Supplement to "Rewards and punishments: Informal contracting through social preferences"

(*Theoretical Economics*, Vol. 11, No. 3, September 2016, 1145–1179)

Sylvain Chassang
Department of Economics, New York University

Christian Zehnder
Department of Organizational Behavior, University of Lausanne

Appendix S.A: Extensions

This appendix presents the following extensions: we clarify the importance of capturing betrayal aversion (Bohnet and Zeckhauser 2004, Bohnet et al. 2008) in our model of social preferences; we establish the robustness of our results to small perturbations in social preferences and in the cost of transfers; finally, we outline a simple model of endogenously incomplete contracts using the model of informal justice developed in this paper as a building block.

S.A.1 *Alternative social preferences*

The principal's social preferences play a key role in our analysis. One central assumption is that the principal treats exogenous uncertainty over payoffs conditional on actions differently from endogenous uncertainty deriving from mixing by players: we assume that the principal evaluates the fairness of every relationship between a player A and a player P independently. This allows us to capture a form of betrayal aversion documented by Bohnet and Zeckhauser (2004) and Bohnet et al. (2008).

This section shows that this modeling choice is essential for informal justice to take into account payoff-irrelevant signals x that are informative of player A's behavior, i.e., for informal justice to depend on assessments of intents. Take $\pi \in \Delta(\{C, D\})$ as given. We now assume that the jury chooses a transfer function T that solves the optimization problem

$$\max_{T \in [-T_{\max}, T_{\max}]^Z} \widehat{V}(\pi, T) \equiv \delta \mathbb{E}[\Phi(u^T)|\pi] + (1-\delta)\Phi(\mathbb{E}[u^T|\pi]),$$

where uncertainty over behavior has been folded into uncertainty over outcomes. In other terms, the principal evaluates fairness at the population level rather than relationship by relationship. We now show that the corresponding transfer function does not

Sylvain Chassang: chassang@nyu.edu
Christian Zehnder: christian.zehnder@unil.ch

DOI: 10.3982/TE2063

depend on side information. Indeed, given a candidate transfer function T, define

$$T^U(u) \equiv \int_Z T_z f_\pi(z|u)\,\mathrm{d}z.$$

Transfer scheme T^U is the expectation of transfer T conditional on payoff outcome $u = (u_A, u_P)$. For any T, we have that

$$\widehat{V}(\pi, T) = -\lambda \int_Z |T_z| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda)T_z| f_\pi(z)\,\mathrm{d}z \\ - (1-\delta)\alpha \left| \int_Z (\Delta u_z - (2+\lambda)T_z) f_\pi(z)\,\mathrm{d}z \right|.$$

By convexity of $|\cdot|$ and Jensen's inequality, we obtain that

$$\widehat{V}(\pi, T) \leq -\lambda \int_Z |T^U| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u - (2+\lambda)T^U| f_\pi(z)\,\mathrm{d}z \\ - (1-\delta)\alpha \left| \int_Z (\Delta u - (2+\lambda)T^U) f_\pi(z)\,\mathrm{d}z \right| \\ \leq \widehat{V}(\pi, T_{|U}).$$

It follows that informal incentives derived from value function $\widehat{V}$ need only depend on payoff outcome u. In this model, the principal cares only about average inequality and does not care about whether she is punishing a player A who took selfish action D. As a result, transfers never depend on side information x.

S.A.2 *Robustness*

Some of our modeling choices, such as the use of linear inequality-averse preferences à la Fehr and Schmidt (1999) or the use of linear transfer costs $-\lambda|T_z|$, make the analysis tractable but induce corner solutions.

We show that our analysis is in fact robust to small perturbations in the environment. Let $c(T) = \lambda|T|$ denote the reference deadweight cost of transfers paid by the transferring party. We consider sequences of social preferences $\Phi_n(\Sigma u, \Delta u)$ and transfer cost functions c_n such that $\lim_{n\to\infty} \|\Phi_n - \Phi\|_\infty = \lim_{n\to\infty} \|c_n - c\|_\infty = 0$, where $\|\cdot\|_\infty$ denotes the uniform norm. We denote by T_n^π the transfer scheme solving

$$\max_T V_n(\pi, T) \equiv \delta\mathbb{E}[\Phi_n(\Sigma u^T, \Delta u^T)|\pi] + (1-\delta) \sum_{a\in\{C,D\}} \pi(a)\Phi_n(\mathbb{E}[\Sigma u^T|a], \mathbb{E}[\Delta u^T|a]).$$

Lemma S.A.1 (Continuity). *Consider any compact set Π included in the interior of $\Delta(\{C, D\})$. Uniformly over $\pi \in \Pi$, transfer schemes $(T_n^\pi)_{n\geq 0}$ converge to T^π under the L^1 norm:*

$$\lim_{n\to\infty} \sup_{\pi\in\Pi} \int |T_n^\pi - T^\pi|\,\mathrm{d}z = 0.$$

Proof. The difficulty here is that we are working with an infinite set of states $z \in Z$, so that the space of possible transfer functions is infinite dimensional and, therefore, not compact under the L_1 norm. Indeed, if instead we were working with a finite set of states Z, Lemma S.A.1 would follow immediately from the uniqueness of optimal transfers and Berge's theorem of the maximum. Proving an extension is possible in our case but requires some work.

The proposed proof is by contradiction. Assume that there exists $\epsilon > 0$ and a sequence $(\pi_n)_{n\in\mathbb{N}}$ such that for all $n \geq 0$,

$$\|T^{\pi_n} - T_n^{\pi_n}\|_1 > 2\epsilon.$$

By compactness of Π, we can assume that the sequence $(\pi_n)_{n\in\mathbb{N}}$ converges to $\pi_\infty \in \Pi$. In addition, we know from Lemma 4 that T^π is continuous in π under the L^1 norm. Hence, up to extraction of a subsequence, we can assume that

$$\|T^{\pi_\infty} - T_n^{\pi_n}\|_1 > \epsilon. \tag{S.1}$$

For concision, we denote $T_n = T_n^{\pi_n}$. It is immediate that $V_n(\pi_n, T)$ converges uniformly over T to $V(\pi_\infty, T)$. Since T_n solves $\max_T V_n(\pi_n, T)$, we obtain that $V(\pi_\infty, T_n)$ converges to $V(\pi_\infty, T^{\pi_\infty})$ as n grows large. Given that $\max_T V(\pi_\infty, T)$ has a unique solution, it is reasonable to expect that this result and (S.1) should lead to a contradiction. The only difficulty is that $(T_n)_{n\geq 0}$ need not have a converging subsequence under the L_1 norm.

Consider the sequence of expected inequality $(\mathbb{E}[\Delta u^{T_n}|C], \mathbb{E}[\Delta u^{T_n}|D])_{n\geq 0}$ under transfer schemes $(T_n)_{n\geq 0}$. Up to extraction of a subsequence, we can assume that this sequence converges to values (Δ_C, Δ_D). Consider first the case where $(\Delta_C, \Delta_D) \neq (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$. For any $\nu > 0$, let $\widehat{T}_\nu$ denote solutions to

$$\max_T \mathbb{E}[-\lambda|T_z| - \alpha\delta|\Delta u_z - (2+\lambda)T_z||\pi] \quad \left| \quad \begin{array}{l} \mathbb{E}[\Delta u^T|C] \in [\Delta_C - \nu, \Delta_C + \nu] \\ \mathbb{E}[\Delta u^T|D] \in [\Delta_D - \nu, \Delta_D + \nu]. \end{array} \right. \tag{S.2}$$

The set of such solutions, parameterized by $(\Delta_C, \Delta_D, \nu)$, is compact under the L_1 norm.[1] Take ν to 0 and consider a sequence of solutions (S.2) converging to a limit transfer scheme T^∞ under the L_1 norm. The fact that T^∞ solves $\max_T V(\pi_\infty, T)$ and the fact that

$$(\mathbb{E}[\Delta u^{T^\infty}|C], \mathbb{E}[\Delta u^{T^\infty}|D]) \neq (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$$

contradict the fact that $\max_T V(\pi_\infty, T)$ has a unique maximizer.

Consider now the case where $(\Delta_C, \Delta_D) = (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$. For any $\nu > 0$, consider solutions $\widehat{T}_\nu$ to

$$\max_T \mathbb{E}[-\lambda|T_z| - \alpha\delta|\Delta u_z - (2+\lambda)T_z||\pi] \quad \left| \quad \begin{array}{l} \mathbb{E}[\Delta u^T|C] \in [\Delta_C - \nu, \Delta_C + \nu] \\ \mathbb{E}[\Delta u^T|D] \in [\Delta_D - \nu, \Delta_D + \nu]. \end{array} \right. \tag{S.3}$$

[1] They take a threshold form as in Proposition 1, and using the fact that the log-likelihood ratio $\log(f(z|D)/f(z|C))$ admits a density (Assumption 1), convergence of the thresholds implies convergence in the L_1 sense.

The set of such solutions is compact and using the fact that $\max V(\pi_\infty, T)$ has a unique solution, it must be that as ν goes to 0, $\widehat{T}_\nu$ converges to T^{π_∞} under the L_1 norm. Consider the Lagrangian $L(z, T_z)$ corresponding to (S.3). It can be written in the form

$$L(z, T_z) = -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| + \widehat{\mu}_D^\nu \pi(D|z) + \widehat{\mu}_C^\nu \pi(C|z) + L_0, \tag{S.4}$$

where L_0 is a constant.

Let μ_C^∞ and μ_D^∞ denote the Lagrangian multipliers associated with problem $\max_T V(\pi_\infty, T)$, as described by Lemma 3. Given that $\widehat{T}^\nu$ must converge to T^{π_∞} under the L_1 norm, it must be that maximizers of Lagrangian (S.4) converge to maximizers of Lagrangian (11) (see Appendix A.2). Hence it must be that $\lim_{\nu\to 0} \mu_C^\nu = (1-\delta)\alpha(2+\lambda) + \mu_C^\infty$ and $\lim_{\nu\to 0} \mu_C^\nu = (1-\delta)\alpha(2+\lambda) - \mu_D^\infty$.

For any value $\nu > 0$, for n large enough, T_n satisfies the constraints in (S.3). We obtain that, by construction,

$$0 \le \mathbb{E}[L(z, \widehat{T}_z^\nu)|\pi_\infty] - \mathbb{E}[L(z, T_{n,z})|\pi_\infty] \le V(\pi_\infty, T^{\pi_\infty}) - V(\pi_\infty, T_n) + 2\nu.$$

There exists a function $\rho_z > 0$ such that for ν small enough, for a.e. z, $L(z, T_{\nu,z}) - L(z, T_{n,z}) \ge \rho_z|T_{\nu,z} - T_{n,z}|$. Furthermore for any $\overline{\eta} > 0$, there exists $\eta \in (0, \overline{\eta})$ such that $\mathcal{L}(z \text{ s.t. } \rho_z \le \eta) \le \eta$. Pick $\eta < \epsilon/(4T_{\max})$. We have that for all n and ν,

$$\begin{aligned} V(\pi_\infty, T^{\pi_\infty}) - V(\pi_\infty, T_n) &\ge -2\nu + \int_Z \rho_z|T_\nu - T_n| f_\pi(z)\,\mathrm{d}z \\ &\ge -\nu + \eta\underline{h}\int_Z |T_\nu - T_n|(1 - \mathbf{1}_{\rho_z<\eta})\,\mathrm{d}z \\ &\ge -\nu + \eta\underline{h}(\epsilon - 2\eta T_{\max}). \end{aligned}$$

Since this holds for ν arbitrarily close to 0, we obtain that the sequence $V(\pi_\infty, T_n)$ remains bounded strictly below $V(\pi_\infty, T^{\pi_\infty})$ even as n grows large—a contradiction.

Hence T_n^π converges to T^π uniformly over $\pi \in \Pi$ under the L_1 norm. □

Since player A's expected payoffs from different actions are continuous in T^π, this implies that any sequence of equilibria $(\pi_n, T^{\pi_n})_{n\ge 0}$ of perturbed games admits a subsequence that converges to an equilibrium of the unperturbed game. Inversely, assume that (π_0, T^{π_0}) is an equilibrium of the unperturbed game such that $\mathbb{E}[u_A^{T^\pi}|C] - \mathbb{E}[u_A^{T^\pi}|D]$ is either nonzero at π_0 or changes sign around π_0. Then there will be a sequence of equilibria (π_n, T^{π_n}) of perturbed games converging to (π_0, T^{π_0}). In this sense, our analysis is robust to small perturbations in the principal's preferences and in the cost of transfers.

S.A.3 *A model of endogenous incompleteness*

This paper develops a model of informal contracting when punishments and rewards are not determined by an ex ante optimal contract, but rather are taken ex post and express the moral sentiment of the principal. An important complementary research agenda would be to endogenize whether incentive schemes will be determined ex ante or ex post. Following work by Dye (1985) and Tirole (2009) we briefly outline a simple ad hoc model of boundedly rational contracting in which the trade-off between ex ante and ex post contracting can be expressed.

Consider the problem of a senior executive overseeing two managers. There are three periods $t \in \{0, 1, 2\}$. At $t = 0$, the executive has the possibility to commit to transfers as a function of observables. At time $t = 1$, a particular environment $\theta \in \Theta$ is selected and becomes common knowledge among players. An environment θ corresponds to both a selection of which manager is the active or the passive player and a specification of the set of outcomes and their distribution (Z^θ, f^θ). For simplicity we assume that all states $\theta \in \Theta$ occur with the same probability $1/\operatorname{card}\Theta$. In period $t = 0$, the senior executive can choose the environments θ for which he wants to commit to an ex ante contract and the environments for which he will determine rewards and punishments ex post. We denote by $\chi(\theta) \in \{0, 1\}$ the executive's decision to specify ex ante a contract conditional on environment θ. This comes at a consideration cost k for each environment in which an ex ante contract is specified.

If an ex ante contract is specified conditional on state θ, then the executive obtains an expected payoff $V^{\text{ex ante}}(\theta)$. In states θ where no ex ante contract is specified, transfers are determined by an equilibrium of the informal justice game studied in this paper. This results in payoffs $V^{\text{ex post}}(\theta)$. Altogether the senior executive's contract completion decision $\chi(\cdot)$ is chosen to maximize

$$-k \sum_{\theta \in \Theta} \chi(\theta) + \frac{1}{\operatorname{card}\Theta} \sum_{\theta \in \Theta} \chi(\theta)[V^{\text{ex ante}}(\theta) - V^{\text{ex post}}(\theta)].$$

A key aspect of this trade-off is that consideration costs are paid regardless of which state happens. As a result, the senior executive will choose to leave contracts incomplete when the set of relevant environments is large and when the payoffs of informal justice approach those of ex ante contracts, for instance, when justice is intent-based and the information available ex post is sufficiently good. Contracts will be completed at states that are likely to happen and for which informal justice is poorly suited to incentivize good behavior (say negative externality environments with poor ex post information).

Appendix S.B: Proofs

S.B.1 *Proofs for Sections 2 and 3*

Proof of Lemma 1. Let us begin with point (i). Values $V(a, T)$ obtainable when implementing action $a = D$ are bounded above by $V(D, 0)$. Consider the transfer scheme defined $\forall z \in \{-1, 1\}$ by $T_z = -6(\gamma + z)/(2 + \lambda)$. Conditional on actions C and D, payoffs to player A under this transfer scheme are

$$\mathbb{E}[u_A^T|C] = \frac{1}{2}\left[4\left(\frac{1}{2} + \gamma\right) - \frac{3\lambda}{2 + \lambda}(2 + \gamma)\right]$$

$$\mathbb{E}[u_A^T|D] = \frac{1}{2}\left[4\left(-\frac{1}{2} + \gamma\right) - \frac{3\lambda}{2 + \lambda}(2 - \gamma)\right].$$

It follows that $\mathbb{E}[u_A^T|C] - \mathbb{E}[u_A^T|D] = 2 - (3\lambda)/(2+\lambda)\gamma > 0$. Therefore, transfer scheme T implements action C and guarantees that there is no difference in expected payoffs across players. The principal's value for implementing action C rather than action D (by an optimal transfer scheme) is bounded below by

$$\begin{aligned} V(C,T) - V(D,0) &= 4\left(\frac{1}{2}+\gamma\right) - \frac{3\lambda}{2+\lambda}(2+\gamma) - 4\left(-\frac{1}{2}+\gamma\right) \\ &\geq 4 - \frac{3\lambda}{2+\lambda}\frac{5}{2} > 0, \end{aligned}$$

where we used the assumption that $\lambda \in (0,2)$. Hence it is always optimal to choose a contract that implements action C.

We now turn to point (ii) and set $\gamma = -\frac{1}{2}$. We know from point (i) that it is optimal to implement action C. The optimal contracting problem boils down to

$$\begin{aligned} &\max_{T_{-1},T_1} -\frac{\lambda}{4}|T_{-1}| - \frac{3\lambda}{4}|T_1| - \alpha(2+\lambda)\left|\frac{1}{4}T_{-1} + \frac{3}{4}T_1\right| \\ &T_{-1}, T_1 \quad \text{s.t.} \quad \mathbb{E}[u_A|C] - \mathbb{E}[u_A|D] + \frac{1}{2}(T_{-1} + \lambda T_{-1}^+) - \frac{1}{2}(T_1 + \lambda T_1^+) \geq 0. \end{aligned}$$

We only need to show that setting $T_{-1} \leq 0$ cannot be optimal. If this were the case, player A's incentive compatibility (IC) constraint implies that $T_1 < 0$. Consider increasing T_1 and T_{-1} by $\Delta > 0$. For Δ small, player A's IC constraint continues to hold and the principal's payoff increases. It follows that the optimal contract must involve setting $T_{-1} > 0$. □

S.B.2 *Proofs for Section 5*

Proof of Lemma 4. Consider a sequence $(\pi_n, f_n)_{n\geq 0}$ converging to (π, f) under the L_1 norm. For concision, let $T^n \equiv T_{f_n}^{\pi_n}$ denote the corresponding transfer scheme. Assume that there exists ϵ such that for all $n \geq 0$, $\|T_f^\pi - T^n\|_1 \geq \epsilon$. We show that this leads to a contradiction.

We know that transfer scheme T^n can be written to take the form

$$T_z^n = \begin{cases} 0 & \text{if } f_n(z|D)/f_n(z|C) \in (\theta_{-,n}^\Delta, \theta_{+,n}^\Delta) \\ -T_{\max} & \text{if } f_n(z|D)/f_n(z|C) < \theta_{-,n}^{\max} \\ T_{\max} & \text{if } f_n(z|D)/f_n(z|C) > \theta_{+,n}^{\max} \\ \Delta u_z^+/(2+\lambda) & \text{if } f_n(z|D)/f_n(z|C) \in (\theta_{+,n}^\Delta, \theta_{+,n}^{\max}) \\ -\Delta u_z^-/(2+\lambda) & \text{if } f_n(z|D)/f_n(z|C) \in (\theta_{-,n}^{\max}, \theta_{-,n}^\Delta). \end{cases}$$

Up to extraction of a subsequence, we can assume that thresholds $(\theta_{-,n}^{\max}, \theta_{-,n}^\Delta, \theta_{+,n}^\Delta, \theta_{+,n}^{\max})$ converge to thresholds $(\theta_{-,\infty}^{\max}, \theta_{-,\infty}^\Delta, \theta_{+,\infty}^\Delta, \theta_{+,\infty}^{\max})$. As a result transfers T^n must converge under the L_1 norm to transfer function

$$T_z^\infty = \begin{cases} 0 & \text{if } f(z|D)/f(z|C) \in (\theta^\Delta_{-,\infty}, \theta^\Delta_{+,\infty}) \\ -T_{\max} & \text{if } f(z|D)/f(z|C) < \theta^{\max}_{-,\infty} \\ T_{\max} & \text{if } f(z|D)/f(z|C) > \theta^{\max}_{+,\infty} \\ \Delta u_z^+/(2+\lambda) & \text{if } f(z|D)/f(z|C) \in (\theta^\Delta_{+,\infty}, \theta^{\max}_{+,\infty}) \\ -\Delta u_z^-/(2+\lambda) & \text{if } f(z|D)/f(z|C) \in (\theta^{\max}_{-,\infty}, \theta^\Delta_{-,\infty}). \end{cases}$$

Indeed, this follows from the fact that $\forall \nu > 0$,

$$\begin{aligned} &\mathcal{L}\left(\left|\frac{f_n(z|D)}{f_n(z|C)} - \frac{f(z|D)}{f(z|C)}\right| > \nu\right) \\ &\qquad \le \frac{1}{\nu}\int_Z \left|\frac{f_n(z|D)}{f_n(z|C)} - \frac{f(z|D)}{f(z|C)}\right| \mathrm{d}z \\ &\qquad \le \frac{1}{\nu}\int_Z \left|\frac{f(z|C)[f_n(z|D) - f(z|D)] + f(z|D)[f_n(z|C) - f(z|C)]}{f_n(z|C)f(z|C)}\right| \mathrm{d}z \\ &\qquad \le \frac{1}{\nu \underline{h}}(\|f_n(\cdot|D) - f(\cdot|D)\|_1 + K\|f_n(\cdot|C) - f(\cdot|C)\|_1) \\ &\qquad \to 0 \quad (\text{as } n \to \infty). \end{aligned}$$

Necessarily, we have that $\|T^\infty - T_f^\pi\|_1 \ge \epsilon$. However, since $V_f(\pi, T)$ is continuous in f, π, and T, we obtain that T^∞ must solve $\max_T V_f(\pi, T)$. This contradicts the fact that T_f^π is the unique solution to $\max_T V_f(\pi, T)$. Hence, it must be that $T_{f_n}^{\pi_n}$ converges to T_f^π under the L_1 norm. □

The following lemma provides sufficient conditions for intent-based justice to exhibit punitive justice.

Lemma S.B.1. *For any fixed $\eta > 0$, as the weight $1 - \delta$ on ex ante fairness approaches 1, all equilibria with $\pi(C) > \eta$ are such that there is punitive justice, i.e., states z such that $T_z > \Delta u_z^+/(2+\lambda)$.*

Proof. As a preliminary step, we characterize the limit of transfer schemes T_δ^π (where we temporarily emphasize dependency on δ), for any π in the interior of $\Delta(\{C, D\})$, as preference parameter δ approaches 0. Consider the limit problem at $\delta = 0$. Optimal transfers $T_{\delta=0}^\pi$ solve the problem

$$\begin{aligned} &\max_{T_z \in [-T_{\max}, T_{\max}]} L(z, T_z, \mu) \\ &\qquad = -\lambda|T_z| + \alpha(2+\lambda)[\pi(D|z) - \pi(C|z)] - \mu_D \pi(D|z) + \mu_C \pi(C|z) \\ &\qquad = -\lambda|T_z| + \left(\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right)[\pi(D|z) - \pi(C|z)] - \frac{\mu_D - \mu_C}{2}, \end{aligned}$$

with μ_D and μ_C such that $\mu_C + \mu_D \leq 2\alpha(2+\lambda)$. For any $\pi(C) \in (0,1)$, solutions to this problem take the following threshold form: there exists $\theta^+ > 0$ and $\theta^- > 0$ such that

$$T^{\pi}_{\delta=0,z} = \begin{cases} 0 & \text{if } f(z|D)/f(z|C) \in [\theta^-, \theta^+] \\ -T_{\max} & \text{if } f(z|D)/f(z|C) < \theta^- \\ T_{\max} & \text{if } f(z|D)/f(z|C) > \theta^+. \end{cases}$$

Consider a sequence of values $(\delta_n)_{n\geq 0}$ converging to 0. A reasoning similar to that of Lemma 4 implies that $T^{\pi}_{\delta_n}$ must converge to $T^{\pi}_{\delta=0}$ under the L_1 norm.

Limit transfer scheme $T^{\pi}_{\delta=0}$ exhibits punitive justice at every state z such that $T^{\pi}_{\delta=0,z} \neq 0$. In addition, transfers $T^{\pi}_{\delta_n}$ converge to $T^{\pi}_{\delta=0}$ under the L_1 norm. Hence, recalling that $\mathcal{L}$ denotes the Lebesgue measure on Z, it must be that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{L}(z \text{ s.t. } |T^{\pi}_{\delta_n,z}| \geq T_{\max} - \epsilon) = \mathcal{L}(z \text{ s.t. } |T^{\pi}_{\delta=0,z}| \geq T_{\max} - \epsilon).$$

Therefore, as δ approaches 0, transfer schemes T^{π}_{δ} must exhibit punitive justice. □

Lemma S.B.2. *Assume that $\gamma = -\frac{1}{2}$ in the numerical example defined Section 3. Player A's equilibrium behavior π is unique and characterized by*

$$\pi(C) = \frac{9\alpha(2+\lambda) - 10\lambda}{9\alpha(2+\lambda) - 4\lambda}.$$

Proof. Recall that $\delta = 0$ in this example. Given behavior $\pi \in \Delta(\{C, D\})$, transfers chosen by the principal solve

$$\max_{T_{-1},T_1} \pi(C)\left[-\frac{1}{4}\lambda|T_{-1}| - \frac{3}{4}\lambda|T_1| - \alpha(2+\lambda)\left|\frac{1}{4}T_{-1} + \frac{3}{4}T_1\right|\right] + \pi(D)\left[-\frac{3}{4}\lambda|T_{-1}| - \frac{1}{4}\lambda|T_1| - \alpha\left|6 - (2+\lambda)\left(\frac{3}{4}T_{-1} + \frac{1}{4}T_1\right)\right|\right].$$

It follows from elementary (though tedious) algebra that whenever $\pi(C) > (9\alpha(2+\lambda) - 10\lambda)/(9\alpha(2+\lambda) - 4\lambda)$, the optimal transfer is $T_{-1} = T_1 = 0$. In contrast, whenever $\pi(C) < (9\alpha(2+\lambda) - 10\lambda)/(9\alpha(2+\lambda) - 4\lambda)$, the optimal transfer sets $T_{-1} \geq 8/(2+\lambda)$ and $T_1 \leq 0$. For such transfers, $\mathbb{E}[u^T_A|C] - \mathbb{E}[u^T_A|D] \geq -1 + 4(1+\lambda)/(2+\lambda) > 0$. It follows that the only equilibrium behavior is $\pi(C) = (9\alpha(2+\lambda) - 10\lambda)/(9\alpha(2+\lambda) - 4\lambda)$. □

References

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser (2008), "Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review*, 98, 294–310. [1]

Bohnet, Iris and Richard Zeckhauser (2004), "Trust, risk and betrayal." *Journal of Economic Behavior & Organization*, 55, 467–484. [1]

Dye, Ronald A. (1985), "Costly contract contingencies." *International Economic Review*, 26, 233–250. [4]

Fehr, Ernst and Klaus M. Schmidt (1999), "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817–868. [2]

Tirole, Jean (2009), "Cognition and incomplete contracts." *American Economic Review*, 99, 265–294. [4]

Co-editor George J. Mailath handled this manuscript.

Submitted 2015-1-7. Final version accepted 2015-11-24. Available online 2015-12-7.