ANNALS OF ECONOMICS AND FINANCE **9-2**, 385–395 (2008)

Inducing Cooperation by Self-Stipulated Penalties

Cheng-Zhong Qin

Department of Economics, University of California, Santa Barbara, CA 93106-9210
E-mail: qin@econ.ucsb.edu

This paper considers an approach for inducing cooperation in prisoner's dilemma. The approach is based on players individually committing to pay self-stipulated penalties for defection. We provide a complete characterization of self-stipulated penalties that are necessary and sufficient to induce the players to cooperate in subgame-perfect equilibrium. An alternative interpretation of the conditions using contract remedies is provided.

Key Words: Penalty for defection; Prisoner's dilemma; Subgame-perfect equilibrium.
JEL Classification Numbers: C72, K12.

1. INTRODUCTION

A fundamental characteristic of the prisoner's dilemma is that each of two players could capture substantial gains through mutual cooperation, but is tempted by even greater gains should the player defect while the other player cooperates. For either player the worst case is to cooperate while the other defects. The result is that both players defect, even though mutual defection leaves each player with a payoff less than the player could have obtained had both players cooperated.

In this paper, we consider an approach to induce cooperation in a prisoner's dilemma game that calls for the players to play an enlarged game with two stages. In stage one, each player independently commits to pay a certain binding amount should he defect while the other player cooperates. For example, a player may leave a good faith deposit with a third-party and specify that the deposit will be paid to the other player when he defects while the other player cooperates.[1] We also consider an alternative

[1]Some employers require an employee to provide a performance bound, which is an amount of money that will be given to the employers if the employee fails to complete

1529-7373/2008

treatment under which each player pays the self-stipulated binding amount when he defects regardless of what the player does. In stage two, players play the prisoner's dilemma game with knowledge of the penalties for defection they each committed to pay.[2]

We say that a penalty configuration "induces" the players to cooperate if there is a subgame-perfect equilibrium (SPE) that involves the players committing to pay penalties prescribed by the penalty configuration in stage one and subsequently cooperating in stage two conditional on them committing to pay these penalties for defection.

The necessary and sufficient conditions for self-stipulated penalties to induce the players to cooperate turn out to require that each player i commit to pay a large enough penalty to deter himself from defecting and, on the other hand, not so large that player $j \neq i$ would rather have player i defect in order to capture his penalty payment than have both of them cooperate. The compatibility between these upper and the lower bounds for each player implies that mutual cooperation is most efficient. Furthermore, as explained in detail in next section, the upper and lower bounds are analogous to the "expectation" and the "disgorgement remedies", respectively.[3] Since expectation remedies measure actual harms, committing to pay more than the actual harm does not survive the incentive compatibility called for by the notion of SPE.[4]

---

certain duties. The employee leaves this bond with the employers or a third-party such as an insurance company before the job begins. See Perloff (2004, pp. 710-712).

[2]Williamson (1983, pp 537-538) discusses the merit of crafting *ex ante* incentive structures for prisoner's dilemma. The idea of inducing cooperation via self-stipulated penalties follows from an ancient technique. Schelling (1960. p. 44) observes that the exchange of hostages served incentive purposes in an earlier age, and suggests that the institution of hostages is an ancient technique that deserves to be studied by game theory (p. 135). For example, during the later part of the Warring States between 475 and 221 B.C. in Chinese history, state Chu and state Qin were the strongest of the states. They could form alliances either with each other or with other weaker states. Doing the former, they could avoid conflict while doing the latter, they could each have the opportunity to become stronger than their counterparts. But, the second strategy would also lead more likely to war. Mutual cooperation could make both Chu and Qin better off than mutual competition. However, each preferred to becoming stronger irrespective of the other one's choice. Realizing the difficulties for achieving mutual cooperation, an advisor from state Qin suggested that the king of Chu offer his heir as a hostage to Qin and that the king of Qin do likewise to induce the two states not to attack each other (see Crump 1996, p. 242-245).

[3]See Section 3.1 for further discussion.

[4]Jackson and Wilkie (2005) consider strategy dependent payoff transfers between the players. Payoff transfers in their paper are not restricted to be achievable through self-stipulated penalties for defection only. The difference between their paper and the present one is that while they focus on what feasible payoff allocations can be achieved in subgame-perfect equilibrium, we focus on what penalty configurations can induce the players to play a particular strategy profile; namely, mutual cooperation.

The rest of the paper is organized as follows. Section 2 introduces the penalty scheme. Section 3 establishes necessary and sufficient conditions for self-stipulated penalties to induce the players to cooperate and provides an interpretation of these conditions in terms of contract remedies. Section 4 concludes the paper.

2. THE PENALTY SCHEME

A generic prisoner's dilemma game has two players each of whom can either cooperate (action C) or defect (action D) with payoffs as in Figure 1.

		Player 2	
		C	D
Player 1	C	(R_1, R_2)	(S_1, T_2)
	D	(T_1, S_2)	(P_1, P_2)

FIG. 1. Prisoner's Dilemma Game with $S_k < P_k < R_k < T_k$, $k = 1, 2$.

The pair (D, D) is the only Nash equilibrium which yields player i a payoff of P_i less than payoff R_i that player i could have obtained had both players cooperated. We assume payoffs are transferable.

Suppose that before playing a prisoner's dilemma game, each player independently commits to pay as penalty a binding amount to the other player should he defect while the other player cooperates. Suppose further payoffs from the play of the prisoner's dilemma game and penalty payments are addable, so that a penalty payment from player i implies a payoff transfer from him to player j conditional on him defecting and j cooperating.

Let $\mathcal{H}_i$ be the set of payoff transfers implied by penalties for defection player i may commit to pay. For simplicity, we assume that the transfer rate is one-to-one. We do not impose any restriction on how much player i can commit to pay in order to study issues such as whether committing to pay more than actual harms one's unilateral defection inflicts upon the other player can survive incentive compatibility. A penalty configuration H changes the prisoner's dilemma game in Figure 1 into game $\Gamma(H)$ in Figure 2, which is the subgame of the enlarged two-stage game that follows penalty configuration H.

		Player 2	
		C	D
Player 1	C	(R_1, R_2)	$(S_1 + H_2, T_2 - H_2)$
	D	$(T_1 - H_1, S_2 + H_1)$	(P_1, P_2)

FIG. 2. The Subgame $\Gamma(H)$ Induced by Penalty Configuration H.

The players can condition choices of actions in the prisoner's dilemma game on penalty configurations. That is, players' choices in $\Gamma(H)$ may depend on H. Denote by Φ_i a mapping that maps each penalty configuration $H \in \mathcal{H}_1 \times \mathcal{H}_2$ into a probability distribution $\Phi_i(H) = (\Pi_i(C, H), \Pi_i(D, H))$ over the action set $\{C, D\}$. Such a mapping specifies a plan of contingent actions in the prisoner's dilemma game for player i. A strategy for player i is thus a pair (H_i, Φ_i) with a penalty H_i he will commit to pay and an action plan Φ_i he will subsequently follow. We consider subgame-perfect equilibrium as the solution concept for our two-stage game.

DEFINITION 2.1. A penalty configuration $H^* = (H_1^*, H_2^*)$ induces the players to cooperate if there are action plans Φ_1^* and Φ_2^* such that (i) the strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is a SPE; (ii) $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.

For a penalty configuration to induce cooperation, Definition 2.1 requires that there exist action plans to be subsequently followed by the players that are *credible*, in the sense that conditional on each penalty configuration H, the action plans specify an action configuration which forms a Nash equilibrium for the subgame $\Gamma(H)$. This requirement prevents each player from using a *non-credible* action plan to make the other player commit to pay a penalty more favorable to him. Next, the penalty configuration must make it incentive compatible for each player to commit to pay the corresponding penalty for defection, given that the other player commits to pay his and given that they both subsequently follow their action plans. Finally, conditional on them committing to pay penalties for defection in the configuration, the players must subsequently cooperate with probability 1.

3. RESULTS

Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. From Figure 2, player 1 receives payoff R_1 and player 2 receives payoff R_2 in the SPE with both players cooperating conditional H^*. Player 1's payoff will be $T_1 - R_1 - H_1^*$ if he defects in $\Gamma(H^*)$, given that player 2 cooperates. Thus, since H^* induces cooperation, H_1^* must satisfy $H_1^* \geq T_1 - R_1$. Next, consider $H_2 \in \mathcal{H}_2$ with $H_2 < T_2 - R_2$. It must be $\Phi_1^*(C, (H_1^*, H_2)) < 1$; otherwise, given player 1's strategy (H_1^*, Φ_1^*), player 2 receives payoff $T_2 - H_2 > R_2$ by committing to pay H_2 and by subsequently defecting in $\Gamma(H_1^*, H_2)$. Hence, given player 1's strategy (H_1^*, Φ_1^*), by committing to pay $H_2 < T_2 - R_2$ and by cooperating in $\Gamma(H_1^*, H_2)$, player 2 receives expected payoff $\Phi_1^*(C, (H_1^*, H_2))R_2 + [1 - \Phi_1^*(C, H_1^*, H_2))][S_2 + H_1^*]$. Since $\Phi_1^*(C, (H_1^*, H_2)) < 1$ as argued above, H_1^* must also satisfy $H_1^* \leq R_2 - S_2$. In summary, we have shown $T_1 - R_1 \leq H_1^* \leq R_2 - S_2$. Similarly, $T_2 - R_2 \leq H_2^* \leq R_1 - S_1$.

LEMMA 1. *Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. Then, for $i \neq j$, $T_i - R_i \leq H_i^* \leq R_j - S_j$.*

When $P_1 - S_1 < T_2 - R_2$, H_1^* must also satisfy $P_2 - S_2 \leq H_1^* \leq T_1 - R_1$. To see this, consider $H_2 \in \mathcal{H}_2$ with $P_1 - S_1 < H_2 < T_2 - R_2$. Then from Figure 2, $H_2 > P_1 - S_1$ implies that given that player 2 defects, the unique optimal action for player 1 in $\Gamma(H_1^*, H_2)$ is C.

If $H_1^* > T_1 - R_1$, then C remains the unique optimal action in $\Gamma(H_1^*, H_2)$ for player 1, given that player 2 cooperates. It follows that $H_1^* > T_1 - R_1$ and $H_2 > P_1 - S_1$ together imply that C is strictly dominant for player 1 in $\Gamma(H_1^*, H_2)$. Since $H_2 < T_2 - R_2$ and since $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma(H_1^*, H_2)$, we must have $\Phi_1^*(C, (H_1^*, H_2)) = 1$ and $\Phi_2^*(C, (H_1^*, H_2)) = 0$. This shows that by deviating from (H_2^*, Φ_2^*) to (H_2, Φ_2^*) with $P_1 - S_1 < H_2 < T_2 - R_2$, player 2 can guarantee himself a payoff equal to $T_2 - H_2 > R_2$, while his payoff would be at most R_2 were he to commit to pay H_2^*. This contradicts the fact that H^* induces the players to cooperate. We conclude $H_1^* \leq T_1 - R_1$ whenever $P_1 - S_1 < T_2 - R_2$.

If $H_1^* < P_2 - S_2$, however, then it together with $H_2 < T_2 - R_2$ implies that action D is strictly dominant for player 2 in $\Gamma(H_1^*, H_2)$. In this case, since $H_2 > P_1 - S_1$ and since $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma(H_1^*, H_2)$, it must be $\Phi_1^*(C, (H_1^*, H_2)) = 1$ and $\Phi_2^*(C, (H_1^*, H_2)) = 0$. It follows that by deviating from (H_2^*, Φ_2^*) to (H_2, Φ_2^*) with $P_1 - S_1 < H_2 < T_2 - R_2$, player 2 can guarantee himself a payoff equal to $T_2 - H_2 > R_2$. Thus, it must be $H_1^* \geq P_2 - S_2$ whenever $P_1 - S_1 < T_2 - R_2$.

In summary, the preceding analysis shows $P_2 - S_2 \leq H_1^* \leq T_1 - R_1$ whenever $P_1 - S_1 < T_2 - R_2$. By analogy, $P_1 - S_1 \leq H_2^* \leq T_2 - R_2$ whenever $P_2 - S_2 < T_1 - R_1$. We thus have:

LEMMA 2. *Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. Then, $P_j - S_j \leq H_i^* \leq T_i - R_i$ whenever $P_i - S_i < T_j - R_j$ for $i \neq j$.*

Conditions in Lemmas 1 and 2 turn out to be not only necessary but also sufficient for penalty configuration H^* to induce the players to cooperate. This result is summarized in the following theorem.

THEOREM 1. *A penalty configuration H^* induces the players to cooperate if and only if*

$$T_i - R_i \leq H_i^* \leq R_j - S_j, \tag{1}$$

and

$$P_j - S_j \leq H_i^* \leq T_i - R_i \text{ whenever } P_i - S_i < T_j - R_j, \tag{2}$$

for $i \neq j$.

Proof. See the Appendix. ▌

Notice that when $P_2 - S_2 \geq T_1 - R_1$ and $P_1 - S_1 \geq T_2 - R_2$, the set of penalty configurations inducing the players to cooperate is determined completely by condition (1). In that case, the set is rectangular.

The lower bound $T_i - R_i$ on H_i^* is needed to deter player i from defecting. On the other hand, the upper bound $R_j - S_j$ on H_i^* is needed for player i to deter player j from committing to pay a penalty smaller than $T_j - R_j$. Player i's action to defect conditional on such smaller penalties committed to pay by player j provide the deterrence. However, such deterrence is not credible when $H_i^* > R_j - S_j$, because then player j would rather have player i defect in which case he receives $S_j + H_i^*$ by subsequently cooperating, than have both cooperate in which case he receives $R_j < S_j + H_i^*$.

Notice that the compatibility of the upper and the lower bounds in (1) imply $T_1 + S_2 \leq R_1 + R_2$ and $T_2 + S_1 \leq R_1 + R_2$. Hence, with the compatibility, mutual cooperation is most efficient, in the sense that the sum of players' payoffs is the greatest.

3.1. An Alternative Interpretation of the Lower and Upper Bounds

In the law of contracts, an "expectation remedy" is a payment that places the victim of a breached contract in the position he would have been in had the other party performed (Cooter and Ulen 2000, pp. 226). Assume that

the prisoner's dilemma game in Figure 1 arises from a contract between players 1 and 2. If player i cooperates (performs) and player j defects (does not perform), player i's payoff is S_i. Player i's payoff would have been R_i if player j had cooperated. Thus to place player i in the position he would have been in had player j cooperated, it would require that player j pay player i the amount $R_i - S_i$. It follows that the expectation remedy when j is held liable for is $R_i - S_i$. The upper bounds are thus analogous to the expectation remedies.

A "disgorgement remedy" is a payment paid to the victim of a breached contract to eliminate the breacher's profit from wrong doing (Cooter and Ulen (2000, pp. 234). From Figure 1, player j gains $T_j - R_j$ units more from defecting given that player i cooperates. Thus to eliminate this gain from wrong doing (not performing), it would require that player j pay the amount $T_j - R_j$. It follows that the disgorgement damage remedy when player j is held liable for is $T_j - R_j$. The lower bounds are thus analogous to the disgorgement remedies.

3.2. A Variant of the Penalty Scheme

Consider a variant of the preceding penalty scheme under which each player commits to pay a binding amount whenever he defects. Denote by $\Gamma'(H)$ the subgame conditional on penalty configuration H. This subgame is shown in Figure 3.

		Player 2	
		C	D
Player 1	C	(R_1, R_2)	$(S_1 + H_2, T_2 - H_2)$
	D	$(T_1 - H_1, S_2 + H_1)$	$(P_1 - H_1 + H_2, P_2 - H_2 + H_1)$

FIG. 3. The Subgame $\Gamma'(H)$ Induced by Penalty Configuration H.

Under this variant necessary and sufficient conditions for penalties to induce the players to cooperate change to:

THEOREM 2. *A penalty configuration H^* induces the players to cooperate under the variant of the penalty mechanism if and only if*

$$T_i - R_i \leq H_i^* \leq \min\{P_i - S_i, R_j - P_j\},\ i \neq j. \tag{1'}$$

Proof. See the Appendix. ▌

Since $S_1 < P_1$ and $S_2 < P_2$, (1′) implies $T_1 - R_1 \leq H_1^* < R_2 - S_2$ and $T_2 - R_2 \leq H_2^* < R_1 - S_1$. This means that (1′) implies (1). However, (1′) does not necessarily imply both (1) and (2) nor conversely.

4. CONCLUSION

We considered self-stipulated penalties for defection as inducements to cooperate. We provided a complete characterization of penalty configurations that are necessary and sufficient to induce the players to cooperate. When players are motivated by their own material payoffs only, penalty configurations consistent with our characterization are equally effective in inducing the players to cooperate. By experimentally testing the effectiveness of these consistent penalty configurations in inducing cooperation, results in this paper are helpful for testing factors other than own material payoffs such as equity and fairness that may affect players' behavior.

Our characterization result establishes a lower and upper bounds on penalty configurations inducing cooperation in prisoner's dilemma. These lower and upper bounds correspond to disgorgement and expectation remedies. In the law of contracts, a liquidated remedy is defined as an amount predetermined by the parties themselves *rather than* imposed upon them by a court as the total compensation to an injured party should the other party breach (see Cooter and Ulen 2000, pp. 225-237). In In practice, courts generally will not enforce a liquidated remedy unless it is a reasonable approximation of the expectation measure of damages. Our lower and upper bounds on penalty configurations inducing the players to cooperate are consistent with the enforceability requirement of liquidated remedies.

APPENDIX

Let $U_1((H_1, \Phi_1), (H_2, \Phi_2))$ and $U_2((H_1, \Phi_1), (H_2, \Phi_2))$ denote the payoffs for player 1 and player 2, respectively, at strategy profile $((H_1, \Phi_1), (H_2, \Phi_2))$. Let $L_i = P_i - S_i$ and $G_i = T_i - R_i$ for $i = 1, 2$.

Proof of Theorem 1: The necessity of (1) and (2) follows directly from Lemma 1 and Lemma 2. Thus, it only remains to prove the sufficiency of these conditions.

Let $H^* \in \mathcal{H}$ be a penalty configuration satisfying conditions (1)-(2). For $H_2 \in \mathcal{H}_2$, let $\Phi_1^*(H_1^*, H_2)$ and $\Phi_2^*(H_1^*, H_2)$ be defined by

$$\Phi_1^*(C,(H_1^*,H_2)) = \begin{cases} 1 & \text{if } H_2 \geq G_2, \\ 0 & \text{if } L_1 \leq H_2 < G_2, \\ 0 & \text{if } H_2 < \min\{L_1, G_2\},\ H_1^* \leq L_2 \\ \frac{H_1^* - L_2}{G_2 - H_2 + H_1^* - L_2} & \text{if } H_2 < \min\{L_1, G_2\},\ H_1^* > L_2 \end{cases} \tag{A.1}$$

and

$$\Phi_2^*(C,(H_1^*,H_2)) = \begin{cases} 1 & \text{if } H_2 \geq G_2, \\ 1 & \text{if } L_1 \leq H_2 < G_2, \\ 0 & \text{if } H_2 < \min\{L_1, G_2\},\ H_1^* \leq L_2 \\ \frac{L_1 - H_2}{H_1^* - G_1 + L_1 - H_2} & \text{if } H_2 < \min\{L_1, G_2\},\ H_1^* > L_2. \end{cases} \tag{A.2}$$

For $H_1 \in \mathcal{H}_1$, let $\Phi_1^*(H_1, H_2^*)$ and $\Phi_2^*(H_1, H_2^*)$ be defined analogously. Finally, for $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$, let $(\Phi_1^*(H), \Phi_2^*(H))$ be any Nash equilibrium for $\Gamma(H)$.[1] By (1), (3), and (4), $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.

Consider $H_2 \in \mathcal{H}_2$. Suppose first $H_2 \geq G_2$. By (1), $H_1^* \geq G_1$. It follows that (C, C) is a Nash equilibrium for $\Gamma(H_1^*, H_2)$. Suppose now $L_1 \leq H_2 < G_2$. In this case, $L_1 < G_2$. Hence, by (1) and (2), $L_2 \leq H_1^*$ and $H_1^* = G_1$. Consequently, (D, C) is a Nash equilibrium for $\Gamma(H_1^*, H_2)$. Suppose finally $H_2 < \min\{L_1, G_2\}$. In this case, if $H_1^* \leq L_2$, then $P_1 - L_1 + H_2 < P_1$ and $P_2 - L_2 + H_1^* \leq P_2$, implying that (D, D) is a Nash equilibrium for $\Gamma(H_1^*, H_2)$. If $H_1^* > L_2$, then (3) and (4) imply that given player 2's strategy (H_2, Φ_2^*), action C and action D yield the same payoff to player 1 in $\Gamma(H_1^*, H_2)$, while given player 1's strategy (H_1^*, Φ_1^*), action C and action D yield the same payoff to player 2 in $\Gamma(H_1^*, H_2)$. Thus, $\Phi^*(H_1^*, H_2)$ as defined by (3) and (4) is a Nash equilibrium for $\Gamma(H_1^*, H_2)$.

In summary, we have shown that for any $H_2 \in \mathcal{H}_2$, $\Phi^*(H_1^*, H_2)$ as in (3) and (4) is a Nash equilibrium for $\Gamma(H_1^*, H_2)$. By analogy, for any $H_1 \in \mathcal{H}_1$, $\Phi^*(H_1, H_2^*)$ is a Nash equilibrium for $\Gamma(H_1, H_2^*)$. Thus, to complete the proof of the sufficiency, it only remains to show that players do not have any incentive to unilaterally change their penalties in H^*.

[1] The specifications for $\Phi^*(H) = (\Phi_1^*(H), \Phi_2^*(H))$ at $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$ are inessential to subgame-perfect equilibrium analysis.

To this end, consider $H_2 \in \mathcal{H}_2$. By (3), (4), and by Figure 2,

$$U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) = \begin{cases} R_2 & \text{if } H_2 \geq G_2, \\ P_2 - L_2 + H_1^* & \text{if } L_1 \leq H_2 < G_2, \\ P_2 & \text{if } H_1^* \leq L_2,\ H_2 < \min\{L_1, G_2\}, \\ R_2' & \text{if } H_1^* > L_2,\ H_2 < \min\{L_1, G_2\}, \end{cases} \tag{A.3}$$

where $R_2' = \Phi_1^*(C, (H_1^*, H_2))R_2 + \Phi_1^*(D, (H_1^*, H_2))(S_2 + H_1^*)$. By (1), $H_1^* \leq R_2 - S_2$ which implies $R_2' \leq R_2$. Consequently, by (5),

$$U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) \leq R_2.$$

This shows that player 2 does not have any incentive to change his penalty. Similarly, player 1 does not have any incentive to change his penalty. ▌

Proof of Theorem 2: Let H^* be a penalty configuration inducing the players to cooperate. By Definition 1, there are action plans Φ_1^* and Φ_2^* such that $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$ and the strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is a subgame-perfect equilibrium. From Figure 3, player 1 receives payoff R_1 and player 2 receives payoff R_2 in this subgame-perfect equilibrium. Player 1's payoff becomes $T_1 - H_1^*$ if he defects in $\Gamma'(H^*)$, given that player 2 cooperates. Thus it must be $H_1^* \geq G_1$.

Suppose $H_1^* > L_1$. Let $H_2 = 0$. From Figure 3, action D is strictly dominant for player 2 in $\Gamma'(H_1^*, 0)$. This together with $H_1^* > L_1$ implies that the unique Nash equilibrium for $\Gamma'(H_1^*, 0)$ is (C, D). Consequently, given player 1's strategy (H_1^*, Φ_1^*), player 2 receives payoff $T_2 > R_2$ by deviating from (H_2^*, Φ_2^*) to $(0, \Phi_2^*)$. This contradicts the fact that H^* induces the players to cooperate. Hence, $H_1^* \leq L_1$ and $\Phi_1^*(C, (H_1^*, 0)) < 1$. Now observe that given player 1's strategy (H_1^*, Φ_1^*), by committing to pay $H_2 = 0$ for defection and by defecting in $\Gamma'(H_1^*, 0)$, player 2's payoff would be $\Phi_1^*(C, (H_1^*, 0))T_2 + \Phi_1^*(D, (H_1^*, 0))[P_2 + H_1^*]$. Since $\Phi_1^*(C, (H_1^*, 0)) < 1$, it must be $P_2 + H_1^* \leq R_2$ or equivalently $H_1^* \leq R_2 - P_2$ for $\Phi_1^*(C, (H_1^*, 0))T_2 + \Phi_1^*(D, (H_1^*, 0))[P_2 + H_1^*] \leq R_2$ to satisfy. In summary, we have shown that H_1^* must satisfy $G_1 \leq H_1^* \leq \min\{L_1,\ R_2 - P_2\}$. By analogy, H_2^* must satisfy $G_2 \leq H_2^* \leq \min\{L_2,\ R_1 - P_1\}$.

Conversely, let H^* be a penalty configuration satisfying $(1')$. We show that H^* induces the players to cooperate. To this end, for $H_2 \in \mathcal{H}_2$, let $\Phi_1^*(H_1^*, H_2)$ and $\Phi_2^*(H_1^*, H_2)$ be defined by

$$\Phi_1^*(C, (H_1^*, H_2)) = \Phi_2^*(C, (H_1^*, H_2)) = \begin{cases} 1 & \text{if } H_2 \geq G_2, \\ 0 & \text{if } H_1 < G_2. \end{cases} \tag{A.4}$$

For $H_1 \in \mathcal{H}_1$, let $\Phi_1^*(H_1, H_2^*)$ and $\Phi_2^*(H_1, H_2^*)$ be defined analogously. Finally, for $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$, let $\Phi^*(H)$ be any Nash equilibrium for the subgame $\Gamma'(H)$. Notice that, since $H_1^* \geq G_1$ and $H_2^* \geq G_2$,

(6) implies $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$. Notice also that the conditions on H^* imply $G_1 \leq H_1^* \leq L_1$ and $G_2 \leq H_2^* \leq L_2$. Thus for $H_2 \in \mathcal{H}_2$, (C, C) is a Nash equilibrium for $\Gamma'(H_1^*, H_2)$ when $H_2 \geq G_2$; (D, D) is a Nash equilibrium for $\Gamma^{p'}(H_1^*, H_2)$ when $H_2 < G_2$. This shows that the action pair $\Phi^*(H_1^*, H_2)$ in (6) is a Nash equilibrium for $\Gamma'(H_1^*, H_2)$, for all $H_2 \in \mathcal{H}_2$. By analogy, for all $H_1 \in \mathcal{H}_1$, the action pair $\Phi^*(H_1, H_2^*)$ is also a Nash equilibrium for $\Gamma(H_1, H_2^*)$. Thus to show that H^* induces the players to cooperate, it only remains to prove that the players do not have any incentive to unilaterally change penalties that constitute H^*.

Consider $H_2 \in \mathcal{H}_2$. By (6) and Figure 3, player 2's payoff at $((H_1^*, \Phi_1^*), (H_2, \Phi_2^*))$ is R_2 whenever $H_2 \geq G_2$ and his payoff is $P_2 - H_2 + H_1^*$ whenever $H_2 < G_2$. Since $H_1^* \leq R_2 - P_2$, it follows that player 2 has no incentive to unilaterally deviate from H_2^*. By analogy, player 1 has no incentive to unilaterally deviate from H_1^* either. ▌

REFERENCES

Cooter, Robert and Ulen, Thomas, 2000. *Law and Economics.* Addison-Wesley, Third Edition.

Crump, James I., 1996. *Chan-Kuo Ts'e.* Revised Edition. The University of Michigan Press.

Jackson, Matthew O. and Wilkie, Simon, 2005. Endogenous Games and Mechanisms: Side Payments Among Players. *Review of Economic Studies* **72**, 543-566.

Perloff, Jeffrey M., 2004. *Microeconomics.* Pearson Addison Wesley, Third Edition.

Schelling, Thomas, 1960. *The Strategy of Conflict.* Cambridge, Massachusetts: Harvard University Press.

Williamson, Oliver E., 1983. Credible Commitments: Using Hostages to Support Exchange. *American Economic Review* **73(4)**, pp. 519-540.