Hostages as a commitment device
A game-theoretic model and an empirical test of some scenarios*

Werner Raub

Utrecht University, Utrecht, Netherlands

Gideon Keren

Free University of Amsterdam, Amsterdam, Netherlands

Received July 1990, final version received May 1992

Hostages posting is a mechanism of cooperation in social and economic relations with incentive problems. A game-theoretic model of hostage posting as a commitment device is presented. Cooperation via hostage posting is individually rational if it is supported by a subgame perfect equilibrium. Maximin properties of such equilibria are also considered. Predictions derived from the model were tested in several experiments. Results support the empirical significance of subgame perfect equilibria and of maximin-strategies.

'The institution of hostages is an ancient technique that deserves to be studied by game theory, as does the practice of drinking wine from the same glass or of holding gang meetings in places so public that neither side could escape if it subjected the other to massacre.' [Schelling (1960, p. 135)]

1. Cooperation via hostages in problematic social situations

In a classic analysis, Schelling (1960) highlighted the use of commitments in social and economic relations. He stressed a paradoxical feature of commitments, namely, that it can be profitable for an actor to incur them:

'The essence of these tactics is some voluntary but irreversible sacrifice of freedom of choice. They rest on the paradox that the power to constrain an adversary may depend on the power to bind oneself; that, in bargaining, weakness is often strength, freedom may be freedom to capitulate, and to burn bridges behind one may suffice to undo an opponent.' [Schelling (1960, p. 22)]

Correspondence to: Werner Raub, Department of Sociology, Utrecht University, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands, email: raub@cc.ruu.nl.

*We thank Hartmut Kliemt, Albert Verbeek, and, in particular, Jeroen Weesie for their helpful comments. Financial support was provided by the Faculty of Social Sciences, Utrecht University, the 'Stichting Empirisch-Theoretsch Onderzoek in de Sociologie' (SETOS), and the Hans Frisch-Stiftung. The paper was written and revised when the first author held visiting appointments at the Center for Study of Public Choice, Fairfax, Virginia, and at the Department of Sociology, University of Chicago. The support of these institutions is gratefully acknowledged.

0167-2681/93/$06.00 © 1993—Elsevier Science Publishers B.V. All rights reserved
In this paper, we analyze a specific kind of commitment device: the use of hostages. Like commitment devices in general, hostages can be used in two broad types of situations. The first type, to which Schelling refers, involves distribution problems. In such situations, the use of a hostage induces an outcome that is more favorable for the actor posting it and less favorable for his or her partner(s), compared to the outcome that would have resulted without the use of hostages.\(^1\) An example from industrial organization and imperfect competition is the problem of entry deterrence [see Dixit (1982)]. Assume that a monopolist, the incumbent, faces a potential entrant in its market. The incumbent has to decide whether to react aggressively by fighting a costly price war, or to share the market. A price war is not profitable for the potential entrant. If the potential entrant considers the incumbent's threat of reacting aggressively to be credible, he or she will thus not enter. However, an incumbent's threat of fighting entrance will be credible only if the incumbent has incentives to implement the threat, given that the relevant contingency – entrance – occurs. One way to establish such credibility is by prior investments in capacity which cannot be costlessly liquidated. These investments of the incumbent, then, serve as a hostage that makes fulfillment of the commitment to fight entry optimal. In this way, the incumbent is able to deter the potential entrant.

We focus on a second type of situations, in which hostages are used in order to solve efficiency problems. In these situations, the use of hostages induces an outcome that is profitable for all actors compared to the outcome which results if hostage posting is not feasible. A sample of examples will be useful to illustrate such settings.

(i) As Williamson (1985, p. 163) puts it, hostages may be used 'to craft intermediate structures, located between discrete market contracting at the one extreme and hierarchical organization on the other, whereby the hazards of bilateral contracting are attenuated with less severe sacrifices [compared to hierarchical organization] in incentive and scale/scope economy respects'. Williamson (1985, chs. 7 and 8) shows that the technique of hostage posting is a rather common and frequently used practice in economic exchange relations that involve investments in transaction specific assets, and are therefore exposed to cancellation hazards. Assume that one of the parties (the seller) has irreversibly invested in some special purpose technology (e.g., by buying a machine) serving the specific demands of a particular partner (the buyer). These specific investments get lost if the business partner cancels an order because he or she meets more favorable market conditions from other sellers. Safeguards against these hazards are needed. Frequently, there

\(^1\)One might argue that 'pledging bonds' is a more appropriate description of these situations than 'posting hostages'. In this paper, we adopt the terminology introduced by Schelling (1960) and extensively used by Williamson (1985, chs. 7 and 8).
is also a need for self-enforcing commitments and self-enforcing agreements
because courts and other third parties are unavailable or too costly as
enforcing agents. In the case of unilateral trade relations, credible commit-
ments (of the buyer) can be made by posting a hostage (e.g., the machine)
with the producer [Williamson (1985, p. 171–176)]. Reciprocal exchange
relations may in principle be exposed to similar hazards in the case where
both parties have made specific investments. Williamson (see 1985, pp. 194–
197) stresses the advantages of using 'hostages' in the bilateral as compared
to the unilateral case. Whereas the unilateral transfer of a hostage exposes
the transferring party to the risk of being exploited, these expropriation
hazards do not occur in the bilateral case.

(ii) A related form of mutual hostage posting is cross-licensing [see Van
Cayseele and Schreuder (1988, p. 1157)]. A firm F operating in region X may
contemplate to license another firm G operating in region Y to which F has
no access. However, an obvious problem arises: Firm G, strong in R&D,
might be able to use the transfer of knowledge associated with licensing to
invent around F's patent or even to come up with a superior technology.
Without additional safeguards, G's promise not to use such practices will be
hardly credible. One such safeguard can be that G licenses F to use a
technological innovation of G in region X. Each firm's own license now
becomes a hostage which deters abuse of the license received from the
partner. In this way, incentives to violate unilaterally a license agreement are
reduced because opportunities for sanctions of violations are created.

(iii) Hostage posting is a common practice in the real estate business in
the Netherlands. The sale of one-family houses is organized such that buyer
and seller first sign a bill of sale. The transfer of ownership itself usually
takes place some months later. To organize the transaction in such a way is
obviously more attractive for both parties than the alternative arrangement
where agreement on the sale and transfer of ownership coincide: the process
of searching a new house proceeds smoothly, both parties gain time to
arrange their respective removals, etc. However, after signing the bill of sale
and before the transfer of ownership, each partner may face, with some
positive probability, incentives to violate the contract. For example, the seller
may find out that his or her own new house will not be ready for moving in
time and may therefore wish to postpone the transfer of ownership. On the
other hand, the buyer may find another house that suits his or her purposes
better. Hostages are used as a precaution against such contingencies. They
provide a mechanism that facilitates the use of the mutually beneficial
'stepwise' sales arrangement. The buyer's hostage is ten percent of the
purchase price which the buyer has to transfer immediately after signing the
bill of sale to the account of a conveyancer. This hostage is given to the
seller should the buyer violate the contract. The seller likewise posts a
hostage. He or she agrees that the buyer has to pay only 90 percent of the
purchase price in case that the seller does not stick to the terms of the agreement.

(iv) Applications of hostage mechanisms are also found in the context of principal–agent relations. In an early paper, Becker (1960) has pointed out that arrangements like pension funds or seniority rules for promotion, can be considered as 'side bets' providing incentives for long-term employment relationships and constraining hasty job change. An elaborate analysis of a very specific use of hostages in principal–agent relations, where the hostage is not a valuable resource but an action alternative, has been provided by Gilson and Mnookin (1990). They consider the 'up-or-out system', used by traditional American corporate law firms as the typical career pattern for new employees which enter the firm as associates. The system can be roughly described as follows [Gilson and Mnookin (1990, p. 210)]:

'Over a period of up to some ten years an associate progresses toward a decision concerning whether he or she will be promoted to partner. This status conveys lifetime tenure and the right to participate in the profits of the firm (...) The peculiar characteristic of this career pattern is what happens if the associate is not promoted to partner. However close the associate may have come, if the partnership standard is not met the associate is fired rather than retained as an associate.'

At first sight, the up-or-out system constitutes a puzzle because it seems to work against the interests of both the firm and the associate. By using the system, the two parties give up options that may prove valuable for both of them at the time of the partnership decision, namely, to continue the employment relation in a capacity other than partner. Gilson and Mnookin (1990, pp. 216–217) argue that 'the up-or-out system, seemingly foolish from the perspective of both the firm and the associate when evaluated at the time when the partnership decision is made, is revealed to provide an effective structural bonding technique when evaluated at the time when the associate is initially hired'. Such an arrangement is needed because the firm needs a lengthy apprenticeship period in order to assess legal skills and personal characteristics of newly hired associates. This is consistent with the associate’s interests at the time of initial hire only if opportunistic behavior of the firm at the time of partnership decision is excluded. However, the firm will face incentives for opportunism at that time:

'The firm is making a substantial profit from the associate’s labor. Promoting the associate to partner (...) rather than continuing the associate as an employee is costly to the firm; it diminishes the profits accruing to existing partners (...) Suppose the firm opportunistically offers only a permanent associate position even though the associate has actually met the partnership standards. So long as the firm’s non-partnership offer is more lucrative than what the associate could receive in alternative employment, an easily satisfied condition because the associate’s firm-specific capital is by definition significantly less valuable to any other employer, the associate is better off taking less than the original promise despite the firm’s opportunistic behavior.' [Gilson and Mnookin (1990, p. 214)]

The rationale for an up-or-out system (backed with, e.g., appropriate outplacement procedures) becomes now more obvious:
By committing itself to fire anyone who is not made a partner, the firm effectively eliminates its incentive to undertake the very opportunistic behavior that creates the associate's uncertainty. With the up-or-out system in place, the firm cannot manipulate the partnership decision so as to retain in some other capacity the services of an associate whose performance really merits partnership. [Gilson and Mnookin (1990, p. 217)]

From a more general viewpoint, hostages are used in our examples as a mechanism of cooperation in situations that are 'problematic' in the sense that individually rational, i.e., incentive-guided behavior can generate outcomes being suboptimal in the Pareto sense. The incentive structure is such that certain combinations of 'cooperative' behavior of the actors entail Pareto efficiency and are profitable for each actor compared to other combinations of 'defecting' behavior. However, for at least some actors, there are individual incentives to abstain from cooperation and to defect. If all of these actors follow these incentives, they produce the suboptimal result (see Raub and Voss (1986) for an explication of the intuitive notion of a 'problematic situation' or a 'social dilemma', as it is frequently referred to in social psychology).

In this paper, we develop and test a formal model of hostage posting. We do not provide a model that is neatly suited to specific characteristics of any one of the above-mentioned examples. Rather, we try to expose as simply as possible driving forces that capture some of the essential characteristics of the mechanism.

The model addresses three basic issues. First, as can be seen from our examples, hostage posting occurs in situations of strategic interdependence among actors. An analysis of hostage posting thus calls for game-theoretic modelling. The strategic aspect implicit in hostage posting is frequently acknowledged and sometimes even stressed - most obviously in Schelling's analysis. Nevertheless, and perhaps somewhat surprisingly, an explicit game-theoretic model is still lacking. In the following, such a model is provided. The straightforward interpretation of the model would be a (conditionally) normative one [e.g., Luce and Raiffa (1957, p. 63)]. Like game theory in general, it tells players how to behave in order to achieve their goals most effectively. Not withstanding, the model may also be interpreted as a descriptive one, attempting to explain and predict how actual choices are being made. To validate such an interpretation, the model has to be tested with regard to how well it approximates actual behavior. The empirical part of this paper describes several experiments that were designed to test the explanatory and predictive power of the model from a descriptive viewpoint.

Second, notice that our examples have a somewhat 'functionalistic' flavor. They stress the efficiency aspect of hostage posting: If hostages are posted,

---

2It seems fair to state that the emphasis of much recent game-theoretic analyses of cooperation in problematic situations is on the effects of repeated interactions. In contrast, we consider prospects of cooperation in non-repeated interactions.
the parties involved do cooperate and are thus all better off compared to a situation where no hostages are posted. But even if it is taken for granted that hostage posting would induce Pareto efficient outcomes, the problem of sufficient individual incentives to post hostages remains. To use Anatol Rapoport's well known labels, we have to distinguish between the collective rationality of hostage posting and the individual rationality of that mechanism. Our model will stress the latter aspect. In this model, hostage posting becomes a feasible option in a (noncooperative) game and we will specify conditions under which it is in the individual interest of a player (and thus self-enforcing) to use that option.

A third basic feature of our model can be related to a comment by Williamson (1985, p. 204):

'Contrary to the prevailing view that hostages are a quaint concept with little or no practical importance to contemporary contracting, the use of hostages to support exchange is widespread and economically important. But hostage creation is only part of the story. Expropriation hazards and prospective maladaptation conditions also have to be considered. Complex governance structures, of which reciprocal trading is one, arise in response to such conditions.'

Williamson seems to allude that an incentive guided actor's willingness to post hostages may depend on specific attributes of hostages. For example, one may suspect that hostage posting occurs only if it cannot be exploited by the partner. Our model establishes this conjecture formally as well as empirically. Explicit conclusions are derived with respect to the kinds of hostages that will or will not be posted by rational actors. The conditions under which a hostage will be forfeited and attributes like the value of a hostage for the actor who posts it as well as for this actor's partner turn out to have a decisive influence on an actor's readiness to make use of a hostage posting option. Such characteristics of hostages depend on the institutional framework for hostage posting. We provide a comparative analysis of institutional arrangements. As will be shown, rational actors will post specific kinds of hostages and will not post specific other kinds of hostages.

The remainder of the paper is organized as follows. A game-theoretic model of hostage posting is developed in the next section. Following, empirical tests of the descriptive adequacy of the model are presented. Some implications are discussed in the final section.

2. A simple game-theoretic model of hostage posting

In this section we provide a theoretical framework for the analysis of...
hostage posting. We envisage a paradigmatic situation with efficiency problems, namely, a symmetric 2-person Prisoner’s Dilemma (PD). Each player $i$ ($i=1,2$) has a strictly dominant and, thus, uniquely determined individually rational strategy $D_i$ (‘defection’). However, the outcome associated with mutual defection is Pareto inefficient. On the other hand, the outcome associated with mutual choice of $C_i$ (‘cooperation’) is Pareto efficient and better for both players than mutual defection.

The PD is a subgame of an enlarged 2-person hostage game. Hostage posting is explicitly modelled as a move in the enlarged game. We thus employ a well-known, but not frequently used, suggestion of Schelling (1960, p. 127) for the representation of ‘actions in the course of the game that irreversibly change the game itself.’ In fact, broadly conceived, we apply the Nash program [Nash (1951)] to explicitly model all kinds of pre-play behavior as moves in the extensive form of an extended noncooperative game and to derive all kinds of cooperative behavior as a Nash equilibrium of that extended game.

The hostage game is a 2-stage noncooperative 2-person game with complete information, where complete information is assumed to be common knowledge. In stage 1, hostage posting decisions are made. For reasons of simplicity, we assume that hostages can be posted without incurring any (transaction) costs. Both players can post a hostage, i.e., we neglect the case where hostage posting is an option for only one party [see Williamson (1985, ch. 7) for examples]. Players have to decide independently on whether or not to post a hostage. Thus, no player is able to condition his own hostage posting decision on hostage posting of his partner. We focus on the simple case where one hostage is available for each player, Thus, we neglect hostage selection problems that arise if players can choose from a set of hostages. Player $i$’s decision to post or not to post a hostage will be denoted by $H_i^+$ and $H_i^-$, respectively. Combinations of players’ moves in stage 1 are denoted by $H=(H_1, H_2)$. After stage 1, players are informed on each other’s first move. Subsequently, players enter stage 2. Here, they play a subgame where they have to decide, independently and simultaneously, on whether to behave cooperatively or defectively. Combinations of second
moves are denoted by $X = (X_1, X_2)$. After stage 2, players receive a payoff $u_i(H, X) = u_i(Z)$ that depends on the payoff associated with $X$ in the original PD, the hostage postage decisions of the players, the institutional arrangements for hostages, and the values of the hostages.

We denote by $K_{ij}$ ($j = 1, 2$; here we may have $i = j$, while subsequently $i \neq j$) the value of player $j$'s hostage for player $i$. Formally, hostage values represent utility differences: their 'location' is well-defined and they are 'scaled' together with the payoffs from the PD. Due to the fact that we presuppose complete information, it seems natural to assume that $K_{ii} > 0$ and $K_{ij} \geq 0$ for $i = 1, 2$, i.e., each player values his or her own hostage positively and the hostage of one of the players is no burden for the other. We also assume $K_{ii} \geq K_{ij}$ for $i = 1, 2$. According to this assumption, each player values his own hostage at least as much as the other player's hostage, so that players can not gain by exchanging their hostages. Given $K_{ii} \geq K_{ij}$, hostage posting can only serve commitment purposes and has no exchange function. Finally, we assume that player $i$'s payoff $u_i(H, X)$ in the hostage game is additive in $i$'s payoffs associated with $X$ in the PD and values of hostages lost and received.

Consider now different kinds of institutional arrangements for hostages. A hostage institution $I$ defines loosely speaking, under what conditions players will lose their hostage and what will happen with a lost hostage. More precisely, a hostage institution determines for each endnode of the hostage game which hostages are owned by which player. We consider a subset of institutions that are characterized by a number of simple properties according to four assumptions.

(A1) Three alternative arrangements: There are three possibilities with respect to what happens with a hostage after the hostage game has been played: the hostage is given back to the owner who posted it ($B$), the hostage is given to the other player ($O$), or the hostage is lost ($L$), i.e., given to some outside party.

These alternatives seem rather natural in situations with exactly one cooperative and one defective action, as opposed to more complex cases where players can choose between actions that differ in their degree of (non-) cooperativeness.

(A2) Voluntary hostage posting: Hostages are never forfeited (i.e., lost or given to the other player) if they have not been posted.

Under (A2), if a player decides not to post a hostage, he or she will never lose it. In this sense, hostage posting is voluntary. An important implication will be that the subgame which is played if no player posts a hostage is identical, also in terms of the payoff function, with the underlying PD itself.

(A3) Ex post incentives for cooperation: A player who cooperates in the subgame played in stage 2 of the hostage game does not lose his or her hostage.
This represents the idea that a player posts a hostage ex ante in order to provide an incentive for own cooperation ex post: he or she receives his hostage back if he cooperates later on.

(A4) Player-symmetric arrangements: Assume that player \( i \) has posted a hostage and defects in stage 2 of the hostage game, while player \( j \) has chosen \( H_j \) and \( X_j \). What happens with \( i \)'s hostage after this outcome does not depend on whether \( i = 1 \) or \( i = 2 \).

Thus, e.g., if players \( i \) and \( j \) both post a hostage and defect afterwards, the same thing happens with both of their hostages: they are either both returned, or are both lost, or are both given to some other party.

Hostage institutions that satisfy assumptions A1–A4 can be uniquely characterized in a simple way. Consider the 8 paths \( Z \) of the hostage game where at least one player \( i \) has posted a hostage and defected. According to A1–A4, hostage institutions can differ only according to what happens with player \( i \)'s hostage after these paths have been played. We classify the paths according to the number of players who posted a hostage and the number of cooperators. We thus get four sets:

\[
\begin{align*}
z_1 &= \{(H_i^+, H_j^+, D_i, C_j)\}, \\
z_2 &= \{(H_i^+, H_j^+, D_i, D_j)\}, \\
z_3 &= \{(H_i^+, H_j^-, D_i, C_j)\}, \\
z_4 &= \{(H_i^+, H_j^-, D_i, D_j)\}.
\end{align*}
\]

A hostage institution according to A1–A4 is defined by its arrangement for \( i \)'s hostage associated with each of these four classes of paths: An institution specifies for each class \( z_k \) whether \( i \) does or does not lose his or her hostage and whether the hostage, if lost, is given to another player or to an outside party. Let \( a_k \in \{B, O, L\} \) denote the arrangement for \( z_k \). A complete hostage institution, then, is a 4-tuple \( l = (a_1, a_2, a_3, a_4) \). Our analysis refers to the set of \( 3^4 = 81 \) hostage institutions \( I \) that satisfy A1–A4.

We focus on a generic set of hostage games. Compare the descriptions of an outcome of a hostage game for player \( i \) in terms of three criteria: (i) whether \( i \) has posted a hostage, (ii) which of the hostages \( i \) owns after the game, and (iii) the strategy combination \( X \) that was chosen in the subgame on stage 2. Then, generically, player \( i \) is assumed never to be indifferent between two outcomes that differ with respect to at least one of these criteria. E.g., given that \( i \) did not post a hostage while \( j \) has posted a hostage, player \( i \) can not be indifferent between the outcome (1) when player \( i \) cooperates while player \( j \) defects and loses his hostage to player \( i \), and the
outcome (2) when both cooperate and the hostage is given back to player \( j \). Indifference between outcomes occurs only if 'end states' are identical in terms of the three criteria. Thus, a player is indifferent between outcome (2) and outcome (3) when both \( i \) and \( j \) cooperate without prior hostage posting of \( i \) or \( j \).

Consider now a solution concept for hostage games which consists of three elements:

(i) Selten's (1965) subgame perfect equilibrium (spe) as the standard refinement of the original 'static' Nash equilibrium concept.

(ii) Payoff-dominance of a spe over other spe's.

(iii) Maximin properties of spe's.

Of course, our central focus is on those hostage institutions \( I \) that provide proper incentives for hostage posting and mutual cooperation if appropriate restrictions for the values of hostages are taken into account. In other words, we wish to specify conditions for hostage games such that hostages are posted and players cooperate on the equilibrium-path (e-path). The following theorem provides the main result.\(^5\)

**Theorem [Weesie and Raub (1992)].** Generically, the hostage game has the following properties:

(i) There exists a spe \( s^* = (s^*_1, s^*_2) \) such that both players post hostages and cooperate on the e-path if and only if

\[
a_1 \neq B,
\]

and for \( i = 1, 2 \)

\[
K_{ii} > T - R \quad \text{and} \quad S < u_i(H^+_i, H^-_j, D_i, D_j).
\]

(ii) There exists at most one spe such that both players cooperate surely on the e-path.

\(^5\)See Weesie and Raub (1992) for a more general theorem on hostage games for (possibly asymmetric) PDs where hostage posting is associated with transaction costs. Also, that paper provides results on hostage posting as a mechanism of cooperation for 2-and \( n \)-person games that differ from the PD, as well as analyses of sequential and unilateral hostage posting. The proof of the theorem used here is somewhat tedious, but straightforward. The crucial part are the conditions in part (i) for a spe \( s^* \). These follow from an analysis of the equilibria in the four subgames that can emerge after stage 1. Trivially, mutual defection is the unique equilibrium in the subgame where no player has posted a hostage. Also, given our generic conditions, (1) and the first part of (2) are necessary and sufficient for an equilibrium in cooperative strategies in the subgame where both players have posted a hostage. This equilibrium payoff dominates all other equilibria in that subgame, if other equilibria exist. Given the first part of (2), both subgames have a unique equilibrium where one player has posted a hostage and the other did not. The equilibrium payoff in these subgames for the player who posted the hostage is smaller than \( R \). Part (2) guarantees that the equilibrium payoff for the other player in these subgames is likewise smaller than \( R \). Parts (ii) and (iii) of the theorem follow immediately. Part (iv) on maximin conditions for reactive strategies can be established by considering that \( P \) is the maximin payoff and by specifying the necessary and sufficient conditions so that strategies \( s^*_i \) have a security level \( P \).
(ii) The strategy $s^*_i$ has the following properties:
- The hostage is posted.
- $C_i$ is played in the subgame where both players have posted a hostage.
- $D_i$ is played in the subgame where no player has posted a hostage.
- Depending on the parameters of the game, $D_i$ is played surely or at least with positive probability in the other two subgames that can be played on stage 2. Thus $D_i$ is played at least with positive probability particularly in that subgame where $i$ has posted a hostage, while $j$ did not post a hostage.

(iii) The strategy $s^*$ payoff dominates all other strategies.

(iv) The strategy $s^*$ is also a maximin point if and only if the following conditions hold for $i = 1, 2$:
\[
\begin{align*}
    a_1 & = 0 \quad \text{and} \quad a_4 = B, \\
    K_{ij} & > P - S, \\
    T - R & < K_{ii} < T - P \quad \text{if} \quad a_3 \neq B.
\end{align*}
\]

The theorem confirms our intuition that it can be individually rational (self-enforcing) to incur commitments via hostage posting and that hostage posting can stabilize cooperation in problematic social situations. The minimal institutional requirement is that a player loses his hostage if both players have posted one and if a unilateral defection occurs on stage 2 (i.e., $a_1 \neq B$). Furthermore, the hostages should be valuable enough so that cooperation becomes an individually rational strategy in the subgame that is played after mutual hostage posting. The critical level for the hostage value is the cost of cooperation $T - R$ in the PD. Also, if $i$ posted a hostage while $j$ did not, it must be individually rational for $i$ not to accept a unilateral defection by $j$ in the subsequent subgame. Therefore, a unilateral defection by $j$ in that subgame has to be costly for $i$ (i.e., $S < u_i(H_i^+, H_j^-, D_i, D_j)$).

The strategy $s^*$ consists of reactive strategies. The players condition their behavior in stage 2 on their partner’s hostage posting. The equilibrium strategies contain the implicit promise of own cooperation after hostage posting of the partner. At the same time, they implicitly threaten a partner who does not post a hostage with (a sufficiently high probability of) own subsequent defection. Of course, the promises are mutually credible due to the Nash equilibrium property of $s^*$, while subgame perfection of the equilibrium guarantees the mutual credibility of threats.

A strategy $s^*$ in reactive strategies need not be unique. In general, there will be other strategies, too. E.g., for many combinations of parameters, a strategy exists such that no player posts a hostage and both defect subsequently. Thus, a strategy $s^*$ in reactive strategies is not necessarily the obvious candidate for the solution of a hostage game. Problems of tacit strategy coordination and equilibrium
An important additional feature of reactive strategies are maximin properties. To appreciate that, notice that a player using a reactive strategy may be vulnerable in a specific sense. First, a player $i$ using a reactive strategy may encounter an opponent $j$ who does not post a hostage and who defects in the subgame that emerges after $i$ has posted a hostage while $j$ did not. In such a setting, the game can end with mutual defection. Under an institution with $a_i \neq B$, $i$ would lose the hostage and end up with a payoff $P - K_{ii} < P$, where $P$ is, trivially, the maximin payoff for the hostage game which $i$ can guarantee by posting no hostage and subsequent defection. Player $i$ may also encounter an opponent who posts a hostage on the first move but defects on the second move. According to $s^*$, $i$ would cooperate against $j$. Under an institution with $a_i \neq O$, $i$ would not receive $j$'s hostage and would end up with a payoff $S < P$. In both cases, reactive strategies do not satisfy the maximin criterion so that a player using these strategies is vulnerable in the sense that he or she may end up with less than the maximin payoff.

One might ask why a rational player $i$ should expect with any positive probability to meet opponents like those in our two examples. Given that reactive strategies are in subgame perfect equilibrium, a rational player $i$ will choose a reactive strategy only if $i$ has sufficiently firm expectations that $j$ will also choose a reactive strategy. It will be difficult for $i$ to entertain such expectations if $i$ believes for one or the other reason that $j$ is an irrational player. For instance, $i$ may believe that there is some positive probability that $j$ does not try to maximize his or her own payoff $u_j$, but tries to minimize $i$'s payoff $u_i$. Given such beliefs and $i$'s problems to predict $j$'s behavior on the basis of such beliefs, choosing a maximin strategy may become a viable option for $i$. In such contexts, maximin properties become relevant because they guarantee that reactive strategies are not vulnerable.

Our theorem provides two crucial conditions for the maximin property of reactive strategies. These refer to institutional arrangements and hostage values. First, a player must receive the other player's hostage after mutual hostage posting and a unilateral defection of the other player (i.e., $a_i = O$). Second, the other player's hostage has to be valuable enough to cover the costs of being unilaterally exploited in the subgame after mutual hostage posting (i.e., $K_{ij} > P - S$). Given that a combination $s^*$ of reactive strategies is not only a subgame perfect equilibrium with payoff dominance but is also a maximin point, it is an obvious candidate for the solution of the hostage game.

A hostage game with incomplete information would be an appropriate model for the explicit analysis of such issues.
3. Descriptive facets of hostage posting: Empirical tests

In the present section, we tentatively interpret the theory as a descriptive one, and test its predictive power against behavioral data. Our goal, thus, is to assess the adequacy of the model in providing satisfactory descriptive approximations. Specifically, the goal is to identify empirically the conditions (as reflected in different scenarios) under which people will tend to post a hostage, and examine the extent to which these tendencies coincide with the theoretical predictions.

Using the game-theoretic model, hostage posting scenarios can be ranked according to how favorable, or less favorable, the conditions are for a rational actor contemplating to apply a reactive strategy. We compare three types of conditions. In the scenario which is least favorable for the use of reactive strategies and, hence, for mutual hostage posting and subsequent mutual cooperation, the reactive strategies are not in subgame perfect equilibrium so that the model predicts no hostage posting. In order to provide for a strong test of the model, it is reasonable to turn, more specifically, to scenarios where reactive strategies are in simple Nash equilibrium but the equilibrium is not subgame perfect. Thus, in such a scenario, the implicit promises to cooperate against a hostage posting partner are credible, while the implicit threats to sanction a partner who does not post a hostage are not credible. If credibility of threats and subgame perfection of equilibria has empirical impact, players should refuse to post hostages under these scenarios. A scenario with more favorable conditions for the application of reactive strategies is one where these strategies are in subgame perfect equilibrium, so that promises and threats are credible, but where reactive strategies do not have the maximin property, so that a player using a reactive strategy is vulnerable. Finally, the scenario with most favorable conditions for hostage posting and cooperation is one where reactive strategies are not only in subgame perfect equilibrium but are also a maximin point.

We present two experiments. In each experiment, three scenarios according to the three different conditions for the use of reactive strategies are studied in a comparative perspective. The experiments employed a similar paradigm. Therefore, we first describe the general setup, followed by a detailed description of both experiments.

3.1. Method

General design: Each of the experiments consisted of two parts: A simple game, played once, which was a PD in terms of monetary outcomes, followed by a hostage game. In this latter game, subjects had an opportunity to pass a predetermined sum of money as a hostage to the Experimenter. The main purpose of the first part, namely playing the PD once, was to
familiarize subjects with the task. Another goal was to test whether subjects' behavior is consistent with an auxiliary assumption that is needed for testing the model. In particular, it is assumed in the present context that utility functions of subjects correspond to their own monetary outcomes (as given in the payoff matrix). A rational subject, whose utility function is compatible with this assumption, should defect in the first game.

The behavior of a subject who prefers to cooperate in the first part, may be accounted for in different ways: Such a subject may have a different utility function, for instance one that is increasing in both own payoffs and payoffs of the other actor (which amounts to some form of altruism). Alternatively, such a subject may simply not be a rational actor. The purpose of the first part was thus to distinguish between those subjects whose behavior is eventually consistent with the underlying assumptions (regarding the utility function and the model), and those whose behavior is not. The succeeding analyses are focused on those subjects who defected on the first part of the experiment. However, as will become evident, the pattern of results remains remarkably stable if the analyses are extended to all subjects, regardless of their decision on the first part of the experiment.

The second part consisted of the hostage game, and its purpose was to test the extent to which choices of subjects were consistent with those derived from the theory. The predictions obtained from different scenarios were tested by varying the rule that specified the provisions under which a hostage was returned or lost, the size of the hostage, and whether a lost hostage was given to the other side (the partner) or to a third party.

General procedure. Subjects were undergraduate students at the University of Utrecht. Each subject was tested individually and was told that the experiment involved two persons. In fact, there was no other subject and the choices that were supposedly made by the other subject were predetermined by the Experimenter and were the same for all subjects in a given condition. The moves of the 'other' person were always those of a rational actor whose behavior is consistent with the theory as developed in the first part of this paper. Subjects were initially paid the sum of 5 Dutch guilders (approximately $2.5) in cash. They were told that additional money could be earned during the experiment. Subsequently, subjects received the first set of instructions.

The instructions for the first part of the experiment, introduced the subject to a version of the PD in the form of a card game. There were two cards, each with a different color (e.g., Blue and Green), and a subject had to choose between one of the two. One of the colors was representing cooperation and the other defection, but care was taken to avoid these terms through the entire experiment. Subjects were never given a payoff matrix in the form presented in fig. 1. Rather, each cell and its corresponding payoffs were presented separately.
Subjects were explicitly instructed to try to gain as much money as possible. They were told that the ‘other’ person received identical instructions, and that each person had to make an independent choice.

Following the completion of a questionnaire designed to ensure that subjects understood the instructions, subjects made their selection of card (i.e., one of the two colors) and provided written reasons for their choice. Independent of the subject’s choice, a subject was always informed that the other person had chosen the card associated with defection. Thus, the first part of the experiment provided also a feedback for subjects which was consistent with the belief that the other person did indeed consider the situation as a PD. Depending on subjects’ choice, they received an immediate cash payment (according to the payoff matrix that was used in the particular experiment), and were then handed the instructions for the second part of the experiment.

The instructions of part two introduced the subject to the option of hostage posting. The size of the hostage, and the specific rule under which the hostage was returned or lost (including whether a lost hostage was given to the other party or not), differed from one experiment to the other. To insure good comprehension of the instructions, subjects were given a short questionnaire that tested whether they were able to calculate the appropriate payoffs for different possible outcomes.

Following the completion of the questionnaire, subjects decided whether or not they wanted to post a hostage, and gave written reasons for their choice. After reporting their decision to the Experimenter, they were informed about the decision of the other player. Subjects who opted for posting a hostage, were required to deposit the hostage with the Experimenter.

Finally, subjects were required to play again the card game, namely to choose one of the two cards associated with cooperation or defection respectively. They were then told about the decision of the ‘other’ person and were paid according to the appropriate payoff matrix.

Depending on their choices, subjects earned at least 5 Dutch guilders and at most (depending on the group to which they were assigned) 19 Dutch guilders. For our student population, the maximal payments represented quite considerable incentives for a task which took between 30 and 45 minutes on the average. Furthermore, there is a considerable difference between high and low payments. Most of the subjects were obviously motivated to take their task seriously.

3.2. Experiment 1

Design and procedure: Three groups were employed in the first experiment. Subjects in each group were presented in the first part of the experiment with an identical payoff matrix such that $T = 10$, $R = 8$, $P = 4$, and $S = 2$ guilders.
In the second part of the experiment, subjects were given the option to post a hostage worth 4 guilders. The rules under which the hostage was returned or forfeited, differed for the three groups. Group 1 was used for testing conditions where reactive strategies are associated with credible promises but not with credible threats, and where reactive strategies are furthermore vulnerable. This group received instructions that were associated with an institution $I = (L, L, L, L)$, and read as follows:

Rule 1: An actor who deposits a hostage receives it back under the condition that in the second stage, he or she have chosen the cooperative move (independent of the other actor's decision). Otherwise the hostage is lost.

Under rule 1, reactive strategies $s_i^*$ are in Nash equilibrium that prescribe to defect in the subgame where $i$ has posted a hostage while $j$ did not. Thus, the implicit promises associated with reactive strategies are credible. Cooperation is even a strictly dominant strategy for both players in the subgame that emerges after both players have posted a hostage. However, the equilibrium is not subgame perfect. This can be easily seen by considering the subgame that is played after $i$ has posted a hostage while $j$ did not. This subgame has a unique equilibrium $(C_i, D_j)$ in dominant strategies. Hence, reactive strategies that require that $i$ defects in this subgame (at least with positive probability) are never associated with best-reply strategies for $i$ in this subgame: the implicit threat is not credible.

Group 2 was used for testing conditions where reactive strategy are again vulnerable, but contain both credible promises and credible threats. The instructions for this group were associated with the institution $I = (L, B, L, B)$, namely:

Rule 2: An actor who deposits a hostage receives it back only if he or she have chosen the cooperative move, or if both actors have chosen to defect. Otherwise the hostage is lost.

Given the specification of the payoffs and the hostage values, reactive strategies fulfill the conditions for a spe according to our theorem. Finally, group 3 received the following instructions that correspond to the institution $I = (O, O, B, B)$:

Rule 3: If one actor posts a hostage and the other one does not, then the hostage is immediately returned before the second part of the game. Given that both actors post a hostage, the hostage is returned to an actor under the condition that on the subsequent move the actor has chosen the cooperative move (independent of the other actor's decision). Otherwise the hostage is lost, and given to the other player.

Under this rule, a lost hostage is transferred to the other actor, and unilateral hostage posting is not allowed. As can be seen from the theorem, reactive strategies do now not only contain credible promises and credible threats but are also maximin strategies and thus invulnerable.
Following their decision, all subjects in group 1 (regardless of whether they decided to post a hostage or not) were told that the 'other' person did not post a hostage. Similarly, all subjects in groups 2 and 3 were told that the 'other' person did post a hostage. With the exception of the rule and the feedback regarding the other subject's decision, the three groups received identical treatments.

Several hypotheses can be formulated with regard to experiment 1. First, the theory predicts that no hostage posting, followed by defection on the subsequent move, will occur under the conditions of group 1. Furthermore, there is a clear prediction for group 3: Subjects will post a hostage and will cooperate (given the feedback that the 'other' person also posted a hostage). Concerning group 2, two alternative conjectures can be stated: According to the strong hypothesis, all subjects should choose the reactive strategy. Thus, they should post a hostage, and subsequently choose for the cooperative move, given that the 'other' party also posted a hostage. Such a prediction, however, is based on several auxiliary assumptions (besides the one mentioned above on the utility function of subjects), namely, an assumption of tacit strategy coordination, and mutual expectations of rational behavior. If, however, these assumptions, and particularly the last one, are not satisfied, then even rational subjects may decline from posting a hostage. Thus, presupposing that the relevant auxiliary assumptions apply at least to a significant number of subjects, one gets a weak version of the hypothesis, namely, that a significant number of subjects will post a hostage.

It is important to emphasize that the predictions of the model always refer to a combination of two acts: The act of posting or not posting a hostage, and the subsequent act of cooperation or defection. For example, for group 1 as described above, the model predicts refusal to post a hostage followed by defection. Similarly, the model's predictions for group 2 are that subjects who will post a hostage will subsequently cooperate (given a feedback that the other subject has also posted a hostage). The behavior of subjects in all experiments was remarkably consistent: With very few exceptions (5 out of 146 subjects, i.e., less than 3.5%), all subjects' decisions with respect to cooperation or defection were consistent with their decisions on hostage posting (and the feedback concerning the other person's first move). Consequently, we treat as 'noise' those few cases in which decisions to cooperate or defect were incompatible with the decisions on hostage posting.

Another complication arises in the comparison of two groups that received different feedback after the hostage posting decision. For example, it is difficult to compare the difference in the number of subjects who posted a hostage and subsequently cooperated, in groups 1 and 2: The reason is that subjects in group 1 were informed that the other subject declined to post a hostage, whereas subjects in group 2 were informed that the other subject posted a hostage. This different feedback may differentially affect subjects'
Table 1

Number of subjects (out of those defecting in part 1) posting a hostage in experiment 1 (numbers in parentheses include subjects who chose to cooperate in part 1).

<table>
<thead>
<tr>
<th>Group</th>
<th>Condition</th>
<th>Posting hostages</th>
<th>Not posting hostages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Only promise is credible</td>
<td>3(^a) (4(^a))</td>
<td>15 (19) (^b)</td>
</tr>
<tr>
<td>2</td>
<td>Promise and threat are credible</td>
<td>6(^a) (11) (^a)</td>
<td>9 (11)</td>
</tr>
<tr>
<td>3</td>
<td>Promise and threat are credible and coinciding with a maximin criterion</td>
<td>18 (25)</td>
<td>1 (3)</td>
</tr>
</tbody>
</table>

\(^a\)One subject did not cooperate after posting a hostage.
\(^b\)One subject did not defect after refusing to post a hostage.

decision on the succeeding move. To simplify matters, and to circumvent potential confoundings, we focus on proportion of subjects posting a hostage (ignoring the subsequent decision of cooperation or defection). Whenever relevant, however, we do report on those few subjects who deviate from the predictions on the second stage.

**Results:** Combining subjects from all groups, 74% chose defection on the first part of the experiment, slightly higher than the mean reported in previous studies [see Colman (1982) for a review]. Same proportions were obtained in the second experiment.

Number of subjects posting a hostage in each of the groups is presented in table 1 (numbers refer to subjects who defected on the first part; numbers in parentheses include all subjects regardless of their move in the first part). Unless otherwise specified, all subjects who posted a hostage cooperated on the subsequent move, and all subjects who declined to post a hostage defected on the subsequent move. Under rule 1, the large majority of the subjects (regardless of whether they defected or cooperated in the first part of the experiment) declined to post a hostage as predicted by the theory. Likewise, with very few exceptions, subjects (regardless of their choice on the first part of the experiment) decided to post a hostage under rule 3. Evidently, favorable conditions for hostage posting that coincide with a maximin criterion result in hostage posting.

Under rule 2, half of the subjects have posted a hostage (regardless of whether we consider the entire sample, or only subjects who defected on the first part). One subject posting a hostage did not cooperate on the subsequent move. The results of group 2 are clearly compatible with the weak but not with the strong hypothesis. Using Fisher's exact test [e.g., Everitt (1978)], the proportion of subjects posting a hostage in group 2 was
significantly larger ($p < 0.05$) than the proportion of subjects posting a hostage in group 1, congruent with the weak hypothesis.

Discussion: Experiment 1 suggests that a credible promise (i.e., an equilibrium in reactive strategies) and a credible threat (i.e., a subgame perfect equilibrium in reactive strategies) are necessary but not sufficient conditions for inducing subjects to post a hostage. To ensure that all subjects will post a hostage, the conditions have also to satisfy a maximin criterion.

The results of group 2 support the weak (but not the strong) prediction. An immediate problem to be raised then is why did a substantial number of subjects in condition 2 decline to post a hostage? In particular, one might ask whether subjects in group 2 were indeed searching for a certain level of security which led them to reject the hostage posting option and to favor maximin behavior. This conjecture gets some support from analyzing the written explanations of subjects for their decision to post a hostage or not. The two authors of this paper, evaluated independently (according to criteria established in advance) whether the reasoning of subjects was mainly based on considerations of safety. Of the 9 subjects who refused to post a hostage (and had defected in the first part of the experiment), 6 were judged to rely mainly on reasons of safety (e.g., I do not know the other player's motives and reasoning; if I do not post a hostage I am certain to win at least 4 guilders (an explicit maximin statement); I want to play it safe). None of the subjects who posted a hostage mentioned any such consideration. There was a high inter-rater reliability among the two authors (initial agreement with regard to 16 out of 18 responses).

Two related questions may be raised concerning the empirical evidence: First, it may be claimed that although the results are unambiguous, the power of our statistical tests [Cohen (1988)] is rather low, due to the relative small number of subjects employed in each condition. Second, one may question whether the results can be generalized beyond the particular conditions that were used (i.e., the specific payoff matrix, and the rules used in experiment 1). The purpose of experiment 2 was to replicate the first experiment, and at the same time to introduce some modifications in order to test the generality of the earlier results.

### 3.3. Experiment 2

Design and procedure: Three groups were employed in experiment 2. All three groups were first exposed to a PD with payoffs $T = 12$, $R = 10$, $P = 4$, and $S = 0$ guilders, that was also used in the second part of the experiment.

The differences in conditions between groups in experiment 1 were primarily obtained by varying the specifications under which a hostage is returned or declared forfeited. In experiment 2 these differences were mainly achieved by varying the size of the hostage. Consider the following rule
Rule 4: An actor who deposits a hostage receives it back under the condition that in the second stage, he or she have chosen the cooperative move (independent of the other actor's decision). Otherwise the hostage is lost, and given to the other party.

Rule 4 is identical to rule 1 except that, unlike in rule 1, a lost hostage is given to the other actor. Group 1 in the present experiment was exposed to rule 4 and the option to post a hostage of 5 guilders. As in group 1 of experiment 1, only the promise is credible under those conditions, and the model predicts no hostage posting followed by defection. Consequently, subjects in this group were always told that the other party did not post a hostage.

Group 2 was also exposed to rule 4, but the size of the hostage was reduced to 3 guilders. As can be seen immediately from our theorem, this combination corresponds to a scenario where reactive strategies are in subgame perfect equilibrium but are not maximin. As in group 2 of experiment 1, both the promise and the threat are credible, but a player who posts a hostage is vulnerable. All subjects in this group were told that the other party has posted a hostage.

Group 3 was exposed to rule 3 (as in experiment 1), except that the hostage in this case was 5 guilders. The conditions correspond to a scenario where reactive strategies are in subgame perfect equilibrium and are likewise maximin, so that the model predicts hostage posting followed by cooperation. Accordingly, all subjects in this group were informed that the other actor has posted a hostage.

The procedure in experiment 2 was further identical to the one used in the first experiment.

Results: Results are presented in table 2. The pattern of results for groups 1 and 3 is very similar to the corresponding groups in experiment 1. The proportion of subjects not posting a hostage in group 1 (0.89) was slightly larger than in the comparable group (0.83) in experiment 1. The proportion of subjects posting a hostage in group 3 (0.89) was slightly smaller than in the comparable group in experiment 1 (0.95). The pattern of results is the same regardless of whether the analyses are performed on all subjects, or are restricted only to those who defected on the first part of the experiment.

Unlike in experiment 1 (and independently of whether all subjects are analyzed or only those who defected on the first part of the experiment), the number of subjects posting a hostage in group 2 was not significantly larger than in group 1. Like in experiment 1 (group 2), however, most of the subjects refusing to post a hostage justify it in terms of vulnerability concerning the other person, and their desire to play it 'safe'. An analysis, similar to the one conducted in experiment 1, was performed on subjects' explanations for their responses and showed a very similar pattern: Out of 13
Table 2
Number of subjects (out of those defecting in part 1) posting a hostage in experiment 2 (numbers in parentheses include subjects who chose to cooperate in part 1).

<table>
<thead>
<tr>
<th>Group</th>
<th>Condition</th>
<th>Posting hostages</th>
<th>Not posting hostages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Only promise is credible</td>
<td>2</td>
<td>16&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>(20)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>2</td>
<td>Promise and threat are credible</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8)</td>
<td>(16)</td>
</tr>
<tr>
<td>3</td>
<td>Promise and threat are credible and coinciding with a maximin criterion</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(21)</td>
<td>(4)</td>
</tr>
</tbody>
</table>

<sup>a</sup>One subject did not cooperate after posting a hostage.
<sup>b</sup>One subject did not defect after refusing to post a hostage.

subjects who refused to post a hostage (and who had defected on the first part of the experiment), 8 were judged to rely mainly on reasons of safety. None of the subjects who posted a hostage mentioned any such consideration.

In order to obtain more power for the analysis, we may combine the data from experiment 2 with the corresponding conditions in experiment 1. Although experiment 2 used slightly different specifications as well as a different payoff matrix, the conditions that were collapsed were corresponding to strategically identical scenarios: Group 1 in experiments 1 and 2 were exposed to conditions of an equilibrium in reactive strategies without subgame perfectness (credible promise only); group 2 in both experiments to a subgame perfect equilibrium in reactive strategies (promise and threat credible), and group 3 in both experiments to the condition where hostage posting also coincided with a maximin criterion.

The data of the combined results were submitted to an arcsine transformation, and a Chi-square test of homogeneity of proportions [e.g., Marascuilo and McSweeney (1977)] was highly significant for both the analysis restricted to subjects who defected on the first part of the experiment ($p<0.005$), as well as for the entire population of subjects ($p<0.005$). Employing multiple comparisons for proportions based on the arcsine transformation, there was a significant difference ($p<0.05$) in hostage posting between all three pairs of comparisons (again, this was true for both subjects who defected on the first part, as well as for all subjects).

The combined data may be used for an additional analysis. It should be realized that the model's predictions regarding the extreme conditions (not posting a hostage if the threat is not credible, and posting a hostage if the maximin criterion is also satisfied), are deterministic. With very few excep-
Wi Raub and G. Keren, Hostages as a commitment device

Table 3

<table>
<thead>
<tr>
<th>Group</th>
<th>Model's prediction</th>
<th>Observed proportion of the model's predictions</th>
<th>Confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Exp. 1)</td>
<td>No hostage posting</td>
<td>0.833</td>
<td>0.761-0.905</td>
</tr>
<tr>
<td>3 (Exp. 1)</td>
<td>Hostage posting</td>
<td>0.947</td>
<td>0.875-1.00</td>
</tr>
<tr>
<td>1 (Exp. 2)</td>
<td>No Hostage posting</td>
<td>0.888</td>
<td>0.817-0.961</td>
</tr>
<tr>
<td>3 (Exp. 2)</td>
<td>Hostage posting</td>
<td>0.888</td>
<td>0.817-0.961</td>
</tr>
</tbody>
</table>

The few exceptions (8 out of 73 subjects in the four groups) can thus be considered as random noise. Assuming that this level of random noise is the same for all four groups, we may calculate a common variance term based on the combined data from all four groups, and correspondingly the 95% confidence intervals for each of the four groups. These are given in table 3. The proportions and corresponding confidence intervals are based only on subjects who defected on the first part of the experiment. A very similar pattern is obtained when subjects who cooperated on the first part are included.

4. Concluding comments

In the first part of this paper a theory was outlined that identifies conditions under which hostage posting is a favorable option for rational actors in order to achieve cooperation in problematic situations. In the second part of the paper the descriptive adequacy of the model was tested. Generally, the predictions derived from the model were quite successful. Thus the theory can serve as a useful framework for explaining behavior regarding hostage posting.

The strongest predictions made by the theory concern the extreme cases under which hostage posting should and should not occur, respectively; these predictions have been robustly corroborated by the empirical tests. Specifically, the theory predicts that no hostage posting should take place unless both the promise and the threat implicit in reactive strategies are credible, or alternatively, unless these strategies are in subgame perfect equilibrium. Virtually all subjects in experiments 1 and 2 (condition 1) refused to post a hostage, thus confirming the model's prediction at this end. This result might be interpreted as a rather strong confirmation of the empirical relevance of the subgame perfectness of equilibria. At the other end, the theory predicts unequivocally that hostages should be posted under maximin conditions,
where in addition to the credibility of the promise and the threat the actor is in fact invulnerable to the other partner's decisions. Again, the pattern of results in experiments 1 and 2 (group 3) is in excellent agreement with those predictions. This supports the empirical significance of maximin properties of strategies.

The predictions of the model with regard to conditions that lie between the two extremes require additional assumptions. Conditions 2 in experiments 1 and 2 can be interpreted as examples of testing the extent to which those additional assumptions are eventually met. The results of group 2 in experiments 1 and 2, in which the conditions for hostage posting are favorable, but require additional assumptions (particularly with regard to beliefs about the partner's behavior) are revealing. On one hand, there were clearly more subjects willing to post a hostage compared with condition 1. On the other hand, it is also apparent that a substantial number of subjects refused to post a hostage under those conditions.

It seems plausible to conclude that whether hostage posting is induced, depends on complex interactions of conditions (e.g., value of hostage, rules for hostage being returned, etc.) that are well captured by the model. Notwithstanding, several questions still remain open. One particularly interesting finding from our empirical research is that subjects are often risk averse in the sense that they prefer the sure thing, and are reluctant to post a hostage unless maximin conditions are satisfied for reactive strategies. While the maximin criterion seems to be a sufficient condition for inducing hostage posting, the question remains whether it is also a necessary one. Hostage games with incomplete information may provide attractive models for an analysis of this problem.

We have stressed the importance of different institutional frameworks for hostage posting. Also, we have suggested to distinguish between the collective rationality of hostage posting ('Do cooperation and efficiency emerge if hostages have been posted?') and the individual rationality of using this mechanism ('Do actors have sufficient individual incentives to post hostages?'). It was argued that the rationality of hostage posting, and in particular its individual rationality, may depend on specific attributes of (institutional arrangements for) hostages. Our analysis has borne out, theoretically as well as empirically, both of these considerations. Compare the scenarios for groups 1 with those for groups 3 in both experiments. In all of these scenarios, cooperation is a strictly dominant strategy for both players in the subgame that emerges after both have posted a hostage. Therefore, rational actors should cooperate under all of these scenarios if such hostages have been posted. The results for groups 3 show that cooperation does indeed emerge empirically if hostages have been posted. However, our theory predicts that it is far from self-evident that subjects would indeed post such hostages. More specifically, the theory predicts that
such hostages will not be posted under the scenarios for groups 1, but that they will be posted under the scenarios for groups 3. Again, our empirical results confirm the prediction. This shows that a careful design of institutional arrangements for hostage posting is required, even if hostages posting techniques are readily available.

This paper has been restricted to simultaneous hostage posting for a symmetrical Prisoner's Dilemma. Extensions could refer to sequential hostage posting as well as asymmetric PDs. Likewise, it would be interesting to find out whether our results can be generalized to other 2-person dilemma situations that differ from the PD. Further generalizations could be obtained by extending the model and the corresponding empirical tests to n-person games. Theoretical research [see Weesie and Raub (1992)] as well as experimental research [see Keren and Raub (1992)] concerning such extensions has begun recently and may serve as an adequate test for the robustness of our findings. Via such extensions, better adapted models could be provided for analyzing in a systematic and sufficiently detailed manner different social situations such as those described in the introduction.

We would like to add a more speculative remark. Generally speaking, hostage posting is a means of changing the (physical) outcomes associated with certain strategy combinations in a problematic situation. Via changing physical outcomes, players are also able to indirectly change the preferences associated with certain strategy combinations in the problematic situation, although preferences with respect to physical outcomes themselves remain constant. It is then instructive to imagine other settings where physical outcomes associated with strategy combinations remain constant, while preferences for outcomes themselves change. Players may be able – at least to some degree – to directly modify their preferences by, e.g., adopting 'altruistic' or 'moral' preferences [see, e.g., Sen (1974), Frank (1988)]. The model outlined in this paper may be used as a starting point to show that, given opportunities for direct preference modifications, such preference modifications can be individually rational. The model can thus be reinterpreted as a model of endogenous preference changes [see Raub (1990)].

References
Cohen, Jacob, 1988, Statistical power analysis for the behavioral sciences (Lawrence Erlbaum, Hillsdale, NJ).
W. Raub and G. Keren, Hostages as a commitment device 67


Luce, R. Duncan and Howard Raiffa, 1957, Games and decisions (Wiley, New York).


