

**ECONOMIC RESEARCH REPORTS**

THE POINT OF NO RETURN:  
A GAME-THEORETIC ANALYSIS  
OF CRISIS STABILITY

by

Steven J. Brams

and

D. Marc Kilgour

R.R. #86-03

February 1986

**C. V. STARR CENTER  
FOR APPLIED ECONOMICS**



**NEW YORK UNIVERSITY  
FACULTY OF ARTS AND SCIENCE  
DEPARTMENT OF ECONOMICS  
WASHINGTON SQUARE  
NEW YORK, N.Y. 10003**

THE POINT OF NO RETURN: A GAME-THEORETIC ANALYSIS OF CRISIS STABILITY

Steven J. Brams\*  
Department of Politics  
New York University  
New York, NY 10003

D. Marc Kilgour\*\*  
Department of Mathematics  
Wilfrid Laurier University  
Waterloo, Ontario  
Canada N2L 3C5

\* Steven J. Brams gratefully acknowledges the financial support of the Ford Foundation under Grant No. 845-0354 and the National Science Foundation under Grant No. SES84-08505.

\*\* D. Marc Kilgour gratefully acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada under Grant No. A8974.

**ABSTRACT**

A model of crisis stability is developed based on a deterrence game defined on the unit square. In this game, the players are assumed to be able to choose any level of initial cooperation or noncooperation; the more cooperative player (if there is one) may then retaliate if he chooses.

In a crisis the players are assumed to have escalated their conflict and desire to stabilize it before it explodes. Their payoffs in this game, as in the original deterrence game, are based on those in Chicken but now depend on how noncooperative the players are in the crisis.

Conditions are found under which each player can, by threatening his opponent, stabilize the crisis (i.e., make it a Nash equilibrium) without necessarily returning to full cooperation in the original game. In addition, points of no return are identified at which neither player can prevent his opponent from preempting further without threatening him at the point where the crisis has already carried the players.

Crisis stabilization is aided by being close to the full-cooperation position, which can always be stabilized by threats. Also, if one player is much more cooperative in the crisis, or values the cooperative outcome relatively highly, crisis stabilization is facilitated. Possible ways of avoiding points of no return, especially in conflicts between the superpowers, are discussed.

## THE POINT OF NO RETURN: A GAME THEORETIC ANALYSIS OF CRISIS STABILITY

Steven J. Brams and D. Marc Kilgour

### 1. Introduction

Prominent in the lexicon of nuclear strategies is the notion of "crisis stability," which means roughly the ability of decision makers in a dispute that escalates to crisis proportions to prevent that crisis from exploding. The principal means to avoid an explosion is to ensure that neither side has an incentive to preempt the other in the crisis.<sup>1</sup>

Nuclear deterrence between the superpowers, as long as it is assured by a second-strike capability, would appear to provide the necessary insurance. What worries the political leaders of each superpower considerably more than a "bolt from the blue" by the other superpower is the possible escalation of a conventional conflict, such as might occur in the Middle East or Western Europe, into a crisis that involves serious threats to their allies or even their own security. Insofar as their ability to stabilize such a conflict lies outside their control, crisis stabilization -- at least on their part -- is jeopardized.

Presumably, decision makers in an escalatory situation would prefer to be able to damp down a conflict, or at least contain it, so that a crisis cannot erupt into full-fledged war in which everybody may suffer egregiously. This is an especially frightening problem in a confrontation between nuclear powers. But even in conventional conflicts, such as that between Iran and Iraq today, the human casualties and material damage on both sides may be horrendous.

The ability to stabilize a crisis or contain a conflict, however, may

be a mixed blessing. For once it is recognized by the protagonists that they may be able to escape disaster, they may be more willing to take risks that they would otherwise avoid if escape from disaster were impossible (Brams and Kilgour, 1985e).

Put another way, crisis stability may encourage provocative behavior, whereas an inability to stabilize crises may induce more cautious choices. In the Cuban missile crisis of October 1962, for example, the leaders of the superpowers assiduously eschewed making explicit nuclear threats as the crisis heightened, perhaps in part because each side's second-strike capability at the time was very limited. In the absence of stabilizing forces, the situation was already extremely delicate, so neither side wanted to do anything to provoke the other side into preempting with nuclear weapons.

Because stability is an eminently game-theoretic concept, it is appropriate to model crisis stability by a game. Accordingly, we shall begin our analysis by setting up a Variable-Retaliatio n Deterrence Game based on Chicken to model deterrence. The so-called Deterrence Equilibrium in this game, as we shall show, may be difficult to sustain if a crisis pushes the players into bellicose postures.

How they can prevent their positions from deteriorating further in a crisis will be one theme of our analysis. Specifically, we shall derive threat functions or lines that illustrate how the players can deter further escalation.

Unfortunately, under certain conditions there may be a point of no return: no threats of reprisal will be able to render this point stable and thereby deter preemption past this point. In effect, the players

will already have gone too far to make a return to mutual cooperation, or even stabilization, worth their while. A second objective of our analysis will therefore be to identify conditions under which deterrence cannot be restored in the Variable-Retaliatio n Deterrence Game after the onset of a crisis. The plausibility of this game as a model of nuclear deterrence, and as a basis for determining when deterrence can be maintained under conditions of crisis, will also be assessed.

## 2. The Variable-Retaliatio n Deterrence Game

The Variable-Retaliatio n Deterrence Game (VRDG) is based on the game of Chicken that, it has been argued, constitutes an appropriate basic model of deterrence between the superpowers (Brams, 1985, ch. 1). This claim has been disputed by Zagare (1985), who contends that Prisoners' Dilemma is a better model.

This dispute, however, is somewhat beside the point, because VRDG adds two significant features to Chicken that make it a truly different game (Brams and Kilgour, 1985a): (i) the players can make quantitative choices of levels of cooperation ( $C$ ) or noncooperation ( $\bar{C}$ ), not just qualitative choices of  $C$  or  $\bar{C}$ ; (ii) once these initial choices, which can be interpreted as levels of nonpreemption or preemption, are made, the less preemptive player (i.e., the player who chose the lower level of preemption initially) can retaliate by choosing a different -- and presumably higher -- level of noncooperation subsequently. The game terminates whenever the initial levels are the same or, if they are not, after the less preemptive player has retaliated.

More formally, VRDG is defined by the following rules:

1. The players do not choose initially between  $C$  and  $\bar{C}$  but instead choose preemption levels  $s$  (Row) and  $t$  (Column) that may range on a continuum between 1 (no preemption) and 0 (maximum preemption). These choices are simultaneous and determine a point on the unit square defined by vertical coordinate  $y$  and horizontal coordinate  $x$  (see Figure 1),

---

Figure 1 about here

---

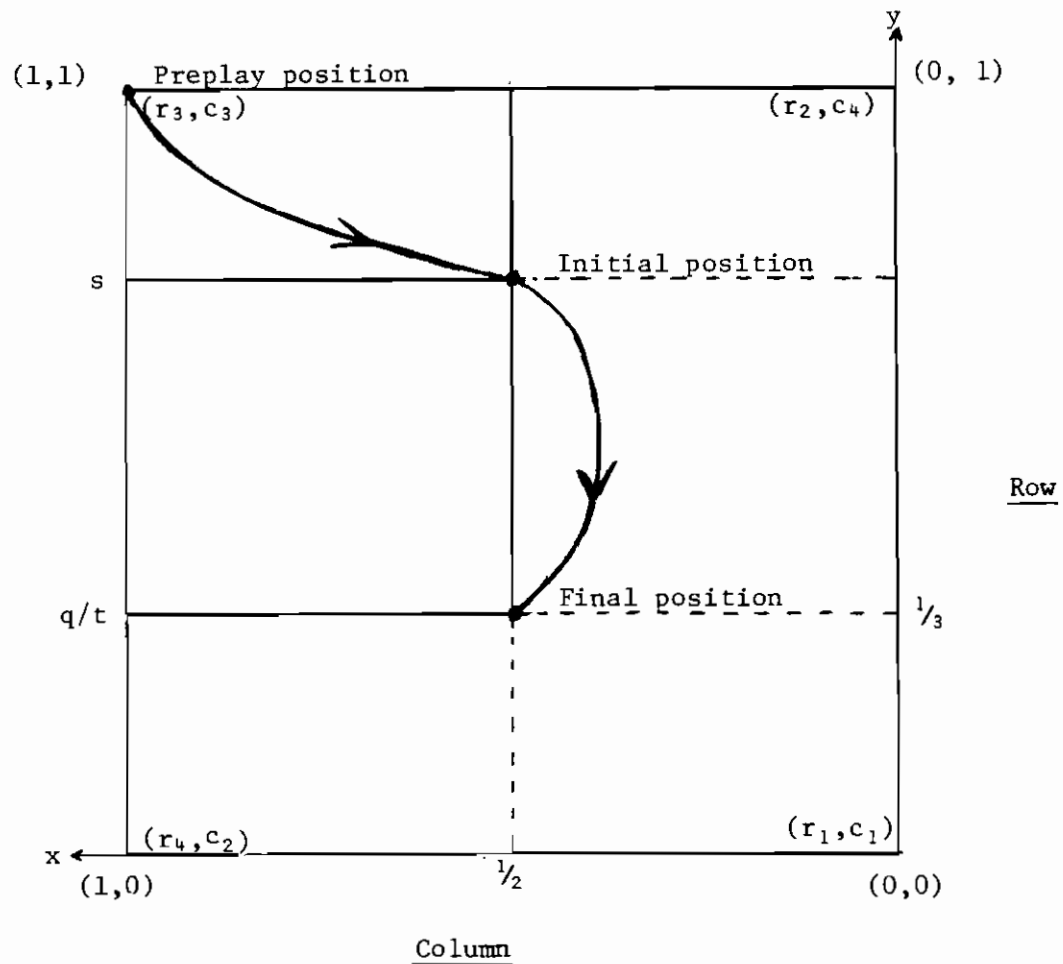
which we call the initial position. It is convenient to assume that, before play commences, the preplay position of the players is  $(1,1)$ , the mutual-cooperation position, but this starting point is not crucial to the analysis of VRDG.

2. The  $2 \times 2$  game of Chicken is embedded in the unit square so that the four corners of the square correspond to the four "pure" states  $(r_i, c_j)$  of Chicken, with  $r_i$  giving the payoff of Row and  $c_j$  the payoff to Column (the subscripts  $i$  and  $j$  give the rankings of payoffs, with 4 being best and 1 being worst):

- i. Both players cooperate (CC) -- next-best outcome for both players:  $(r_3, c_3)$  at upper-left corner in Figure 1.
- ii. One player cooperates and other does not ( $C\bar{C}$  and  $\bar{C}C$ ) -- best outcome for the player who does not cooperate and next-worst for the player who does:  $(r_2, c_4)$  and  $(r_4, c_2)$  at upper-right and lower-left corners, respectively.
- iii. Both players do not cooperate ( $\bar{C}\bar{C}$ ) -- worst outcome for both players:  $(r_1, c_1)$  at lower-right corner.

3. In Figure 1, the square is drawn so that the  $(1,1)$  position, corresponding to the CC outcome of Chicken, is at the upper left and

FIGURE 1  
VARIABLE-RETALIATION DETERRENCE GAME



Key:  $(r_i, c_j) = (\text{payoff to Row, payoff to Column})$

$r_4, c_4 = \text{best}$ ;  $r_3, c_3 = \text{next best}$ ;  $r_2, c_2 = \text{next worst}$ ;  $r_1, c_1 = \text{worst}$

$s, t = \text{initial strategy choices of Row and Column, respectively}$

$q(t) = \text{subsequent strategy choice of Row (more cooperative player initially)}$

and the (0,0) position, corresponding to the  $\overline{CC}$  outcome of Chicken, is at the lower right. Points of the square not at the four corners represent mixtures of two or more outcomes of Chicken.

4. If  $s = t$ , the game terminates at point  $(t,s)$  in the unit square. If  $s > t$ , the more cooperative player, Row, retaliates against Column by changing his level of cooperation, according to the retaliation function  $q(t)$ , to make the final position of the game  $(t,q(t))$ . Row's retaliation function thus depends on Column's initial choice  $t$ ; however, it is the value of  $s$ , as compared with  $t$ , that determines whether Row or Column (if either) can retaliate. If  $s < t$ , Column is the player who is more cooperative initially and thereby can retaliate; Column's level of cooperation changes from  $t$  to  $p(s)$ , where  $p(s)$  is Column's retaliation function, and the final position is  $(p(s),s)$ . The players' retaliation functions are chosen at the same time as their initial choices,  $s$  and  $t$ .

5. Payoffs to Row (R) and Column (C) are defined by both the Chicken payoffs at the four corners of the unit square and the final position  $(x,y)$  of the players on the square (the route by which the final position was attained from the preplay and initial positions is assumed to be irrelevant):

$$P_R(x,y) = xyr_3 + x(1-y)r_4 + (1-x)yr_2 + (1-x)(1-y)r_1$$

$$P_C(x,y) = xyc_3 + (1-x)yc_4 + x(1-y)c_2 + (1-x)(1-y)c_1$$

At the four corners of the unit square,  $P_R$  and  $P_C$  exactly coincide with the corresponding payoffs of Chicken. At any other point on the square,  $P_R$  and  $P_C$  are averages of the four payoffs at the corners, weighted according to the distance of the point from the corner.

It is worth noting that the payoff functions are bilinear (linear

in each coordinate): in calculating payoffs at  $(x,y)$ , the payoffs at each corner are weighted by the product of the distances, parallel to the axes, from  $(x,y)$  to the opposite corner. The four weights so obtained are non-negative and sum to 1, but unlike barycentric coordinates they are not unique. This is because barycentric coordinates would be defined in terms of the four corner points of the unit square, whereas uniqueness is preserved only if three points are used to define coordinates in the 2-space of the unit square.

As an illustration of this weighting system, consider the coordinates of the final position shown in Figure 1, where  $x = \frac{1}{2}$  and  $y = \frac{1}{3}$ . The weights of the four corner values of Chicken in  $P_R(x,y)$  and  $P_C(x,y)$  are:

$$\begin{aligned} (1,1): xy &= \frac{1}{6}; & (1,0): x(1-y) &= \frac{1}{3}; \\ (0,1): (1-x)y &= \frac{1}{6}; & (0,0): (1-x)(1-y) &= \frac{1}{3}, \end{aligned}$$

which sum to 1.

The two lower payoffs in Figure 1,  $(r_4, c_2)$  and  $(r_1, c_1)$ , which correspond to corners equidistant from  $(\frac{1}{2}, \frac{1}{3})$ , are equally weighted by factors of  $\frac{1}{3}$  in  $P_R$  and  $P_C$ ; these factors correspond to distances from  $(\frac{1}{2}, \frac{1}{3})$  in the  $(x,y)$ -coordinate system. Similarly, the two upper payoffs,  $(r_3, c_3)$  and  $(r_2, c_4)$ , which also correspond to corners equidistant from  $(\frac{1}{2}, \frac{1}{3})$ , are equally weighted by factors of  $\frac{1}{6}$ . Not unreasonably, the upper factors have half as much "pull" as the lower factors, for they are twice as distant from  $(\frac{1}{2}, \frac{1}{3})$  on the vertical axis.

Different metrics would simply deform the continuous surface generated by our payoff functions, but these deformations would remain anchored at the four corners of the unit square. They would not alter the basic nature of the Nash (1951) equilibrium results that we shall describe.

An advantage offered by the present formulation, in addition to mathematical tractability, is that the four weights in  $P_R$  and  $P_C$  can be

interpreted as probabilities because they are non-negative and sum to 1. Consequently,  $P_R$  and  $P_C$  may be thought of as expected values, the corner values being the four "pure" states of the game that can arise. As probabilities of being at the four corners, the four factor weights, when multiplied by a player's payoffs at each corner, give the expected payoff at any point of the unit square.

In a non-probabilistic interpretation, the unit square of Figure 1 might be considered a board game. The players choose  $(x,y)$ -coordinates, according to rules 1-4, that define a point on this board. As given by rule 5, the payoffs to the players are the heights of the continuous surfaces, above the  $(x,y)$ -plane, generated by  $P_R$  and  $P_C$ . A Nash equilibrium then has the property that the player who was more cooperative initially cannot increase his payoff by changing his coordinate when his opponent's coordinate is held constant.

VRDG has several qualitatively different Nash equilibria, but only one equilibrium, which we call the Deterrence Equilibrium (Brams and Kilgour, 1985a), involves no preemption:  $(x,y) = (1,1)$ , with payoffs  $(r_3, c_3)$  to Row and Column. Assume henceforth that the payoffs are normalized so that  $r_4 = c_4 = 1$ ,  $r_1 = c_1 = 0$ , and the intermediate payoffs satisfy the following inequalities:  $0 < r_2 < r_3 < 1$  and  $0 < c_2 < c_3 < 1$ . Then the initial strategies and retaliation functions that render this outcome stable are

$$s = 1, q(t) \leq q_1(t); \quad t = 1, p(s) \leq p_1(s),$$

where

$$q_1(t) = \frac{c_3 - tc_2}{1 - t(1 - c_3 + c_2)}, \quad 0 \leq t < 1; \quad p_1(s) = \frac{r_3 - sr_2}{1 - s(1 - r_3 + r_2)}, \quad 0 \leq s < 1.$$

At the Deterrence Equilibrium both players never preempt ( $s = t = 1$ ), but

Row threatens retaliation at level  $q(t) \leq q_1(t)$  if preempted at level  $t < 1$ , and Column threatens retaliation at level  $p(s) \leq p_1(s)$  if preempted at level  $s < 1$ . The "1" subscripts on the right sides of each retaliation-function inequality indicate that the preempted player initially chooses  $s = y = 1$  (Row) or  $t = x = 1$  (Column). Note that the retaliation levels,  $q(t)$  and  $p(s)$ , that Row and Column threaten are functions of the levels of preemption,  $t$  and  $s$ , of their opponents; we shall have more to say about this dependence later.

Like other Nash equilibria, DE possesses the property of "mutually fulfilled expectations": the choice of  $q(t) \leq q_1(t)$  by Row and  $p(t) \leq p_1(t)$  by Column associated with DE ensures that any level of preemption by each player's opponent will be at least as costly (after retaliation) as no preemption. Hence, the players do at least as well and generally better choosing no preemption initially ( $s = t = 1$ ).

In an earlier model of deterrence (Brams and Kilgour, 1985b, 1985c), we did not permit retaliation to be a function of the level of preemption of the more preemptive player. Instead, in a Deterrence Game (without variable retaliation), we showed that "optimal deterrence" could be achieved by threats of retaliation independent of the actual level of preemption.<sup>2</sup> The introduction of variable retaliation may be viewed as a refinement of the earlier model in that it ties retaliatory threats to the level of preemption rather than assuming that the same threat will suffice in all situations.

Since one player can, by his choice of his DE strategy, make the compromise outcome uniquely attractive for his opponent, and vice versa, there is a strong reason for both players to settle on DE. Because DE

depends fundamentally on threats, however, it is not surprising that it is neither perfect nor subgame-perfect in the sense of Selten (1975).<sup>3</sup> This is because if deterrence for any reason should fail, it may be irrational to retaliate since retaliation could lead to a worse outcome for the threatener, having to carry out his threat, as well as for the player who preempted and thereby provoked retaliation.

The problem of imperfectness is "solved," in a sense, if the players can irrevocably precommit themselves to carrying out their threats. In the case of nuclear deterrence between the superpowers, both sides have institutionalized detailed procedures for responding to a nuclear attack that are designed to ensure -- insofar as possible -- that retaliation will occur, even if communication, command, control, and intelligence (C<sup>3</sup>I) capabilities are damaged by the attack (Bracken, 1983).<sup>4</sup>

We shall assume that the rules of VRDG permit precommitments by the players that are regarded as credible -- the threatened player believes that the threatener will in fact retaliate at the level he says he will. In the next section, we shall take a closer look at the nature of the retaliatory threats necessary to sustain not only DE but also, should a crisis cause the players' positions on the unit square to deteriorate, a position below and to the right of DE in Figure 1.

### 3. Deterrence in a Crisis

We assumed earlier that the preplay position of the players in VRDG is (1,1). A crisis is an event, or series of events, that changes this position to  $(x_0, y_0)$ , where either  $x_0 < 1$  or  $y_0 < 1$  or both -- that is, at least one player preempts so that  $(x_0, y_0) \neq (1,1)$ . We assume for now that the new preplay position is near [but not at] (1,1), in a sense

to be made precise later. In essence, the crisis has created a situation in which the players find themselves, more or less simultaneously, to be more hostile toward each other than previously, perhaps because one player partially preempted and the other responded with some noncooperative countermove.

If the players wish to return to DE at (1,1) eventually, an option for each is to try to do so immediately. By virtue of the fact that DE is a Nash equilibrium, if either player credibly announces his DE strategy, his opponent can do no better than to return to his own DE strategy. Thus, if Column were to announce  $t = 1$ ,  $p(s) \leq p_1(s)$ , Row maximizes his payoff by returning to  $s = 1$ . The desire to protect himself against possible future preemption by Column would then motivate Row to threaten retaliation as well -- that is,  $q(t) \leq q_1(t)$ .

One problem with effecting stabilization immediately at DE is that it would necessarily involve threats against the status quo  $(x_0, y_0)$ . By assumption, the game is not at (1,1) after the crisis has occurred, so some intimidation would be required to return to this point. In the middle of a crisis, however, a threat of retaliating against the status quo might indeed be dangerous since it could be interpreted as moving up the escalation ladder. Moreover, to create an incentive for an opponent to return to (1,1), the threatener would have to change to a cooperative position himself; this may not be easy to do in a crisis.

It may be preferable for the players to try first to stabilize the status quo  $(x_0, y_0)$ , postponing further ameliorative measures for

restoring the no-preemption position at (1,1) until the crisis atmosphere has cleared. This might be done by indicating a temporary acceptance of the current position, and using milder threats to deter further preemption.

Such a tactic, in forestalling a deepening of the crisis by arresting it at the point where it presently is, affords the players the opportunity to buy more time. Yet it poses a new question: Is it always possible, given a preplay position near but not at full cooperation, for one or both players to stabilize  $(x_0, y_0)$  using less provocative threats than those supporting full deterrence?

If the players perceive their problem as one in which the status quo  $(x_0, y_0)$  brought on by the crisis is the best (i.e., most cooperative) position they can hope for in the short term, then we may consider the revised VRDG to be a game played on a reduced "game board," comprising those points that lie below and to the right of  $(x_0, y_0)$  rather than (1,1) (see Figure 2). In the Appendix, this short-term deterrence problem is

---

Figure 2 about here

---

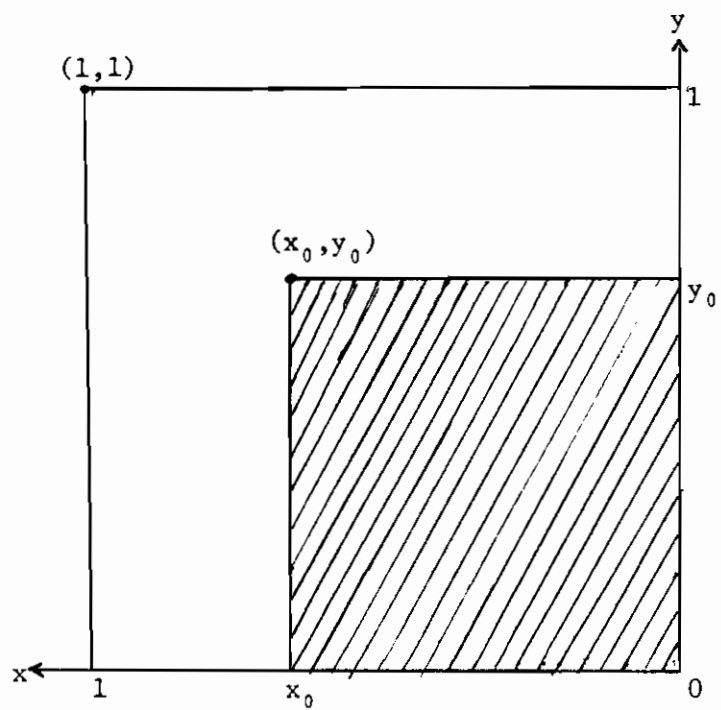
modeled by a Reduced Variable-Retaliatio n Deterrence Game (RVRDG) -- or simply Reduced Game, to avoid tongue twisters -- wherein the preplay position is assumed to be  $(x_0, y_0)$ .

It is shown that, provided that  $(x_0, y_0)$  is not too far from (1,1) [specifically,  $x_0 > x^* = r_2 / (1 - r_3 + r_2)$  and  $y_0 > y^* = c_2 / (1 - c_3 + c_2)$ ], this game has a Deterrence Equilibrium defined by

$$s = y_0, q(t) \leq q_{x_0, y_0}(t); \quad t = x_0, p(x) \leq p_{x_0, y_0}(s),$$

where

FIGURE 2  
REDUCED GAME BOARD AT  $(x_0, y_0)$



$$p_{x_0, y_0}(s) = \frac{P_R(x_0, y_0) - sr_2}{1 - s(1 - r_3 + r_2)}, \quad 0 \leq s < y_0$$

and

$$q_{x_0, y_0}(t) = \frac{P_C(x_0, y_0) - tc_2}{1 - t(1 - c_3 + c_2)}, \quad 0 \leq t < x_0.$$

It is this Reduced Deterrence Equilibrium (RDE) at  $(x_0, y_0)$  that will serve as the cornerstone of our model of crisis stabilization.

#### 4. Crisis Stabilization

We now examine more closely the properties, feasibility, and strategic implications of RDE. Several immediate observations will help. First, notice that  $p_{1,1}(s) = p_1(s)$ , and similarly  $q_{1,1}(t) = q_1(t)$ . This means that, as  $(x_0, y_0)$  approaches  $(1,1)$ , RDE at  $(x_0, y_0)$  approaches DE, the Deterrence Equilibrium of the original Variable-Retaliatio n Deterrence Game, which we shall refer to more simply as the Original Game.

Two mathematical properties of RDE in the Reduced Game are important. Like  $p_1(s)$  and  $q_1(t)$  in the Original Game,  $p_{x_0, y_0}(s)$  and  $q_{x_0, y_0}(t)$  are continuous, strictly increasing, and convex: as the levels of initial cooperation,  $s$  and  $t$ , increase, the minimal threats of retaliation,  $1 - p_{x_0, y_0}(s)$  and  $1 - q_{x_0, y_0}(t)$ , necessary to deter will decrease at a decreasing rate. In other words, players must be relatively more threatening at low levels of initial provocation than at high levels, sometimes threatening even greater retaliation than the provocation at these levels. The nature of such threats is analyzed in detail in Brams and Kilgour (1985a).

The second property is a limit result:

$$\lim_{s \rightarrow y_0} p_{x_0, y_0}(s) = x_0; \quad \lim_{t \rightarrow x_0} q_{x_0, y_0}(t) = y_0.$$

This means that as the level of initial cooperation of one player approaches his position at  $(x_0, y_0)$ , the minimal threatened level of retaliation of the other player also approaches the status quo  $(x_0, y_0)$ . In other words, for initial choices near  $x_0$  or  $y_0$  by the opponent, the minimal "threats" of retaliation will also be small. Thus, when the preemptor stays near RDE, so does the retaliator.

Perhaps the simplest way to understand deterrence equilibria is by means of threat lines. Each player threatens the other with just enough retaliation (at the threshold) to make him prefer no preemption to unilateral preemption at any level. The threat line shows how this minimum level of retaliation depends on the level of preemption.

For example, at the DE in the Original Game, Column threatens to punish any level of preemption by Row (i.e., any  $s < 1$ ) by cooperating at level no more than  $p_1(s)$ . Column's threat line (really a curve) at this equilibrium, called the basic threat line, is given by coordinates

$$(x, y) = (p_1(y), y) \text{ for } 0 \leq y < 1$$

and is shown in Figure 3. Similarly, Column's threat line supporting the

Figure 3 about here

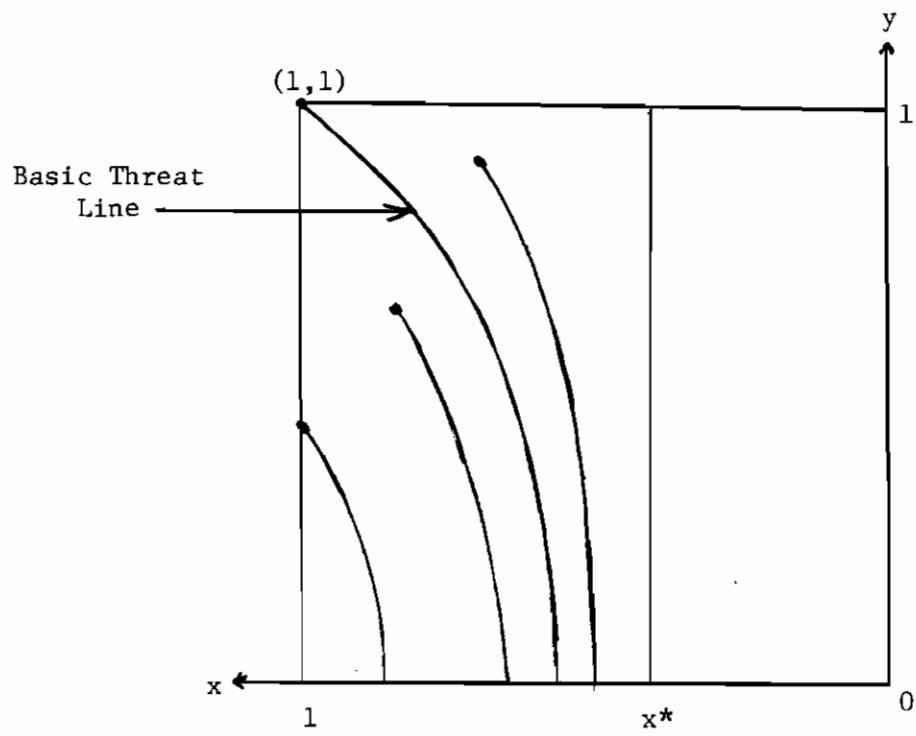
RDE at  $(x_0, y_0)$  in the Reduced Game is given by coordinates

$$(x, y) = (p_{x_0, y_0}(y), y) \text{ for } 0 \leq y < y_0.$$

The upper bound,  $y_0$ , on  $y$  means that at RDE, Column does not threaten to retaliate against the status quo  $y_0$ , just as Column does not retaliate against  $y = 1$  at DE in the Original Game.

In Figure 3 we illustrate several threat lines of Column (besides his basic threat line) for several typical RDEs. Row has a corresponding threat

FIGURE 3  
SEVERAL THREAT LINES OF COLUMN



line at each RDE; in general, the threat line of Row supporting an RDE at  $(x_0, y_0)$  is given by coordinates

$$(x, y) = (x, q_{x_0, y_0}(x)) \text{ for } 0 \leq x < x_0.$$

We next show that a player's threat lines supporting RDEs at different status quo points  $(x_0, y_0)$  never cross -- they are either coincident or disjoint. Fix  $(x_0, y_0)$  satisfying  $x_0 > x^*$  and  $y_0 > y^*$ , where the lower bounds,  $x^*$  and  $y^*$ , are as defined in section 3.

These bounds ensure that  $(x_0, y_0)$  is sufficiently close to  $(1, 1)$  that, near  $(x_0, y_0)$ , it will always be immediately rational for the less preemptive player to retaliate against the more preemptive player (Brams and Kilgour, 1985a). At higher levels of preemption, by comparison, retaliation by the less preemptive player sufficient to deter his opponent is irrational in the sense that it hurts the retaliator as well as his opponent. In effect, we consider only whether a crisis can be stabilized at  $(x_0, y_0)$  in the Reduced Game when rational retaliatory threats by both players are possible.

Given the lower bounds on  $x_0$  and  $y_0$ , consider any  $y$  satisfying  $0 \leq y < y_0$ . The point at height  $y$  on Column's threat line against Row (see Figure 3) is given by coordinates  $(p_{x_0, y_0}(y), y)$ . Row's payoff at this point (see Appendix) is

$$\begin{aligned} P_R(p_{x_0, y_0}(y), y) &= yr_2 + [1 - y(1 - r_3 + r_2)]p_{x_0, y_0}(y) \\ &= yr_2 + [P_R(x_0, y_0) - yr_2] \\ &= P_R(x_0, y_0), \end{aligned}$$

so any point on Column's threat line from  $(x_0, y_0)$  gives Row the same payoff as  $(x_0, y_0)$ . This, of course, is what should be expected since the point on the threat line at height  $y$  represents precisely the minimum

retaliation that would make Row indifferent between no preemption and preemption at level  $y$ .

We have demonstrated that, if two threat lines have a common point, then one must be contained in the other. In particular, if  $(x_0, y_0)$  lies below the basic threat line [originating at  $(1,1)$ ], then the entire threat line from  $(x_0, y_0)$  does also. Similarly, if  $(x_0, y_0)$  lies above the basic threat line, so does its threat line. This situation shows up in the "parallel" threat lines of Figure 3.

We can now answer the question posed in section 3: When can one or both players attempt an immediate crisis stabilization of a status quo point  $(x_0, y_0)$ ? We shall consider two cases. In the first case, the basic threat lines of both Row and Column take the form shown in Figure 4.

---

Figure 4 about here

---

Observe that all points in the unit square close to  $(1,1)$  are either below Column's basic threat line or to the right of Row's (or both). Therefore at least one player [and sometimes both -- in the region of overlap between  $(1,1)$  and  $(x', y')$ ] can practice temporary crisis deterrence.

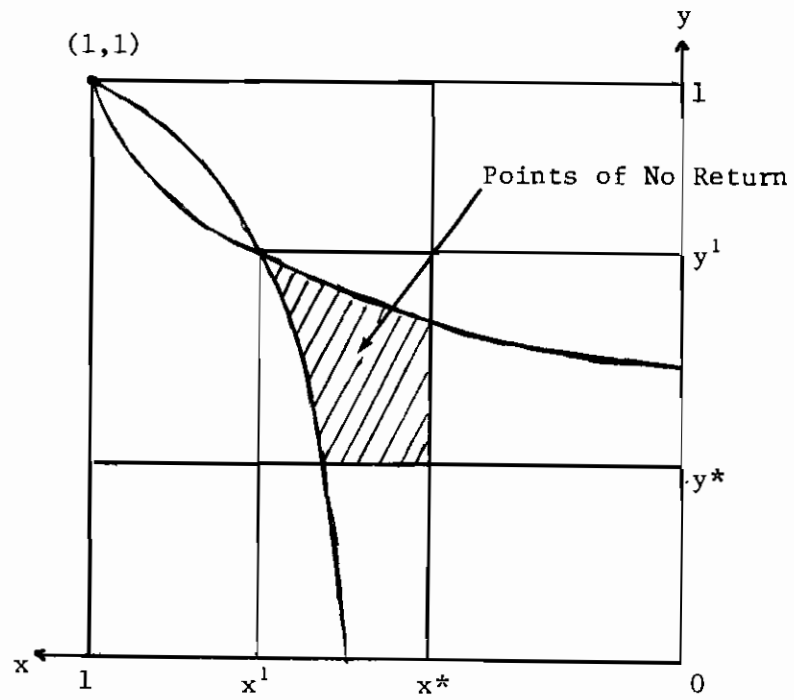
The case illustrated in Figure 4 occurs whenever, in the neighborhood of  $(1,1)$ ,  $(p_1(y), y)$  lies to the right of  $(x, q_1(x))$ . It is shown in the Appendix that this occurs if and only if

$$\frac{1 - c_3}{c_3 - c_2} < \frac{r_3 - r_2}{1 - r_3}. \quad (1)$$

It is noteworthy that, for fixed  $r_2$ ,  $c_2$ , and  $c_3$ , inequality (1) holds if  $r_3$  is large enough, and, similarly, for fixed  $c_2$ ,  $r_2$ , and  $r_3$ , (1) holds

FIGURE 4

CONFIGURATION OF BASIC THREAT LINES IN WHICH CRISIS STABILIZATION IS ALWAYS POSSIBLE NEAR  $(1,1)$



if  $c_3$  is large enough. For example, if  $r_3 = 0.8$  and  $c_3 = 0.6$  and  $r_2 = c_2 = 0.3$ , then inequality (1) is satisfied.

If inequality (1) does not hold, Figure 5 applies and there are

---

Figure 5 about here

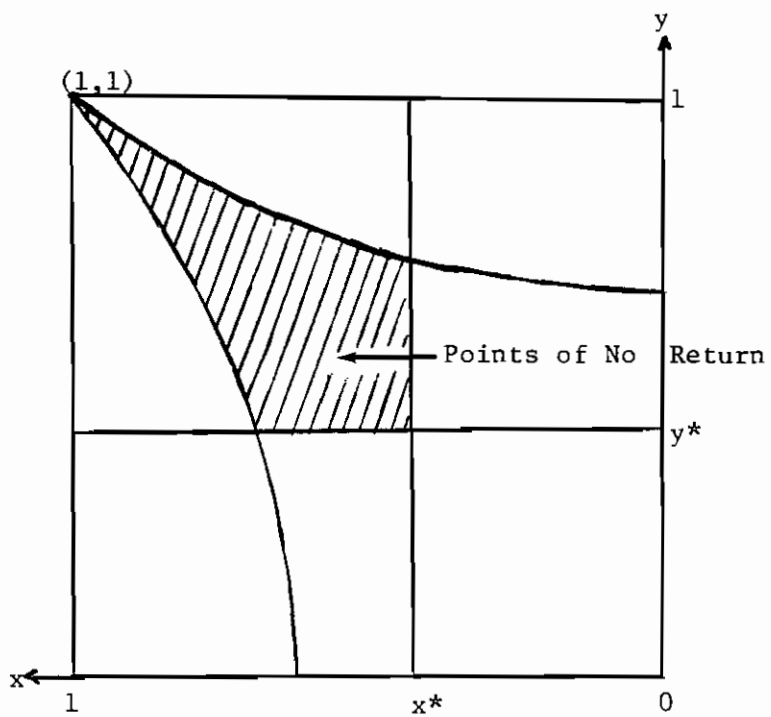
---

initial points close to  $(1,1)$  -- specifically, near the main diagonal where the threat lines diverge -- from which neither player can attempt crisis stabilization. On the other hand, if  $(x_0, y_0)$  is sufficiently close to one of the edges of the unit square, this status quo point will be either below Column's basic threat line (near the vertical axis), or to the right of Row's (near the horizontal axis), and one player can stabilize  $(x_0, y_0)$ .

A point of no return is one that is neither below Column's basic threat line nor to the right of Row's, which in Figure 5 falls below and to the right of  $(1,1)$ . In Figure 4 there are also points of no return southeast of  $(1,1)$ , but they are not in the neighborhood of  $(1,1)$ . Instead, they are points below and to the right of the interior point of intersection  $(x', y')$  of the two basic threat lines.

Actually, all points in the regions where the curves "flare outward" [from  $(x', y')$  in Figure 4 and from  $(1,1)$  in Figure 5] are points of no return. It is noteworthy that, whether Figure 4 or Figure 5 is applicable, crisis stabilization is always facilitated if  $(x_0, y_0)$  falls near one of the edges of the unit square. This is because the preemptor has more to lose and is thus more vulnerable to threats. Hence, a crisis in which both players are more or less equally noncooperative is generally harder to stabilize -- except when inequality (1) holds and the players are close to  $(1,1)$  and therefore protected.

FIGURE 5  
CONFIGURATION OF BASIC THREAT LINES IN WHICH CRISIS STABILIZATION MAY  
NOT BE POSSIBLE NEAR (1,1)



In general, crisis stabilization is enhanced when the advantages of mutual cooperation in the Original Game are substantial for at least one player, thereby ensuring in the Reduced Game that the status quo at  $(x_0, y_0)$  is not too far from  $(1, 1)$ . In addition, crisis stabilization is abetted if  $(x_0, y_0)$  significantly favors one player; this makes his opponent's threat of retaliation potentially more harmful and hence likely to be taken more seriously.

## 5. Conclusions

It is important to emphasize that crisis stabilization is always possible in the Variable-Retaliatio n Deterrence Game (Original Game) if either player credibly threatens retaliation at a level given by the Deterrence Equilibrium of this game. The new question posed in this paper concerns simultaneous departures: If both players should, because of a crisis, find themselves at point  $(x_0, y_0) \neq (1, 1)$  of the unit square, can they stabilize this point using milder threats, making it a new status quo from which neither player will have an incentive to depart unilaterally?

The calculations of the Original Game can be carried over to this Reduced Game, with  $(x_0, y_0)$  now assumed to be the preplay position. If the goal of the players is first to stabilize  $(x_0, y_0)$ , and not necessarily induce a return to  $(1, 1)$  immediately, then this goal cannot always be realized without threatening  $(x_0, y_0)$  itself.

This goal of temporary stabilization seems a sensible one for the players to aspire to if preventing a further deterioration of their positions in a crisis is a sine qua non of eventually easing out of the crisis and moving back to the status quo ante at  $(1, 1)$ . Numerous

crises in international politics have been abated by just such incremental steps.

For example, the eventual settlement of the Cuban missile crisis in October 1962 depended initially on the removal by the Soviets of the missiles they were about to install in Cuba. Only after the Soviets promised to remove the missiles did the United States agree to lift its blockade of Cuba, after which the crisis subsided.

Our game-theoretic analysis showed that crisis stabilization may not always be possible, even in the case of small departures from (1,1). The cooperative outcome must also be valued highly by at least one, and preferably both, players. However, if this is not the case, a more surprising condition can facilitate stabilization: the crisis places the players in asymmetrical positions whereby one player is substantially more antagonistic, or preemptive, than the other at  $(x_0, y_0)$ . The more cooperative player is then in a position to reduce his threats, indicating acceptance of  $(x_0, y_0)$  and thereby stabilizing it.

In the Cuban missile crisis, the Soviets initially reaped a significant advantage by their preemption, but in the Caribbean the United States could threaten them with strong military reprisal, both of a conventional kind against Cuba and of a nuclear kind against the Soviet Union itself. More perilous would be a potential crisis in which there was rapid escalation on both sides, and neither side could pose an overwhelmingly greater threat against the other, perhaps because the confrontation occurred in neither's "sphere of influence."

It is perhaps fortunate that most preemptive moves made by the super-powers have been against foes that were perceived to be within their

spheres of influence. The greatest danger, according to our model, is when there is sudden and serious escalation on both sides, which seems most likely to occur when the confrontation is seen to be outside one's immediate sphere of influence, such as in the Middle East or even outer space.

Stabilizing such a crisis without threatening the status quo, much less returning to the status quo ante, may be impossible for rational players. We hope a recognition that there are such points of no return, and that there may be no escape from them without aggravating the crisis further, will help to avert such crises.

## APPENDIX

We begin by defining various "deterrence games," specifying the strategy choices available to the players, the rules for determining final outcomes, and the payoff functions. The game board of each game is its set of possible final outcomes. For example, the Variable-Retaliatioin Deterrence Game (VRDG) has the unit square  $([0,1]^2)$  as its game board, and its strategy choices, final outcome, and payoffs are as given in Table 1  $[(1_S), (1_F), \text{ and } (1_P), \text{ respectively}]$ . Note that  $r_2, r_3, c_2,$  and  $c_3$  are constants satisfying  $0 < r_2 < r_3 < 1$  and  $0 < c_2 < c_3 < 1$ .

Fix  $x_0$  and  $y_0$  satisfying  $0 < x_0 < 1$  and  $0 < y_0 < 1$ . When the game board is reduced from the unit square to  $[0, x_0] \times [0, y_0]$ , the problem faced by the players can be modeled by a reduced game. We define the Reduced VRDG at  $(x_0, y_0)$  to be Game 2 of Table 1. Game 2 is played just

---

Table 1 about here

---

as the original game; the only difference is in the "shrinking" of  $(x,y)$  by factors of  $x_0$  and  $y_0$  in  $(2_F)$  to obtain a final outcome which always lies on the reduced game board.

It is possible to think of Game 2, the Reduced VRDG at  $(x_0, y_0)$ , as being played entirely within the new game board. This is so because Game 3 is equivalent to Game 2, using the transformations  $s' = sy_0,$   $t' = tx_0,$   $q'(t') = y_0q(t'/x_0),$  and  $p'(s') = x_0p(s'/y_0)$ . Although Game 3 might be seen as the more natural way to reduce VRDG to the game board  $[0, x_0] \times [0, y_0]$ , we shall continue to work with Game 2 for convenience.

TABLE 1  
FOUR DETERRENCE GAMES

<u>Game 1 - Variable-Retaliatio n Deterrence Game (VRDG)</u>	
(1 <sub>S</sub> ) Strategies:	Row $s \in [0, 1]$ and, if $s > 0$ , $q: [0, s] \rightarrow [0, 1]$ ; Column $t \in [0, 1]$ and, if $t > 0$ , $p: [0, t] \rightarrow [0, 1]$ .
(1 <sub>F</sub> ) Final Outcome:	$(x, y) = \begin{cases} (p(s), s) & \text{if } s < t \\ (t, s) & \text{if } s = t \\ (t, q(t)) & \text{if } s > t \end{cases}$
(1 <sub>P</sub> ) Payoffs:	Row $P_R(x, y) = xy(r_3) + x(1-y)(1) + (1-x)y(r_2) + (1-x)(1-y)(0)$ $= x + yr_2 - xy(1 - r_3 + r_2)$ Column $P_C(x, y) = y + xc_2 - xy(1 - c_3 + c_2)$ .
<u>Game 2 - Reduced VRDG at <math>(x_0, y_0)</math>: Board Reduction</u>	
(2 <sub>S</sub> ) Strategies:	as (1 <sub>S</sub> )
(2 <sub>F</sub> ) Final Outcome:	$(x', y') = (x_0x, y_0y)$ , where $(x, y)$ as in (1 <sub>F</sub> ).
(2 <sub>P</sub> ) Payoffs:	Row $P_R(x', y')$ ; Column $P_C(x', y')$ , where $P_R$ and $P_C$ as in (1 <sub>P</sub> ).
<u>Game 3 - Reduced VRDG at <math>(x_0, y_0)</math>: Strategy Transformation</u>	
(3 <sub>S</sub> ) Strategies:	Row $s' \in [0, y_0]$ and, if $s' > 0$ , $q': [0, x_0s'/y_0] \rightarrow [0, y_0]$ ; Column $t' \in [0, x_0]$ and, if $t' > 0$ , $p': [0, y_0t'/x_0] \rightarrow [0, x_0]$ .
(3 <sub>F</sub> ) Final Outcome:	$(x', y') = \begin{cases} (p'(s'), s') & \text{if } s'/y_0 < t'/x_0 \\ (t', s') & \text{if } s'/y_0 = t'/x_0 \\ (t', q'(t')) & \text{if } s'/y_0 > t'/x_0 \end{cases}$
(3 <sub>P</sub> ) Payoffs:	as (2 <sub>P</sub> ).
<u>Game 4 - Reduced VRDG at <math>(x_0, y_0)</math>: Payoff Transformation</u>	
(4 <sub>S</sub> ) Strategies:	as (1 <sub>S</sub> ).
(4 <sub>F</sub> ) Final Outcome:	as (1 <sub>F</sub> ).
(4 <sub>P</sub> ) Payoffs:	Row $Q_R(x, y) = P_R(x', y')/x_0$ ; Column $Q_C(x, y) = P_C(x', y')/y_0$ , where $P_R$ and $P_C$ as in (1 <sub>P</sub> ), and $(x', y')$ as in (2 <sub>F</sub> ).

Similarly, Game 4 is strategically equivalent to Game 2, for the strategy sets and resulting position  $(x,y)$  are identical in both, and the payoffs in Game 4 are constant multiples of the payoffs in Game 2. Thus Games 2, 3, and 4 are strategically identical.

We next determine the precise conditions under which Games 1 and 2 are equivalent. First compare the payoff functions of Games 1 and 4 for Row only -- the situation for Column is analogous. From  $(4_p)$  and  $(2_F)$  it is easy to verify that

$$\begin{aligned} Q_R(1,1) &= [x_0 + y_0 r_2 - x_0 y_0 (1 - r_3 + r_2)] / x_0, & Q_R(0,1) &= y_0 r_2 / x_0, \\ Q_R(1,0) &= 1, & Q_R(0,0) &= 0. \end{aligned}$$

Define  $r'_2 = Q_R(0,1)$  and  $r'_3 = Q_R(1,1)$ . Then

$$Q_R(x,y) = xy(r'_3) + x(1-y)(1) + (1-x)y(r'_2) + (1-x)(1-y)(0) \quad (5)$$

since the right side of (5) equals

$$\begin{aligned} & \{xy[x_0 + y_0 r_2 - x_0 y_0 (1 - r_3 + r_2)] + x(1-y)x_0 + (1-x)y[y_0 r_2]\} / x_0 \\ &= [xx_0 + yy_0 r_2 - xx_0 yy_0 (1 - r_3 + r_2)] / x_0 = P_R(xx_0, yy_0) / x_0 = Q_R(x, y) \end{aligned}$$

because of  $(2_F)$  and  $(4_p)$ .

Comparison of (5) with  $(1_p)$  shows that Row faces the same strategic problem in Game 4 as in Game 1, provided that  $0 < r'_2 < r'_3 < 1$ . Since

$$r'_3 - r'_2 = \frac{x_0 + y_0 r_2 - x_0 y_0 (1 - r_3 + r_2)}{x_0} - \frac{y_0 r_2}{x_0} = 1 - y_0 (1 - r_3 + r_2) > 0,$$

the games are strategically equivalent for Row iff (if and only if)

$r'_3 < 1$ , which is equivalent to

$$x_0 + y_0 r_2 - x_0 y_0 (1 - r_3 + r_2) < x_0.$$

Since  $y_0 > 0$ , this inequality holds iff

$$x_0 > \frac{r_2}{1 - r_3 + r_2} = x^*.$$

Combining with the analogous conditions for Column's payoffs, we conclude that Game 1 and Game 2 are strategically equivalent iff

$x_0 > x^*$  and  $y_0 > y^*$ .

Now assume that  $x_0 > x^*$  and  $y_0 > y^*$ . Our previous analysis of Game 1 (VRDG) applies equally to the Reduced VRDGs at  $(x_0, y_0)$  because of the strategic equivalence we have demonstrated. For Game 4, the Deterrence Equilibrium is

$$(DE) \quad s = 1, q(t) \leq q_1(t); \quad t = 1, p(x) \geq p_1(s),$$

where  $p_1(s) = \frac{r'_3 - sr'_2}{1 - s(1 - r'_3 + r'_2)}$ ,  $0 \leq s < 1$ , and  $q_1(t)$  is similar. Of

course (DE) is identical in Game 2, for  $(4_S)$  and  $(2_S)$  are the same.

The transformation given above shows, after some simplification, that in Game 3 (DE) becomes

$$(RDE) \quad s' = y_0, q'(t') \leq q_{x_0, y_0}(t'); \quad t' = x_0, p'(s') \leq p_{x_0, y_0}(s'),$$

where  $p_{x_0, y_0}(s') = \frac{P_R(x_0, y_0) - s'r_2}{1 - s'(1 - r_3 + r_2)}$ ,  $0 \leq s' < y_0$

and  $q_{x_0, y_0}(t') = \frac{P_C(x_0, y_0) - t'c_2}{1 - t'(1 - c_3 + c_2)}$ ,  $0 \leq t' < x_0$ .

Here  $P_R(x_0, y_0)$  and  $P_C(x_0, y_0)$  are as given by  $(1_p)$ .

Finally we show that, near  $(1,1)$ , Column's basic threat line  $(p_1(y), y)$  lies to the right of Row's basic threat line  $(x, q_1(x))$

iff

$$\frac{1 - c_3}{c_3 - c_2} < \frac{r_3 - r_2}{1 - r_3}. \quad (6)$$

(See Figures 4 and 5 in the text.) Recall that

$$p_1(y) = \frac{r_3 - yr_2}{1 - y(1 - r_3 + r_2)}, \quad 0 \leq y < 1; \quad q_1(x) = \frac{c_3 - xc_2}{1 - x(1 - c_3 + c_2)}, \quad 0 \leq x < 1.$$

Since  $\lim_{y \rightarrow 1} p_1(y) = \lim_{x \rightarrow 1} q_1(x) = 1$ , both threat lines pass through  $(1,1)$ .

We now show that, at  $(1,1)$ , the slope of  $(x, q_1(x))$  is less than that of  $(p_1(y), y)$  iff (6) holds. Since

$$\frac{dq_1}{dx} = \frac{(c_3 - c_2)(1 - c_3)}{[1 - x(1 - c_3 + c_2)]^2},$$

we have that the limiting slope of  $(x, q_1(x))$  at  $x = 1$  is

$$\left. \frac{dq_1}{dx} \right|_{x=1} = \frac{(c_3 - c_2)(1 - c_3)}{(c_3 - c_2)^2} = \frac{1 - c_3}{c_3 - c_2}.$$

Similarly,  $\left. \frac{dp_1}{dy} \right|_{y=1} = \frac{1 - r_3}{r_3 - r_2}$ , so that the limiting slope of  $(p_1(y), y)$

at  $y = 1$  is

$$\left( \left. \frac{dp_1}{dy} \right|_{y=1} \right)^{-1} = \frac{r_3 - r_2}{1 - r_3}.$$

Comparison of these two slopes now yields (6).

## FOOTNOTES

1. A more extended discussion of crisis instability can be found in O'Neill (1985b), who reviews different efforts to formalize crisis instability and defines axiomatically a game-theoretic index of this concept. See also O'Neill (1985a) and Intriligator and Brito (1977), whose analysis of a "cone of mutual deterrence" is somewhat akin to our geometric analysis of areas of stability and instability later.

2. In related work, Dacey (1985) analyzes within a decision-theoretic framework the use of probabilistic bribes, threats, and tit-for-tat combinations in a number of variable-sum games. Probabilistic variations on tit-for-tat are developed in Brams and Davis (1985) and Brams, Davis, and Straffin (1979), which can also be found in Brams (1985, chs. 3 and 4).

3. See also Shubik (1982, pp. 265-270) for a recent discussion of perfect equilibria and related concepts.

4. Gauthier (1984) claims that such precommitments are not necessary to deter aggression, but threats which are not credible are empty, and empty threats invite attack. His calculus of deterrence, we believe, is sensible only when his retaliator's threats will assuredly be implemented because of precommitments.

## REFERENCES

- Bracken, Paul (1983). The Command and Control of Nuclear Weapons. New Haven: Yale University Press.
- Brams, Steven J. (1985). Superpower Games: Applying Game Theory to Superpower Conflict. New Haven: Yale University Press.
- Brams, Steven J., Morton D. Davis, and Philip D. Straffin, Jr. "The Geometry of the Arms Race." International Studies Quarterly 23, no. 4 (December 1979): 567-588.
- Brams, Steven J., and D. Marc Kilgour (1985a). "Deterrence by Variable Retaliation: A Game-Theoretic Analysis." Mimeographed.
- Brams, Steven J., and D. Marc Kilgour (1985b). "Optimal Deterrence." Social Philosophy and Policy 3, no. 1 (Autumn): 118-135.
- Brams, Steven J., and D. Marc Kilgour (1985c). "The Path to Stable Deterrence." In Dynamic Models of International Conflict, ed. Urs Luterbacher and Michael D. Ward. Boulder, CO: Lynne Rienner, 1985, pp. 11-25.
- Brams, Steven J., and D. Marc Kilgour (1985d). "Rational Descalation." Proceedings of the Conference on Evolution, Games, and Learning, (forthcoming).
- Brams, Steven J., and D. Marc Kilgour (1985e). "Winding Down If Preemption or Escalation Occurs: A Game-Theoretic Analysis." Mimeographed.
- Dacey, Raymond (1985). "Ambiguous Information and the Manipulation of Plays of the Arms Race Game and the Mutual Deterrence Game." In Interaction and Communication in Global Politics, ed. Claudio Cioffi-Revilla, Richard L. Merritt, and Dina Zinnes. Beverly Hills, CA: Sage (forthcoming).

- Gauthier, David (1984). "Deterrence, Maximization, and Rationality." Ethics 94, no. 3 (April): 474-495.
- Intriligator, Michael D. and Dagobert L. Brito (1976). "Formal Models of Arms Races." Journal of Peace Science 2, no. 1 (Spring): 77-88.
- Nash, John (1951). "Non-cooperative Games." Annals of Mathematics 54: 286-295.
- O'Neill, Barry (1985a). "Applications of a Crisis Instability Index: Arms Control Agreements and Spaced-Based Missile Defenses." Discussion Paper No. 651, Center for Mathematical Studies in Economics and Management Science, Northwestern University, mimeographed.
- O'Neill, Barry (1985b). "A Measure for Crisis Instability." Discussion Paper No. 652, Center for Mathematical Studies in Economics and Management Science, Northwestern University, mimeographed.
- Selten, Reinhard (1975). "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." International Journal of Game Theory 4, no. 1: 25-55.
- Shubik, Martin (1982). Game Theory in the Social Sciences: Concepts and Solutions. Cambridge, MA: MIT Press.
- Zagare, Frank C. (1984). "Limited-Move Equilibria in  $2 \times 2$  Games." Theory and Decision 16, no. 1 (January): 1-19.
- Zagare, Frank C. (1985). "Toward a Reformulation of the Theory of Mutual Deterrence." International Studies Quarterly 29, no. 2 (June): 155-169.