskin et al. (1991), Bushman & Baumeister (1998)]. It is argued that hostile aggression is an expression of the self's rejection of ego-threatening evaluations received from other people [*e.g.* Baumeister et al. (1996)]. People with high self-esteem usually hold confident and highly favorable ideas about themselves, *i.e.* they exhibit ego-involvement, and react belligerently to ego-threatening feedback from others [Baird (1977), Shrauger & Lund (1975) and Korman (1969)]. These behavioral reactions have been found to be stronger the higher the perceived bias and the lower the perceived quality of the feedback source [*e.g.* Steelmann and Rutkowski (2004)].

In this paper we analyze the impact of aggressive reactions to ego-threatening feedback on principal-agent relationships. More specifically, we show how people's desire to protect their self-esteem can explain the existence of contractual relationships in environments with unobservable effort and subjective measures of performance. We concentrate on situations in which neither effort nor output can be measured objectively as these constitute exactly the settings in which disagreements about effort and performance arise.

In reality, it is very often impossible to objectively measure workers' and especially managers' individual contributions to the success of projects or firm values. Therefore it is widely prevalent to (also) take into account subjective evaluations in performance pay. Already in 1981 the Bureau of National Affairs reports, for example, that pay for performance systems involving subjective measures are more common than those involving only objective performance signals. Furthermore, Milkovich and Wigdor (1991) and Levine (2003) cite more recent evidence on the wide usage of subjective performance appraisal systems in performance pay in *e.g.* investment banks, law firms and consultancies.

Against this background, consider the following example. Suppose a principal

wants to motivate an agent to spend effort on a complex good or service. Neither the agent's effort nor the outcome of the project (the quality of the good or service) is observable. The only information the principal and the agent receive are private, *i.e.* subjective, signals about the effort of the agent. These signals are imperfectly correlated with each other and to the actual effort level. To motivate the agent to spend positive effort, a contract has to specify payments which increase in the subjective signal of the principal (an increase in the reported signal of the agent would just motivate him to misrepresent his information). However, due to the imperfect signal technology the principal can credibly report that he has received a signal of low effort regardless of his actual private information. As payments increase in the subjective signal of the principal, he is always better off by misrepresenting his information and pay the agent the minimum wage. This will be anticipated by the agent and subgame-perfect equilibrium efforts are zero, *i.e.* no principal agent relationship can be established.

In a recent paper [MacLeod (2003)], it has been assumed that the principal can credibly promise to make payments to a third party (contingent on the signal configuration). In the simplest case of two different performance signals, the agent is indifferent between telling the truth and lying (*i.e.* he tells the truth by default) and the principal promises to pay a third party if he pays the agent according to a bad signal and the agent reports a good signal in the optimal contract. The payments to the third party are thereby fixed in such a way that the principal's truth-telling constraint is satisfied. The complete flexibility of third-party payments thereby ensures that a relationship (*i.e.* a positive effort level) can be established regardless of the parameters of the model (*e.g.* the correlation between the principal's and the agent's signal, the size of the project etc.). Of course this result crucially depends on the credibility of payments to the third party. In particular, while the principal cannot credibly promise the agent to report his signal truthfully, it is assumed that he can make such a promise to the third party. This reminds a little bit of the mediator function of mafia clans in business relations - even though one cannot promise to be honest with a client, no-one will lie to the godfather. To explain the widespread use of subjective information in particular in labor market relations far away from enforcement through the Corleone family, MacLeod (2003) refers to the third party payments as anticipations of future conflict in an un-modelled dynamic game.

In this paper, we illuminate the un-modelled conflict in Macleod (2003) and show that principal-agent relationships can be established on the basis of subjective performance evaluations, if agents try to defend their self-esteem through the creation of trouble or aggressive actions. In line with the aforementioned psychological evidence, we assume that agents perceive a negative psychological payoff from ego-threatening performance evaluations through the principal (*i.e.* the agent suffers from a bad performance evaluation by the principal, if he does not share this opinion based on his own subjective signal), and that he can reduce this negative psychological payoff through trouble/aggression. If the agent creates trouble, the principal will face costs of conflict.[1] The costs of conflict play the very

[1]This mechanism could be interpreted as negative reciprocity. Unlike the existing models of

same role as MacLeod (2003)'s third-party payments - they enforce truth-telling by the principal. In our setting, however, they depend on the agent's sensitivity to ego-threats, the quality of the information technology etc. and therefore do not exhibit the same flexibility as third-party payments in MacLeod (2003) that can be optimally chosen. Our model identifies conditions on conflict levels, project returns, quality of information, and sensitivity to ego-threats which promote or rule-out the implementation of positive effort levels.

Our model is closely related but conceptually different from Ellingsen & Johannesson (2007)'s model of self-esteem. In contrast to us, they model a situation in which agents sense a psychological payoff from being esteemed by others. Agents in their setting take pride in what opponents think about them. More formally, agents derive utility from their belief about the others' perceptions about their type. In contrast to this, in our setting agents' sense a threat to their self-esteem, if their own positive self-perception is not confirmed by the feedback of others.

In the baseline model, we assume that the principal designs a contract, but has no influence on the quality of the information technology. We show that the interval of credible bonus payments in case of a positive evaluation by the principal enlarges in the agent's sensitivity to ego-threats, the level of conflict faced by the principal, and the correlation of subjective signals. In particular, higher bonuses become credible if the level of conflict increases. However, we demonstrate that the bonus which makes it incentive compatible for the agent to choose a certain effort level does not have to be credible. In particular, the incentive compatible bonus increases in the agent's sensitivity to ego-threats and the probability of conflict (*i.e.* the principal does not only compensate the agent for (marginal) effort costs but also for (marginal) psychological cost). To guarantee the existence of a positive implementable effort level, the information technology has to be sufficiently unbiased (*i.e.* the relative probability of conflicting signals has to be low), and the ratio of psychological costs for the agent and conflict costs for the principal has to be sufficiently small (*i.e.* the aggressive action has to be 'effective'). If there exists a positive implementable effort level, we demonstrate that a principal agent relationship is established if project returns are sufficiently large. However, the principal-agent relationship establishes a first best solution if and only if signals are perfectly correlated. Hence, the additional agency costs due to a positive probability of ego-threats and the endogenously determined interval of credible bonuses introduce important frictions which unambiguously reduce welfare.

In an extension of this model, we allow the principal to also (costlessly) choose the quality of his signal. We formalize the findings of Steelmann and Rutkowski (2004) and assume that psychological costs of the agent are decreasing in the quality of the signal but do not vanish in the limit of a perfect performance signal of the principal. As an immediate consequence, the bias of the information technology, the psychological costs for the agent, and expected conflict with the principal

reciprocity [Rabin (1993), Dufwenberg & Kirchsteiger (2004) and Falk & Fischbacher (2006)], however, what is considered to be fair or unfair in our model does not depend on beliefs about strategies and their associated outcomes, but rather on the perceived fairness/correctness of (reported) signal constellations.

decrease in the quality of the principal's signal. While a lower bias and lower psychological costs ceteris paribus promote the implementability of a positive effort level (see above), lower expected costs of conflict diminish the set of credible bonuses. These countervailing effects yield the following results. For sufficiently small costs of conflict for the principal, there is no signal quality and no contract which could implement a positive effort and yield positive profits for the principal. If costs of conflict are sufficiently large, a positive effort level will be implemented, but the principal does not choose a perfect signal quality (even though this is assumed to be costless), neither does he implement a first best solution. For intermediate levels of conflict the principal will implement a first best solution if and only if signals are perfectly correlated. Otherwise he may choose a perfect signal but never achieves a first best solution. The non-feasibility of the first best solution is again driven by the additional (psychological) agency costs.

In sum, our model demonstrates that principal-agent relationships may well be feasible due to the agent's eagerness to protect their self-esteem through aggressive acts. However, this psychological mechanism is not sufficiently flexible in order to allow for an implementation of a first best solution(which would be achievable in our setting if signals were observable). Moreover, a break-down of the relationship (the sole implementability of an effort of zero) cannot be ruled out even if the principal can costlessly choose the quality of his own signal.

In section 2 we present the model, the timing of the game and the first best solution. In section 3 we define the optimal contract, comparative statics and a welfare analysis under the assumption that the quality of the principal's subjective performance signal is exogenously given. In section 4 we enrich the analysis by assuming that the principal can choose among different evaluation procedures that differ in quality. With section 5 we conclude.

2 The model

In this section we present the building blocks of our model and derive the first-best outcome.

Production Technology Consider a principal who decides upon undertaking a project which generates a value of $\phi > 0$ if successful. The project requires effort of an agent. Assume that if the agent spends effort $p \in [0, 1]$, the project will be successful (create value ϕ) with probability p. The project is a complex good or service and its success is not verifiable, *i.e.* contracts contingent on the generation of ϕ are not feasible.

Information Technology Neither principal nor agent can observe whether the project is successful or not. Both rather receive a private signal about the agent's performance. The principal receives $s_P \in S_P$, where $S_P = \{U, A\}$, *i.e.* performance is regarded as either acceptable (A) or unacceptable (U). Analogously, the agent receives $s_A \in S_A$ with $S_A = \{U, A\}$. The signals s_P and s_A are non-verifiable private pieces of information of the principal and the agent, respectively.

The signals are informative with respect to the success of the project in the following sense. If the project is not successful (which happens with probability $(1-p)$), principal and agent receive the signal $s_P = s_A = U$. Now denote by γ_{kl} the probability that $s_P = k$ and $s_A = l$ given that the project is a success. Then, the ex-ante probability for the signal pair $s_P = U$ and $s_A = A$, for instance, will be $p\gamma_{UA}$. Following [MacLeod (2003), p.227], we introduce the *perceived bias* of the relationship by

$$\alpha = \frac{\gamma_{UA}}{\gamma_{AA}}$$

where $\alpha = 0$ indicates that the principal always agrees with the agent upon a good signal, while $\alpha = \infty$ would imply that agreement never occurs. *I.e.*, the *perceived bias* is a likelihood ratio which represents the agent's belief about the principal perceiving his performance as acceptable conditional on his own perception of an acceptable performance.

Assumption 1 *We assume that the principal's and agent's signals are positively correlated which each other, i.e.* $\gamma_{AA}\gamma_{UU} - \gamma_{UA}\gamma_{AU} > 0$. *In particular, this implies that* $\alpha < \infty$.

The Game The timing of the game is as follows:

1. The principal offers a contract to the agent and the agent decides upon acceptance.[2] Upfront payments are arranged.
2. The agent decides upon effort p.
3. The project generates value ϕ with probability p.
4. The principal receives s_P and the agent receives s_A. The principal and the agent report (not necessarily truthfully) on s_P and s_A. Denote the reports by t_P and t_A, respectively. t_P and t_A are verifiable.
5. Payments contingent on t_P and t_A are arranged.
6. Contingent on s_A and received payments, the agent decides upon retaliation (*i.e.*, spends effort q).

Agent For an effort of p the agent incurs costs $v(p)$ with $v(p) : [0,1] \to R^+$ in C^2, $v'(0) = 0$, $v''(p) > 0$ and $\lim_{p\to 1} v(p) = \infty$.

First Best Effort Level Would the principal have access to the agent's production technology, his effort choice would solve $v'(p) = \phi$. For further reference, we will denote the first best effort level by p_{FB}. Our assumptions on $v(p)$ ensure that $p_{FB} \in (0,1)$.

[2] In section 4, the principal can also influence the perceived bias of the relationship - for instance, by choosing the quality of his own signal.

Psychological Payoffs The agent is risk-neutral and senses a psychological payoff that depends on his own private signal about performance, s_A, and the reported signal of the principal, t_P. More specifically, the agent's utility function reads:

$$U = c - v(p) - Y(s_A, t_P, \alpha)(1 - q) - w(q) \tag{1}$$

Thereby, c denotes the wage payment and $Y(s_A, t_P, \alpha)$ represents the agent's psychological costs for a given configuration of (reported) signals and the *perceived bias* of the relationship, α. With respect to $Y(s_A, t_P, \alpha)$ we proceed with the following specification.

Assumption 2 *(i)* $Y(s_A = U, t_P, \alpha) = 0$ *for all* t_P *and* α.

(ii) $Y(s_A, t_P = A, \alpha) = 0$ *for all* s_A *and* α.

(iii) $Y(A, U, \alpha) \in C^1$, $Y(A, U, 0) > 0$, *and* $\frac{\partial Y(A,U,\alpha)}{\partial \alpha} > 0$.

Part (i) and (ii) formalize the concept of an ego-threat. Individuals with low self-esteem (here, represented by $s_A = U$) do not exhibit ego involvement and show no reaction to feedback (be it confirming or threatening) [see Baumeister, Smasrt & Boden (1996)]. If the individual is sensitive to ego-involvement ($s_A = A$), it uncritically accepts positive (or confirming) feedback [see Baumeister (2005)] - formalized with zero psychological payoffs in this case - but suffers from negative (or threatening) assessments [see e.g. Bushman and Baumeister (1998)] - which amounts to non-zero psychological costs in our model. Finally, Part (iii) follows the findings of *e.g.* Steelmann and Rutkowski (2004) in assuming that psychological costs increase in the perceived bias of the relationship (or the quality of the information technology).

q is the level of conflict (or retaliation) created by the agent with $w(q) \in C^2$, $w(0) = 0$, $w'(0) = 0$, $w'' > 0$ and $w'(1) < \infty$.

For further reference, we summarize some results concerning the agent's optimal conflict level.

Lemma 1 *Let* $c > 0$, $p \in (0, 1)$, *and* $Y(s_A, t_P, \alpha)$ *satisfy Assumption 2.*

(i) Suppose $s_A = U$ *and/or* $t_P = A$. *Then,* $Y(s_A, t_P, \alpha) = 0$ *and the agent chooses* $q = 0$.

(ii) Suppose $Y(s_A, t_P, \alpha) \geq w'(1)$. *Then, the agent chooses* $q = 1$.

(iii) Suppose $0 < Y(s_A, t_P, \alpha) < w'(1)$. *Then, the agent chooses*

$$q = \arg\left(Y(s_A, t_P, \alpha) = w'(q)\right) > 0.$$

In this case $\frac{dq}{dY} > 0$ *holds for the optimal* q.

Proof. Follows from Eqn. 1 and Assumption 2. ■

With Assumption 2 and Lemma 1, the agent retaliates (i.e., $q > 0$) if and only if $s_A = A$ (ego-involvement) and $s_P = U$ (ego-threat). If psychological costs are larger than maximal marginal costs of conflict (Part (ii)), the agent exerts maximal level of conflict $q = 1$. If psychological costs are below that level, $q \in (0, 1)$ holds. For further reference we abbreviate $Y(A, U, \alpha) \equiv Y$.[3] Moreover, q^* will henceforth denote the conflict level for the configuration $t_P = U$ and $s_A = A$ - as $q = 0$ for all other configuration, no confusion should arise.

Principal The principal's expected profit is given by:

$$\Pi = p\phi - E\{c\} - E\{q\}\psi, \tag{2}$$

where $p\phi$ is the expected benefit which the agent generates, $E\{c\}$ are the expected wage cost of employing the agent, and $E\{q\}\psi$ are the expected costs of conflict due to the reciprocal behavior of the agent. As our assumptions on $w(q)$ ensure that $q \in [0, 1]$, we can interpret q as the probability with which the agent creates costs of $\psi > 0$ for the principal. First best profits are given by $\Pi_{FB} = p_{FB}\phi - v(p_{FB})$.

Contracts A contract Γ specifies payments contingent on verifiable events, *i.e.* $\Gamma = \{c_{kl} \mid k \in S_P, l \in S_A\}$. The agent accepts a contract if he expects a (weakly) positive utility from it (individual rationality) and chooses p as to maximize his utility (incentive compatibility). In this case, we say that Γ *implements* p. Principal and agent report their signal truthfully if and only if they weakly benefit from doing so.

3 Cost Minimizing Contracts

In this section we characterize cost minimizing contracts which implement a certain effort level p (*i.e.,* satisfy individual rationality and incentive compatibility constraints) and discuss their feasibility in the presence of truth-telling constraints for the principal and the agent.

Reduced Form Contracts For a given contract Γ and signals s_P and s_A, the principal and agent decide upon their report. Let $\sigma_P : S_P \to \Delta(S_P)$ and $\sigma_A : S_A \to \Delta(S_A)$ be the principal's and agent's reporting strategies (*i.e.,* mappings from the set of signals S_P and S_A to the set of probability distributions over S_P and S_A, respectively). Suppose that (σ_P^*, σ_A^*) is the pair of optimal reporting strategies for contract Γ. Then, the revelation principle implies that there exists a contract $\hat{\Gamma}$ which implements the same effort at the same costs and induces truthful reports by principal and agent. We will, henceforth, restrict our analysis to this type of (revelation) contracts. The following results further simplify the analysis.

[3] In section 4, α is endogenized and we will refer to $Y(A, U, \alpha)$ as $Y(\alpha)$.

Lemma 2 *Suppose there exists a contract* Γ *which implements* $p > 0$*. Then, there always exists a contract* $\hat{\Gamma}$ *which implements* p *at weakly lower costs such that*

(i) $c_{kl} = c_{km} \equiv c_k$ *for all* $l, m \in S_A$ *and* $k \in S_P$*. In particular, the principal and the agent tell the truth.*

(ii) $c_A > c_U$*.*

Proof. See Appendix ■

For convenience, we will from now on write $c_A = f + b$ and $c_U = f$ and a contract reads $\Gamma = (b, f)$. By Lemma 2(ii), $b > 0$. [4]

Then, the principal's and agent's truth-telling decisions constitute the following simple normal-form game (with the principal being the row- and the agent being the column player, the action-space is given by S_A and S_P, respectively).

	A	U
A	$p\phi - f - b,\ f + b$	$p\phi - f - b,\ f + b$
U	$p\phi - f - q^*\psi,\ f$	$p\phi - f,\ f$

Truth-Telling Notice that as the agent's monetary compensation is independent of his own report, he is indifferent between reporting either of his signals for any given report of the principal. We assume that he tells the truth in case of indifference such that the agent's truth-telling constraint is trivially fulfilled (see Lemma 2(i)).

Suppose $s_P = A$. Then, the principal tells the truth, whenever his payoff from doing so (which reads $p\phi - f - b$) is larger than his payoff from reporting $t_P = U$ (which reads $p\phi - f - Pr(s_A = A \mid s_P = A)q^*\psi$). This means the principal reports $t_P = A$ if

$$b \leq \frac{\gamma_{AA}}{(\gamma_{AA} + \gamma_{AU})} q^*\psi \equiv b^{max}. \tag{3}$$

On the other hand, suppose $s_P = U$. Then, the principal tells the truth, whenever his payoff from doing so (which reads $p\phi - f - Pr(s_A = A \mid s_P = U)q^*\psi$) is larger than his payoff from reporting $t_P = A$ (which reads $p\phi - f - b$). In other words, the principal reports $t_P = U$ if

$$b \geq \frac{\gamma_{UA}}{(\gamma_{UA} + \gamma_{UU})} q^*\psi \equiv b^{min}. \tag{4}$$

For further reference we collect the following properties of b^{max} and b^{min}.

Lemma 3 *Comparative Statics of* b^{max} *and* b^{min}

(i) $b^{max} > 0$*.*

[4] f can be interpreted as an up-front payment which implies a payment of zero if the principal reports $t_P = U$ and a payment of b (a bonus) if he reports $t_P = A$. Henceforth, we assume that f is chosen in such a way that the contract is individually rational.

(ii) $b^{max} > b^{min}$.

(iii) $\Delta b \equiv b^{max} - b^{min}$ *is monotone increasing in* q^*, ψ, *and the correlation between* s_P *and* s_A.

(iv) b^{min} *and* b^{max} *are monotone increasing in* q^*, ψ. b^{min} *is monotone increasing in* γ_{UA} *and* b^{max} *is monotone increasing in* γ_{AA}.

Proof. (i) and (ii) follow from the positive correlation of signals, *i.e.,* $\gamma_{AA}\gamma_{UU} > \gamma_{AU}\gamma_{UA}$.

(iii) Follows from $\Delta b = \frac{\gamma_{AA}\gamma_{UU} - \gamma_{AU}\gamma_{UA}}{(\gamma_{AA}+\gamma_{AU})+(\gamma_{UA}+\gamma_{UU})} q^* \psi$.

(iv) Follows from Eqns. 3 and 4. ■

Part (i) and (ii) of Lemma 3 imply that $[b^{min}, b^{max}]$ is always a non-empty interval, *i.e.,* the principal can credibly offer a bonus $b \in [b^{min}, b^{max}]$. From Lemma 3(iii) and (iv) it follows that the distance between b^{max} and b^{min} (the maximal and minimal credible bonuses) gets larger and the respective interval is shifted towards larger bonuses as q^* or ψ increases. Hence, the larger the potential conflict level, the higher are the bonuses that can be implemented. In fact, for every bonus b there is a conflict level ψ such that b is credible. The distance between b^{min} and b^{max} also becomes larger as the correlation between the principal's and the agent's signal increases.

Incentive Compatibility For a given contract $\Gamma = (f, b)$, the agent chooses effort p so as to maximize his utility (see Eqn. 1) while anticipating the generation of ex-post conflict at level q^* as depicted in Lemma 1. Hence, he maximizes

$$U(p) = p(\gamma_{AA} + \gamma_{AU})b + f - v(p) - p\gamma_{UA}(Y(1 - q^*) + w(q^*))$$

which induces the first order condition[5]

$$b(p) = \frac{v'(p) + \gamma_{UA}(Y(1 - q^*) + w(q^*))}{\gamma_{AA} + \gamma_{AU}}. \tag{5}$$

Note that $\frac{d^2U(p)}{dp^2} = v''(p) > 0$ such that the agent's optimization problem is well-behaved. For further reference we collect the following properties of $b(p)$.

Lemma 4 *Comparative Statics of* $b(p)$

(i) Suppose $p > 0$. *Then,* $b(p) > 0$.

(ii) $\lim_{p \to 0} b(p) > 0$ *if* $\gamma_{UA} > 0$.

(iii) $\frac{db(p)}{dp} > 0$.

(iv) $\frac{db(p)}{dY} > 0$.

[5]We denote a bonus which implements an effort level of p by $b(p)$.

(v) $\frac{db(p)}{d\gamma_{UA}} > 0$.

Proof. Follows from Eqn. 5. ■

Eqn (5) shows that the bonus has to overcome marginal effort costs *and* marginal psychological costs. If the principal wants to induce a positive effort level, he has to offer a positive bonus (Part (i)). However, the required bonus does not vanish in the limit of small efforts, because marginal psychological costs do not vanish for $p = 0$. Parts (iii)-(v) of Lemma 4 indicate that the necessary bonus increases in target effort p, psychological costs Y, and the conditional probability of conflict (γ_{UA}).

Individual Rationality The agent accepts a contract $\Gamma = (f, b)$ whenever his expected utility from it is weakly positive, i.e.

$$p(\gamma_{AA} + \gamma_{AU})b + f - v(p) - p\gamma_{UA}(Y(1 - q^*) + w(q^*)) \geq 0.$$

To maximize her profits, the principal sets for a given bonus b the upfront payment to

$$f(b) = -p(\gamma_{AA} + \gamma_{AU})b + v(p) + p\gamma_{UA}(Y(1 - q^*) + w(q^*)).$$

The upfront-payment can well be negative (franchise fee) as the agent is not protected by limited liability. Note in particular that $f(b)$ can always be fixed such that the agent does not receive any rents from the relationship.

Implementable Efforts We call a certain effort level $p > 0$ *implementable* if $b(p) \in [b^{min}, b^{max}]$ and state the following result.

Lemma 5 *There exists an implementable effort level $p > 0$ if and only if $b^{max} > b(0)$, i.e., $q^*\psi > \alpha(Y(1 - q^*) + w(q^*))$.*

Proof. "⇒". Suppose $p > 0$ is implementable. Then, $b(p) \in [b^{min}, b^{max}]$. According to Lemma 4(iii), $\frac{db(p)}{dp} > 0$ such that $b(0) < b^{max}$.

"⇐". Suppose $b^{max} > b(0)$. Then, by continuity of $b(p)$ in p, Lemma 3(iii), and Lemma 4(iii), there exists a $p > 0$ with $b(p) \in [b^{min}, b^{max}]$ such that $p > 0$ is implementable. ■

Lemma 5 shows that there exists an implementable effort level $p > 0$ whenever the perceived bias α is sufficiently low or the ratio of costs of conflict for the principal to psychological costs for the agent is sufficiently large.[6] In particular, no conflict (*i.e.*, $\psi = 0$) or no psychological costs ($Y = 0$) and therefore no retaliation $q^* = 0$ imply non-implementability of an effort level $p > 0$. The existence of implementable effort levels is promoted by a small perceived bias of the relationship and by high costs of conflict for the principal and low but non-zero costs for the agent.

[6]Note that a perfect correlation of signals (*i.e.* $\alpha = 0$) or a sufficiently high level of conflict ψ guarantees the existence of an implementable positive effort level.

Effort Costs To implement an (implementable) effort $p > 0$ the principal's costs are $C(p) = f + p(\gamma_{AA} + \gamma_{AU})b(p) = v(p) + p\gamma_{UA}((1 - q^*)Y + w(q^*))$. Note that $C(p)$ is convex and that $C(0) = 0$. Moreover, we adopt the convention that an effort $p > 0$ which is not implementable requires infinite costs.

Optimal Effort The principal's profit now reads

$$\Pi(p) = p\phi - p\gamma_{UA}q^*\psi - C(p)$$

which is concave for all implementable $p > 0$ and zero for $p = 0$. We denote the maximum of $\Pi(p)$ on $[0, 1]$ by p^*. p^* will be referred to as the optimal effort level chosen by the principal.

Proposition 1 *Optimal Effort Level*

(i) Suppose $b(0) \geq b^{max}$. Then, $p^* = 0$.

(ii) Suppose $b(0) < b^{max}$. Then, there exists $\underline{\phi} > 0$ such that $p^* = 0$ for $\phi \leq \underline{\phi}$ and $p^* > 0$ for $\phi > \underline{\phi}$.

Proof. *Part (i).* This follows directly from Lemma 5.
Part (ii). With $b(0) < b^{max}$ it follows from Lemma 5 that there exists an implementable effort level $\overline{p} > 0$, *i.e.,* $C(\overline{p}) < \infty$. Observe that $\Pi(p)$ is a linear increasing function of ϕ for a given implementable effort $p > 0$. Hence, there exists $\underline{\phi}$ such that $\Pi(\overline{p}) = 0$. Now take the implementable effort $\overline{p} > 0$ which leads to the lowest such $\underline{\phi}$. As $\Pi(p) < 0$ for $\phi = 0$ and $p > 0$, it follows that $\underline{\phi} > 0$. By construction, there is no positive implementable effort level for any $\phi \leq \underline{\phi}$ which leads to positive profits. Hence, $p^* = 0$ in this case. As $\Pi(\overline{p})$ is monotone increasing in ϕ, $p^* > 0$ if $\phi > \underline{\phi}$. ■

According to Proposition 1, there will be no principal-agent relationship (i.e., $p^* > 0$) whenever no effort is implementable or returns to the project are too small. Recall that the first best solution always requires a positive effort level ($p_{FB} > 0$) which implies a welfare loss due to the subjectivity of information. In case of Part (i), the perceived bias of the relationship is too large or the ratio of costs of conflict for the principal and the agent is too small to overcome the truth-telling problem of the principal (see the discussion of Lemma 5). In Part (ii), implementable effort levels exist but agency costs are too high to generate positive profits for the principal.

In the reminder of this section, we will further analyze the case of existing implementable effort levels (i.e., a sufficiently low perceived bias α or a sufficiently large ratio of conflict costs for principal and agent) which also generate positive profits for the principal (i.e., sufficiently high returns ϕ). In particular, we are interested in the comparative statics of p^* with respect to the parameters of our model.

To this end we neglect for the moment the truth-telling constraints of the principal, *i.e.,* the principal's profit is given by $\Pi(p)$ with $C(p) = v(p) + p\gamma_{UA}((1 -$

$q^*)Y + w(q^*))$ for all $p > 0$. This profit function is concave and we denote the unique maximum by $\tilde{p}$. Furthermore, denote by p^{min} the (unique) effort level for which $b(p) = b^{min}$ and by p^{max} the (unique) effort level for which $b(p) = b^{max}$.[7] Then, the following cases can be distinguished.

Lemma 6 *Suppose $p^* > 0$.*

(i) Binding Lower Truth-Telling Constraint: If $0 < \tilde{p} < p^{min}$, then the principal implements $p^* = p^{min}$ by paying b^{min} [Figure 1].

(ii) Binding Upper Truth-Telling Constraint: If $\tilde{p} > p^{max}$, then the principal implements $p^* = p^{max}$ by paying b^{max} [Figure 2].

(iii) Non-Binding Truth-Telling Constraint: If $\tilde{p} \in [p^{min}, p^{max}]$, then the principal implements $p^* = \tilde{p}$ by paying [Figure 3]:

$$b(\tilde{p}) = \frac{v'(\tilde{p}) + \gamma_{UA}(Y(1-q^*) + w(q^*))}{\gamma_{AA} + \gamma_{AU}}.$$

Proof. Follows from Proposition 1 ■

[Figures 1-3 here]

With Lemma 6, the comparative statics of p^* follows from the respective results for p^{min}, p^{max}, and $\tilde{p}$ which are implicitly determined by Lemma 3 and Lemma 4.

Consider first the impact of conflict costs ψ. Recall that $b(p)$ is independent of ψ while b^{min} and b^{max} are monotone increasing. But as $b(p)$ increases in p, it follows that p^{min} and p^{max} are increasing in ψ. Moreover, the increase of p^{max} is steeper than the increase of p^{min}. Hence, higher costs of conflict for the principal shift the interval of implementable efforts towards larger efforts and increases the distance between p^{min} and p^{max}.

Another clear cut result can be derived for the impact of γ_{AA}. b^{max} increases in γ_{AA}, b^{min} is independent of γ_{AA}, and $b(p)$ decreases. Hence, the larger γ_{AA} the larger p^{min} and p^{max}. Moreover, the increase of p^{max} is steeper such that a larger probability of consensus about acceptable efforts (measured by γ_{AA}) also shifts the interval of implementable efforts towards larger efforts and increases the distance between p^{min} and p^{max}.

The impact of Y and γ_{UA} is more subtle. On the one hand, $b(p)$ is increasing in Y and γ_{UA} such that larger bonuses are needed for the implementation of a given effort level. As b^{max} is independent of γ_{UA}, this implies that p^{max} decreases in γ_{UA}. However, b^{min} is increasing in γ_{UA} such that ceteris paribus the bonus has to be larger in order to be credible. Depending on *e.g.,* the level of ψ, one or the other effect dominates and p^{min} is increasing or decreasing in the conditional probability of conflict γ_{UA}. It is, however, clear that an increase in γ_{UA} reduces the distance between p^{min} and p^{max}. The comparative statics with respect to Y

[7] Uniqueness follows directly from the monotonicity of $b(p)$ in p and the respective independence of b^{min} and b^{max}.

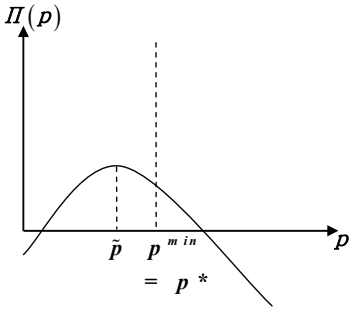

Figure 1: Binding Lower Truth-Telling Constraint

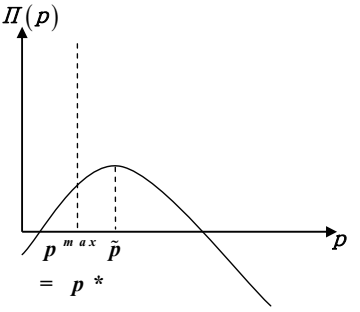

Figure 2: Binding Upper Truth-Telling Constraint

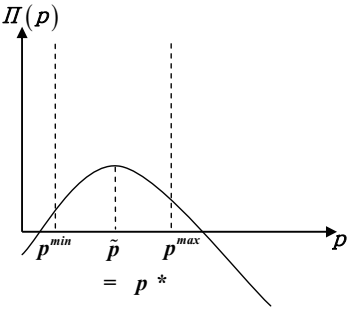

Figure 3: Non-Binding Truth-Telling Constraint

is similar because b^{min} and b^{max} both increase in Y (because q^* increases) and so does $b(p)$. Again, it depends on the other parameters (*e.g.*, the level of ψ) which of these effects dominates.

The comparative statics of $\tilde{p}$ is more straightforward and can be summarized as follows.

Lemma 7 *Comparative Statics of* $\tilde{p}$

$$\textit{(i)}\ \tfrac{d\tilde{p}}{d\phi} > 0,\ \textit{(ii)}\ \tfrac{d\tilde{p}}{d\psi} < 0,\ \textit{(iii)}\ \tfrac{d\tilde{p}}{dY} < 0\ \textit{and (iv)}\ \tfrac{d\tilde{p}}{d\gamma_{UA}} < 0.$$

Proof. Consider

$$\Pi(p) = p\phi - p\gamma_{UA}q^*\psi - C(p)$$

with $C(p) = v(p) + p\gamma_{UA}((1-q^*)Y + w(q^*))$. We use the first order condition

$$\frac{d\Pi}{dp} = \phi - \gamma_{UA}q^*\psi - v'(p) - \gamma_{UA}(Y(1-q^*) + w(q^*) = 0. \tag{6}$$

as an implicit function of $\tilde{p}$. We get

$$\frac{d\tilde{p}}{d\phi} = -\frac{1}{-v''(\tilde{p})} > 0,$$

$$\frac{d\tilde{p}}{d\psi} = -\frac{-\gamma_{UA}q^*}{-v''(\tilde{p})} < 0,$$

$$\frac{d\tilde{p}}{d\gamma_{UA}} = -\frac{-q^*\psi - (Y(1-q^*) + w(q^*))}{-v''(\tilde{p})} < 0,$$

$$\frac{d\tilde{p}}{dY} = -\frac{-\gamma_{UA}\psi\frac{dq^*}{dY} - \gamma_{UA}(1-q^*)}{-v''(\tilde{p})} < 0.$$

■

Lemma 7 demonstrates that comparative statics of the optimal effort level is straightforward if the truth-telling constraints do not bind. In this case, the optimal effort level is certainly increasing in project returns ϕ, and decreasing in conflict costs ψ. The probability of conflict γ_{UA} and the psychological costs of the agent also reduce profits because they increase expected costs of conflict *and* agency costs.

Welfare Analysis Finally, we want to comment on the welfare effects of self-esteem and conflict in our model. Recall that the agent is always left with rents of zero such that a welfare analysis amounts to a discussion of the principal's profits. We have already shown that the assumptions of our model ensure that $p_{FB} > 0$ and $\Pi_{FB} > 0$. Until now, we have identified two different sources for welfare-losses. First, implementable efforts do not have to exist (see Lemma 5) - in which case $p^* = 0$ and $\Pi(p^*) = 0$. This may happen because the bias of the relationship is too large or the ratio of conflict costs of the principal to psychological costs of the agent is too small. Second, even though implementable effort levels exist,

agency costs are substantial and may make it preferable for the principal not to hire the agent at all (see Proposition 1(ii)) - this is in particular the case for too small project returns. As the following result demonstrates, welfare losses are not restricted to the cases of $p^* = 0$ but are a common feature of all parameter configurations of our model.

Proposition 2 *Suppose p_{FB} is implementable*[8]

(i) $p^* = p_{FB}$ if and only if $\gamma_{UA} = 0$. Then, $\Pi_{FB} = \Pi(p^*)$.

(ii) Suppose that $\gamma_{UA} > 0$. Then, $p_{FB} > p^*$ and $\Pi_{FB} > \Pi(p^*)$.

Proof. Observe that $\Pi(p) = p\phi - p\gamma_{UA}q^*\psi - C(p)$ equals π_{FB} if $\gamma_{UA} = 0$. According to Lemma 7(iv), $\frac{d\tilde{p}}{d\gamma_{UA}} < 0$. This implies Part (i) and $p_{FB} > p^*$ whenever $\gamma_{UA} > 0$. $\Pi_{FB} > \Pi(p^*)$ follows from the strict optimality of p_{FB}. ■

A first best outcome can only be achieved with perfectly correlated signals (Part(i)). If the signals are imperfectly correlated, expected costs of conflict *and* agency costs (*i.e.,* the compensation of the agent's psychological costs) lead to optimal effort levels (and profits) strictly below the first best (Part(ii)).

4 The Evaluation Process

Until now, we have investigated optimal contract design for a *given* information technology. In reality, however, the quality of the evaluation process is to a large extent endogenously determined. The principal can, for example, decide how much time he spends on supervising the agent in the accomplishment of his project. He could *(i)* sit next to the agent during the whole project, or *(ii)* close the door to his office and only look at the result. Intuitively, the quality of the signal might be much better under the first evaluation procedure.[9] As a benchmark, Proposition 2 indicates that the principal should choose a perfectly correlated signal if p_{FB} can be implemented with a credible bonus. In general, however, the feasibility of a perfectly correlated signal and the implementability of p_{FB} cannot be taken for granted (for implementability see Lemma 3). This will be the starting point of the following investigation.

In this section, we assume that the principal not only fixes the terms of contract (*i.e.*, the bonus b and fixed payments f), but can also modify the information technology. To be specific, we follow MacLeod (2003) in parameterizing the conditional probabilities γ_{s_P,s_A} with the probability that the principal receives the

[8]Observe that there is always a conflict level ψ which guarantees this (see the discussion of Lemma 3).

[9]Note that we explicitly avoid terms like *control* and *(dis)trust* here (as e.g. used in Falk & Kosfeld (2006) and Ellingsen & Johannesson (2007). The choice of the quality of the evaluation procedure has an influence on how well the principal can observe an acceptable effort given that the project is a success. Therefore, the higher the quality of the principal's evaluation process, the higher the probability that the agent is rewarded in case of success. A higher quality is, hence, not regarded as negative by the agent.

signal $s_P = A$ denoted by g, the probability with which the agent has the same evaluation as the principal denoted by ρ and the probability with which the agent receives $s_A = A$ given that his signal is independent from the evaluation of the principal denoted by x [see also MacLeod (2003): 228]. Hence, g measures the quality of the principal's signal, ρ indicates the correlation between the agent's and the principal's signal - or the probability of an independent judgment - and x quantifies the quality of the agent's signal if he forms an independent judgment.

As an illustration for this parameterization consider an evaluation process for end-of-the-year bonuses for managers. The board fixes a checklist with elements of managerial performance which are assumed to be correlated with firm success. The stronger the relation between the points on the checklist and actual firm success, the better the boards (the principal's) signal (*i.e.* the larger g). For a given checklist, there is the probability ρ that the agent comes to the same judgment (about firm success) following the list. Finally, independent of the checklist, the agent has an independent judgment of his impact on firm success and the extent to which he deserves the bonus. While the latter cannot be influenced by the principal, it seems reasonable to assume that the design of the checklist can influence g (and perhaps ρ which will be discussed at the end of this section). In what follows, we assume that $\rho < 1$ and $x > 0$, and that the principal can costlessly choose g.

Perceived Bias of the Relationship Using g, ρ, and x, we get

$$\gamma_{AA} = g\left(\rho + (1-\rho)\,x\right) \text{ and } \gamma_{UA} = (1-g)\,(1-\rho)\,x.$$

Thus, the perceived bias of the relationship $\alpha = \frac{\gamma_{UA}}{\gamma_{AA}}$ is given by

$$\alpha = \frac{(1-g)}{g}\frac{(1-\rho)\,x}{(\rho + (1-\rho)\,x)}. \tag{7}$$

Eqn. (7) demonstrates that the principal can choose any α between 0 and ∞ with an appropriate choice of g. This implies that he can influence the agent's psychological costs $Y(\alpha)$ and, hence, the optimal conflict level $q^*(\alpha)$ as follows.

Lemma 8 *Comparative Statics w.r.t. g*

(i) $\frac{d\alpha}{dg} < 0$, $lim_{g\to 0}\alpha = \infty$ *and* $lim_{g\to 1}\alpha = 0$.

(ii) $\frac{dY(\alpha)}{dg} < 0$ *and* $lim_{g\to 1}Y(\alpha) > 0$.

(iii) $\frac{dq^*(\alpha)}{dg} \leq 0$ *and* $lim_{g\to 1}q^*(\alpha) > 0$.

(iv) $\frac{\partial C(p)}{\partial g} \leq 0$.

Proof. *(i)* follows directly from Eqn. (7). *(ii)* follows directly from *(i)* together with the fact that $\frac{dY(\alpha)}{d\alpha} > 0$ and $Y(\alpha) < 0$ for $\alpha = 0$. *(iii)* follows directly from *(ii)* together with the fact that $\frac{dq^*}{dY} \geq 0$ and teh properties of $w(q)$. *(iv)* follows from the definition of $C(p)$, Part (ii) and $\frac{d\gamma_{UA}}{dg} < 0$. ■

The lower the quality of the principal's signal, the higher the perceived bias of the relationship (Part(i)) and the larger the agent's psychological costs (Part(ii)). More intuitively, the lower the quality of the evaluation procedure that the principal uses to assess the agent's performance, the angrier the agent gets and the more he is willing to harm the principal whom he regards responsible for the choice of g.

Implementable Efforts As discussed in the previous section, an effort $p > 0$ is implementable for a given information technology if and only if the incentive compatible bonus $b(p)$ is between the maximal and minimal bonus b^{max} and b^{min} which can credibly guarantee truthtelling by the principal. The following Lemma displays the comparative statics of

$$b^{min} = \frac{(1-\rho)x}{(1-\rho x)} q^* \psi,$$

$$b^{max} = (\rho + (1-\rho)x) q^* \psi, \text{ and}$$

$$b(p) = \frac{1}{g}(v'(p) + (1-g)(1-\rho)x(Y(1-q^*) + w(q^*))).$$

with respect to g.

Lemma 9 *(i)* $\frac{db^{min}}{dg} \leq 0$ *and* $\lim_{g \to 0} b^{min} < \infty$.

(ii) $\frac{db^{max}}{dg} \leq 0$ *and* $\lim_{g \to 0} b^{max} < \infty$.

(iii) $\frac{db(p)}{dg} < 0$ *and* $\lim_{g \to 0} b(p) = \infty$.

Proof. (i) and (ii) follow directly from $\frac{dq^*}{dg} \leq 0$ and $q^* \leq 1$. (iii) is an immediate consequence of $\frac{dY}{dg} < 0$ and the definition of $b(p)$. ■

Lemma 8 and 9 determine the possible scenarios for implementability of an effort level $p > 0$ as depicted in Figure 4.

Lemma 10 *We can distinguish the following cases:*

(i) Case 1. $b(p) > b^{max}$ *for all* g.
Then, $p > 0$ *can not be implemented.*

(ii) Case 2. $b(p) \leq b^{max}$ *for some* $g < 1$ *but* $b(p) > b^{max}$ *for* $g = 1$.
Then, $p > 0$ *is implemented with the maximal* g *for which* $b(p) = b^{max}$.

(iii) Case 3. $b(p) \leq b^{max}$ *for some* $g < 1$ *but* $b(p) < b^{min}$ *for* $g = 1$.
Then, $p > 0$ *is implemented with the maximal* g *for which* $b(p) = b^{min}$.

(iv) Case 4. $b(p) \in [b^{min}, b^{max}]$ *for* $g = 1$.
Then, $p > 0$ *is implemented with* $b(p)$ *at* $g = 1$.

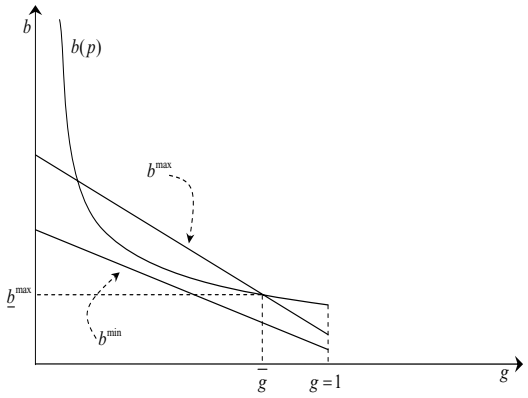

Figure 4: The Quality of the Evaluation Process: *Case 2.*

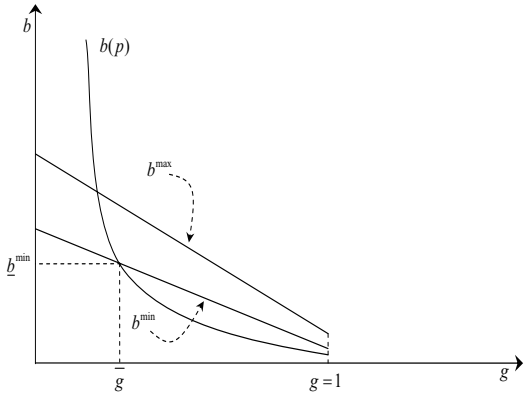

Figure 5: The Quality of the Evaluation Process: *Case 3.*

Proof. Follows from Lemma 8 and 9. ■

[Figures 4 and 5 here]

Figure 4 shows *Case 2.* in which $b(p) \leq b^{max}$ for some $g < 1$. The optimal bonus and signal quality is denoted $\underline{b}^{max}$ and $\overline{g}$. Figure 5, on the other hand, shows *Case 3.* in which $b(p) \leq b^{max}$ for some $g < 1$ but $b(p) < b^{min}$ for $g = 1$. In this case the optimal bonus and signal quality is respectively denoted as $\underline{b}^{min}$ and $\overline{g}$.[10] Figures 4 and 5 demonstrate that an effort level $p \in [p^{min}, p^{max}]$ is implementable (with an individually rational and incentive compatible contract and an appropriate signal quality g) if for any $g \in [0, 1]$ there exists a $b(p)$ such that $b(p) \leq b^{max}$. The principal implements p most efficiently, *i.e.* at lowest cost, by choosing the highest possible g (recall that $\frac{\partial C(p)}{\partial g} < 0$).

The following example shows that all cases can occur in our model.

Example 1 *Let* $w(q) = \frac{1}{2}q^2$, $v(p) = \frac{1}{2}p^2$, $Y = (0.1 + \alpha)$. *Then,* $q^* = (0.1 + \alpha)$ *if* $(0.1 + \alpha) \leq 1$ *and* $q^* = 1$ *otherwise. Moreover,* $p_{FB} = \phi$.

Fix, $\rho = 1/2$, $x = 1/2$, *and* $\phi = 1/2$.

- *Let* $\psi = 1$. *Then,* $b(p) > b^{max}$ *for all* g *(Case 1).*
- *Let* $\psi = 4$. *Then,* $b(p) \leq b^{max}$ *for some* $g < 1$ *and* $b(p) > b^{max}$ *for* $g = 1$ *(Case 2).*
- *Let* $\psi = 100$. *Then,* $b(p) \leq b^{max}$ *for some* $g < 1$ *and* $b(p) < b^{min}$ *for* $g = 1$ *(Case 3).*
- *Let* $\psi = 10$. *Then,* $b(p) \in [b^{min}, b^{max}]$ *for* $g = 1$ *(Case 4).*

As suggested by the example, *Case 1.* (*Case 3.*) will be the relevant description of implementability if the level of conflict ψ is sufficiently small (large) as the following result indicates.

Lemma 11 *Suppose* $p > 0$. *Then, Case 1 holds whenever* ψ *is sufficiently small and Case 3 holds, whenever* ψ *is sufficiently large.*

Proof. Follows directly from Lemma 3, Lemma 9(iii), and the fact that $b(p)$ does not depend on ψ. ■

Lemma 11 implies that the principal will not choose $g = 1$ to implement a certain effort level whenever ψ is too large. This explains the endogenous choice of an imperfect information technology by the principal even if the quality of the signal is costless.[11]

[10] We only focus on *Case 2.* and *3.* in the graphical representation for simplicity. Note that *Case 1.* and *4.* can likewise be analyzed in the same setting.

[11] To see that an effort level which is not implementable with $g = 1$ can indeed be optimal consider the case of very large conflict levels ψ. Then, $b^{min} > b(p)$ at $g = 1$ for all effort levels $p > 0$. If ϕ is sufficiently large (for instance larger than the level of conflict), a positive effort level will nonetheless be optimal.

Welfare Implications The previous paragraph demonstrated that if the principal can decide upon the quality of his own signal, the set of implementable efforts will be larger as compared to the situation in which the information technology was exogenously given. However, a certain effort level does not have to be implementable (Case 1), or is not implementable at $g = 1$ (Cases 2 and 3). This holds in particular for p_{FB} which leads to the following result.

Proposition 3 *Let $\rho < 1$ and suppose Case 1, 2 or 3 describes implementability of p_{FB}.*[12] *Then, $p^* < p_{FB}$ and $\Pi(p^*) < \Pi_{FB}$.*

Proof. Consider Case 1. As p_{FB} is not implementable, $p^* < p_{FB}$ (possibly $p^* = 0$) will be implemented and $\Pi(p^*) < \Pi_{FB}$ follows from the unique optimality of p_{FB}.

Consider Case 2 and 3. Then p_{FB} can not be implemented with $g = 1$. Hence, marginal costs of effort implementation $C'(p_{FB}) > v'(p_{FB})$ which implies $p^* < p_{FB}$ and thereby $\Pi(p^*) < \Pi_{FB}$. ■

According to Proposition 3, the first best solution will not necessarily be implemented by the principal even if he can choose any signal quality at zero costs. As indicated by Lemma 11 this will be the case for instance whenever conflict level ψ is above a certain threshold.

How do these results translate if the principal can choose correlation ρ? First of all, an information technology with perfectly correlated signals (*i.e.,* $\rho = 1$) becomes feasible. For $\rho = 1$, $b^{min} = 0$, $b^{max} = q^*\psi$, and $b(p) = \frac{v'(p)}{g}$. Suppose for the moment that g is fixed. In this case, the principal will implement p_{FB} whenever $b(p_{FB}) \leq b^{max}$. Otherwise (and this will be the case if ψ is sufficiently small) he will - analogously to Proposition 3 - implement a lower effort level ($p^* < p_{FB}$) which indicates a welfare loss. This argument does not change if the principal can choose both g and ρ, because $b^{max} = q^*\psi$ can still be smaller than $b(p) = v'(p_{FB})$ which is *e.g.,* the case whenever $\psi < \phi$. *I.e.*, if the return of the project is bigger than the level of conflict, the principal will not implement a first best effort level - even if he can decide upon the quality of his own signal and the correlation of signals.

5 Concluding Remarks

The analysis of our model revealed that self-esteem and the individual's eagerness to protect it may facilitate principal-agent relationships even if performance signals are subjective and no third-party can enforce truth-telling. However, only if signals are perfectly correlated, a first best can be achieved - even if the principal can costlessly choose the quality of his own signal. For imperfectly correlated signals, positive effort levels will be implemented by the principal if profits and costs of conflict are sufficiently large. As an incentive compatible contract has to compensate the agent for effort costs *and* expected psychological costs, the implemented effort level will be below the first best effort.

[12] Example 1 and Lemma 11 show that this holds true for an appropriate choice of ψ.

This qualifies to some extent the results in MacLeod (2003) which claim the existence of implementable effort levels regardless of the details of the relationship. The positive result of MacLeod (2003) is crucially depending on the credibility and flexibility of the third-party payments. While in his model, every payment to a third-party was a credible promise, the specific nature of conflict in our setting imposes tighter constraints on the set of feasible contracts. Moreover, following the interpretation of third-party payments as endogenous costs of conflict [see MacLeod (2003), p.229], our analysis demonstrates that the feasibility of welfare-optimal solutions in MacLeod (2003) crucially hinges on the fact that conflicts do not impose any costs on the agent. If - as in our model - conflicts entail some costs for the agent, the need to compensate for these costs raises agency costs above the first best level and prevents welfare optimal solutions even if the truth-telling problem is not an obstacle.

In the extended version of the model we assume that the principal has control over the choice of his evaluation procedure. More precisely, we assume that the principal does not only have to choose the optimal compensation scheme, but can also choose among different evaluation procedures that differ in the quality of their subjective performance signal. In particular, the agent's psychological costs increase in the bias of the information technology. This resembles a case of procedural concerns as conceptualized by Sebald (2007) in a general class of models with belief-dependent utility. Interestingly, our model shows that it may be optimal for the principal to choose a procedure which is not minimizing the agent's psychological costs - even if it is costless to do so - but rather facilitates the credible implementation of a positive effort level.

Our assumptions on psychological costs are rather ad-hoc. We simply formalized the results from the literature in social psychology in a straightforward functional form. We opted for this approach as the main purpose of this paper is the discussion of promoting and inhibiting factors for principal-agent relationships in which neither effort nor output can be measured objectively.

Furthermore, we have chosen to model the agent as risk-neutral and with unlimited liability. While this obviously promoted expositional ease, it focuses on the special case of a principal-agent relationship which never leaves a rent to the agent. In case of risk-averse or limitedly liable agents, a non-trivial dependence of the agent's rents on his sensitivity to ego-threats and the quality of the information technology is to be expected and definitely worth an investigation. Our results with respect to break-downs of the relationship are, however, not expected to depend on these assumptions.

Finally, it is known since long [see Malcomson (1984)] that the problem of non-enforceable contracts in the presence of subjective performance measures is easily solved if the principal has to deal with a team of agents and can pay them according to a ranking with pre-committed payments for each rank. If agents do not suffer from psychological costs in these kind of tournaments, a first best can be achieved and performance pay as characterized in this paper is always inferior. However, it is an empirical question whether tournaments actually lead to lower psychological costs. If self-esteem is threatened fiercely by the explicit

announcement that someone-else is better, incentive compatible payments in the tournament have to compensate the corresponding expected psychological costs. This may well lead to an inferiority of such a scheme and promote performance pay as discussed in our paper, where self-esteem is not threatened by a relative performance measure but by an absolute evaluation. In this respect, new laboratory experiments could shed some light on the optimal design of payment schemes in the case of subjective performance evaluation.

6 References

1. Baumeister, R. (2005), *Self-Concept, Self-esteem and Identity*, in Derlega, J., Jones, W. & Winstead, B. (2005), *Personality: Contemporary Theory and Research*, Wadsworth Publishing Company, 3rd Edition.

2. Baumeister, R., Smart, L. & Boden, J. (1996), *Relation of threatened egotism to violence and aggression: The dark side of high self-esteem*, Psychological Review, 103, 5-33.

3. Baird, L. (1977), *Self and Superior Ratings of Performance: As Related to Self-Esteem and Satisfaction with Supervision*, The Academy of Management Journal, 20(2), 291-300.

4. Bénabou, R. & Tirole, J. (2002), *Self-Confidence And Personal Motivation*, The Quarterly Journal of Economics, 117(3), 871-915.

5. Bureau of National Affairs (1981), *Wage and Salary Administration*, PPF Survey No. 131, Washington, D.C.

6. Bushman, B. & Baumeister, R. (1998), *Threatened Egotism, Narcissism, Self-Esteem, and Direct and Displaced Aggression: Does Self-Love or Self-Hate Lead to Violence?*, Journal of Personality and Social Psychology, 75(1), 219-229.

7. Compte, O. & Postlewaite, A. (2004), *Confidence-Enhanced Performance*, American Economic Review, 94(5), 1536-1557.

8. Dufwenberg, M. & Kirchsteiger, G. (2004), *A theory of sequential reciprocity*, Games and Economic Behavior, 47, 268-298.

9. Ellingsen, T. & Johannesson, M. (2007), *Pride and Prejudice: The Human Side of Incentive Theory*, forthcoming in American Economic Review.

10. Falk, A. & Fischbacher, U. (2006), *A theory of reciprocity*, Games and Economic Behavior, 54(2), 293-315.

11. Greenwald, A. (1980), *The totalitarian ego: Fabrication and revision of personal history*, American Psychologist, 35, 603-618.

12. Korman, A. (1969), *Toward a Hypothesis of Work Behavior*, Journal of Applied Psychology, 54, 31-41.

13. Köszegi, B. (2006), *Ego Utility, Overconfidence, and Task Choice*, Journal of the European Economic Association, 4(4), 673-707.

14. Levine, J. (2003), *Relational Incentive Contracts*, American Economic Review, 93(3), 835-857.

15. MacLeod, B. (2003) *Optimal Contracting with Subjective Evaluation*, American Economic Review, 93(1), 216-240.

16. Malcomson, J. (1984), *Work Incentives, Hierarchy, and Internal Labor Markets*, Journal of Political Economy, 92(3), 486-507.

17. Milkovich, G. & Wigdor, A. (1991), *Pay for Performance: Evaluating Performance Appraisal and Merit Pay*, eds, National Academy Press, Washington, D.C.

18. Raskin, R., Novacek, J. & Hogan, R. (1991), *Narcissism, self-esteem, and defensive self-enhancement*, Journal of Personality, 59(1) 19-38.

19. Rabin, M. (1993), *Incorporating fairness into game theory and economics*, American Economic Review, 83, 12811302.

20. Sebald, A. (2007), *Procedural Concerns in Psychological Games*, ECORE Discussion Paper, 2007/62, Brussels, Belgium.

21. Shrauger, J. & Lund, A. (1975), *Self-evaluation and reactions to evaluations from others*, Journal of Personality. 43, 94-108.

22. Steelman, L. & Rutkowski, K. (2004), *Moderators of employee reactions to negative feedback*, Journal of Managerial Psychology, 19(1), 6-18.

Appendix

Proof of Lemma 2: To save on notation, we denote $Y(t_P = k, s_A = l, \alpha)(1 - q^*) + w(q^*) \equiv Y_{kl}$ throughout this proof.

Part (i). Without loss of generality, suppose that Γ is a revelation contract, *i.e.,* the principal and the agent tell the truth under contract Γ. As Γ implements $p > 0$, the incentive compatibility constraint

$$\Sigma_{k \in S_P, l \in S_A}(Y_{kl} + c_{kl}) \frac{dPr\{s_P = k, s_A = l\}}{dp} = v'(p)$$

is satisfied. Consider a contract $\hat{\Gamma}$ which fixes payments of $\hat{c}_k = \sum_{l \in S_A} c_{kl} Pr\{s_P = k, s_A = l\}$ if the principal receives signal $s_P = k$, i.e., payments are independent of s_A. These payments also satisfy the incentive compatibility constraint (see

above).[13] Moreover, the agent always tells the truth due to indifference. Finally, the principal's truth-telling constraint is also satisfied under $\hat{\Gamma}$. To see this observe that the principal reports k given that he has received k under contract Γ if

$$\begin{aligned} &Pr\{s_A = A|s_P = k\}(c_{oA} - c_{kA}) + Pr\{s_A = U|s_P = k\}(c_{oU} - c_{kU}) \\ &\geq Pr\{s_A = A|s_P = k\}((q^*\psi)_{kA} - (q^*\psi)_{oA}) \\ &+Pr\{s_A = U|s_P = k\}((q^*\psi)_{kU} - (q^*\psi)_{oU}) \end{aligned} \tag{8}$$

for all $o \in S_P$ (where $(q^*\psi)_{t_A,t_P}$ denotes the anticipated conflict costs for a reported configuration (t_A, t_P)). This set of inequalities holds because Γ implements truth-telling by assumption. $\hat{\Gamma}$ implements truth-telling if

$$\begin{aligned} \hat{c}_o - \hat{c}_k \geq \quad & Pr\{s_A = A|s_P = k\}((q^*\psi)_{kA} - (q^*\psi)_{oA}) \\ & +Pr\{s_A = U|s_P = k\}((q^*\psi)_{kU} - (q^*\psi)_{oU}). \end{aligned} \tag{9}$$

holds for all $o, k \in S_P$. Inserting $\hat{c}_k$ and $\hat{c}_o$ yields

$$\begin{aligned} &Pr\{s_A = A|s_P = k\}(c_{oA} - c_{kA}) + Pr\{s_A = U|s_P = k\}(c_{oU} - c_{kU}) \\ &\geq Pr\{s_A = A|s_P = k\}((q^*\psi)_{kA} \\ &-(q^*\psi)_{oA}) + Pr\{s_A = U|s_P = k\}((q^*\psi)_{kU} - (q^*\psi)_{oU}). \end{aligned}$$

which coincides with System 8 and therefore shows that for $\hat{\Gamma}$ the principal's truthtelling constraint is satisfied as well. Hence, any revelation contract Γ can be substituted by a revelation contract $\hat{\Gamma}$ with c_{kl} independent of l which also implements $p > 0$ and leaves the principal weakly better off.

Part (ii). Suppose by contradiction that Γ implements $p > 0$ with $c_A = g$ and $c_U = g + \epsilon$ with $\epsilon \geq 0$. Then, the incentive compatibility constraint of the agent can be written as

$$\epsilon = \frac{v'(p) - \gamma_{UA} Y_{UA}}{(\gamma_{UA} + \gamma_{UU} - 1)}. \tag{10}$$

Observe that the numerator of the rhs is positive for every $p > 0$ and vanishes for $p = 0$ while the denominator is negative. Hence, the rhs is negative and the incentive compatibility constraint is not satisfied for any $p > 0$ and $\epsilon > 0$. For $\epsilon = 0$, the incentive compatibility constraint is solved by $p = 0$. A contradiction. ■

[13] Individual rationality is trivially fulfilled as expected payments for the agent are the same under Γ and $\hat{\Gamma}$.

Short Curriculum Vitae

Alexander Sebald was born on October 28, 1976 in Bergisch Gladbach, Germany. In 1996 Alexander finished high school (Gymnasium) and started his national service in the German Navy. From 1997 till 2002 he studied International Economic Studies at Maastricht University. During that time Alexander also spend a year (2000 - 2001) in Milan, Italy, participating in an Erasmus exchange at the Universita Commerciale Luigi Bocconi and working as an intern at McKinsey & Company. After his graduation from Maastricht University in 2002 and a long voyage to India, China and South East Asia, Alexander started as a PhD-student at the Department of Economics of Maastricht University in 2003. Without leaving Maastricht, Alexander followed his supervisor Georg Kirchsteiger to the European Center for Advanced Research in Economic and Statistics at the Université Libre de Bruxelles in 2004.

His research interests are: Behavioral Economics, Psychological Game Theory, Experimental Economics and Contract Theory.