Chapter 1
In Search of an Index

OUR FIRST TASK will be to find some convenient way of studying behavior in Prisoner's Dilemma in relation to the payoffs. Clearly it is desirable to have some index derived from the payoffs and to relate observed behavior to this index.

The matrix which represents Prisoner's Dilemma has eight entries in it. The magnitudes of these entries can vary independently within certain constraints. Therefore any behavioral index derived from "performances" on this game can be conceived as a function of eight independent variables. Such a relationship is far too complex to be grasped intuitively. Fortunately there is a natural way to simplify the situation. Four of the entries are the payoffs to one of the players, and four are the payoffs to the other. We can confine our attention only to those forms of the game which are symmetric with regard to the players, that is, look exactly alike from the point of view of each of the players. To realize this symmetry, we let the payoffs corresponding to each outcome depend only on how the players have chosen, not on how the players are labeled. To be sure, this symmetry in the entries does not necessarily represent psychological symmetry, since the payoff units may have different meanings (utilities)[10] for each of the players. But equalizing the entries is the best we can do.

Our game matrix now appears as shown in Matrix 6. The letters representing the payoffs are meant to be suggestive. *R* stands for reward; it refers to the payoff each of the players receives as reward for co-

	C_2	D_2
C_1	R,R	S,T
D_1	T,S	P,P

Matrix 6.

operating. S stands for sucker's payoff. This is the payoff received by the player who cooperated while the other defected. T stands for temptation, the payoff a player may hope to get if he can defect and get away with it. P stands for punishment, meted out to both players when both have defected.

In order to have the situation defined as Prisoner's Dilemma the following inequalities must be satisfied:

$$S < P < R < T. \tag{1}$$

Specifically, when a player gets the sucker's payoff S, he must be motivated to switch to the defecting strategy so as to get at least P. If he gets the cooperator's payoff R, he must be motivated to defect so as to get still more, T. If he gets the defector's punishment P, he may wish there were a way of getting R, but this is possible only if the other defector will switch to the cooperative strategy together with him.

Besides these inequalities, we shall wish to introduce one additional constraint, namely

$$2R > S + T. \tag{2}$$

If this inequality did not hold, that is, if $S + T$ were equal to or were greater than $2R$, the players would have at their disposal more than one form of tacit collusion. One form of collusion is the tacit agreement to play CC, which is the expected "cooperative solution" of the Prisoner's Dilemma game. However, if $S + T \geq 2R$, there is also another form of collusion, which may occur in repeated plays of the game, namely alternation between CD and DC. Each such alternation

gives each player $S + T$, i.e., at least $2R$, which each player would have obtained from two consecutive *CC* outcomes. The question of whether the collusion of alternating unilateral defections would occur and, if so, how frequently is doubtless interesting. For the present, however, we wish to avoid the complication of multiple "cooperative solutions." Accordingly the inequality (2) will be assumed to hold throughout.

Having reduced the Prisoner's Dilemma game to a matrix with four parameters, independent within the constraints imposed, we make one additional step toward simplification. Namely, we set $S = -T$. This cuts the number of independent parameters to three. In the three-parameter game, S and T become mirror images of each other, and therefore whatever effect they exert on the performance, they exert together. In all games with which we shall be concerned, we shall have $T > 0$, consequently $S < 0$. Further, since in all our games $T + S = 0$, we must have $R > 0$, since $2R > T + S$ [cf. (2)]. Therefore it should be kept in mind that when we write "S increases," this will imply that the numerical value of S *decreases*, i.e., that the sucker's punishment becomes smaller. Mutatis mutandis "S decreases" will mean that the sucker's punishment becomes larger. Similar remarks apply to P when $P < 0$, as it will be in all our games. We shall sometimes refer to different Prisoner's Dilemma games as "mild" or "severe." Mild games are those where T is not much larger than R or where P is numerically large, i.e., those where it does not pay very much to defect. Severe games are those where temptation to defect is strong or the punishment for double defection is weak, or both.

We are now in a position to ask a straightforward question. Are the motivations inherent in the relations among the payoffs reflected in the performances?

Specifically, R rewards cooperation. Is it true that if the other parameters are held constant, while R increases, more cooperation will be observed?

T represents temptation to defect. Is it true that as the other parameters are held constant, cooperation will decrease as T is increased? Note that the numerical magnitude of S, i.e., the magnitude of the sucker's punishment, increases together with that of T in all the games to be considered here. Hence in the three-parameter model any effect attributed to T can with equal justification be attributed to S.

Finally, P represents punishment for failure to cooperate. Is it true that cooperation increases as the magnitude of this punishment becomes greater, i.e., as P decreases?

We shall offer answers to these questions in terms of the frequencies of cooperative responses observed in many repeated plays of Prisoner's Dilemma averaged over many pairs of subjects. We shall refer to the cooperative response as the C response. When there is no danger of confusion, C will also mean the frequency of the C response and its unconditional probability. Similarly D will mean the defecting response, its frequency, and its probability.

Experimental Procedure

We let seventy pairs of University of Michigan students (all males) play Prisoner's Dilemma games three hundred times in succession. The pairs were matched randomly. Except in a very few cases, the members of a pair were not acquainted with each other. The instructions read to the subjects are given in Appendix I. Following each play, the outcome was announced. Each pair played only one variant of the game. There were seven such variants, so that ten pairs were assigned to each. The seven payoff matrices are shown

below. We have kept the number designations of the games to avoid confusion. Games VI through X are not shown here. They were used in other experiments.

	C	*D*
C	9,9	−10,10
D	10,−10	−1,−1

Matrix 7.
Game I.

	C	*D*
C	1,1	−10,10
D	10,−10	−9,−9

Matrix 8.
Game II.

	C	*D*
C	1,1	−10,10
D	10,−10	−1,−1

Matrix 9.
Game III.

	C	*D*
C	1,1	−2,2
D	2,−2	−1,−1

Matrix 10.
Game IV.

	C	*D*
C	1,1	−50,50
D	50,−50	−1,−1

Matrix 11.
Game V.

	C	*D*
C	5,5	−10,10
D	10,−10	−1,−1

Matrix 12.
Game XI.

	C	*D*
C	1,1	−10,10
D	10,−10	−5,−5

Matrix 13.
Game XII.

From the matrices of the seven games, the following can be observed. In Games III, XI, and I, *T*, *S*, and *P* are held constant, while *R* increases from 1 to 5 to 9. One would expect, therefore, that in these games *C* would be largest in Game I and smallest in Game III.

Next observe that in Games IV, III, and V, R and P are held constant while the magnitude of S and T increases from 2 to 10 to 50. One would therefore expect that in these games C would be largest in Game IV and smallest in Game V. Finally in Games III, XII, and II, R, S, and T are kept constant while P decreases from -1 to -5 to -9. Therefore we would expect C to increase from Game III to Game XII to Game II.

We combine these conjectures into Hypothesis 1. *If other payoffs are kept constant, C increases as R and S increase and decreases as T and P increase.*

Comparisons of Games III, XI, and I, Games IV, III, and V, and Games III, XII, and II are shown in Figure 1. Without inquiring into the significance of the comparisons, we conclude that the results tend to corroborate Hypothesis 1.

Hypothesis 1 implies certain rank orderings with respect to C among the games. But the rank order of some of the games is not established by the hypothesis. For example, nothing is implied about the rank order of Games I and II, II and IV, IV and XI, or XI and XII.

To get a theoretical rank ordering of all seven games, we need an index composed of all our parameters, T, R, P, and S, i.e., an index which is a function of four independent variables. It would be all but futile to try to get an idea of this function by examining our data, for that would mean plotting values of C in five-dimensional space. We can, however, make some a priori arguments about the general character of such a function on the basis of some available theory.

Assuming an Interval Scale of Utilities

The theory in question was developed in a purely formal context. To our knowledge no conclusive evidence has ever been offered favoring the conclusion that the theory is valid in *behavioral* contexts. We are re-

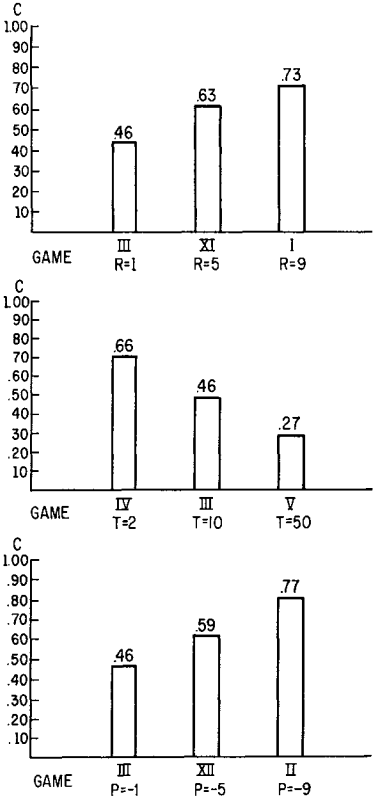

Figure 1.

ferring to the concept of utility, as it has been used in the theory of games. In game theory, utilities (which are supposed to be the entries in the payoff matrices) are almost always assumed to be measured on interval scales at least in two-person games.[11] This, in turn, means that, if all the entries u ($u = T$, R, P, or S) in a game matrix are replaced by $au + b$ (where a and b are constants, $a > 0$), all the conclusions about the game resulting from this transformation should be identical to the corresponding conclusions about the

original game. It seems reasonable to expect this with respect to rational players. Suppose, for example, a given two-person game is transformed by adding a constant to each of the entries. If this constant is positive, this means that each of the players gets a fixed amount just for playing the game (regardless of how he plays it). If the constant is negative, this means that each player must pay a fee for the privilege of playing the game each time, again regardless of how he plays. Either way, these fixed amounts should not make any difference in the *strategic* analysis of the game, precisely because the fixed reward or fee does not depend on what the players do.

Consider next the case where the entries of the payoff matrix are all multiplied by the same positive number. This can be interpreted as simply changing the units of the payoffs. Again the transformation should make no difference in the strategic structure of the game. Combining the two transformations, we have the transformation $au + b$.

Whether such a transformation does not in fact change the players' conception of the game is a psychological question, not a game-theoretical one. But as we have seen, game theory is totally devoid of psychology. It assumes "rational players," that is, "perfect players." These cannot be expected to be human and therefore cannot be expected to have a psychology. We shall, to be sure, be concerned with real subjects and with their observed behavior in the games under discussion. But it will be useful to forget this for the moment, to assume players devoid of psychology and to stay with this assumption as long as we can. Accordingly we shall assume that the behavior of our players remains invariant if the payoff matrices of our Prisoner's Dilemma games are subjected to a linear

transformation, i.e., to a transformation of the form $au + b$ $(a > 0)$.

An immediate consequence of this assumption is that our behavioral variable C should depend not on the parameters, T, R, P, and S individually but rather on the ratios of their differences.[12] For example, $(R - P)/(T - S)$ is one such ratio; $(T - P)/(P - S)$ is another. Formally speaking, thirty such interval ratios can be formed from the four parameters. Of course fifteen of these will be reciprocals of the other fifteen and so can be immediately dismissed from consideration as independent variables. It can be further shown that only two of these interval ratios can be independent: the remaining ones can all be derived from just two. The two can be chosen in many ways. We chose the following:

$$r_1 = \frac{R - P}{T - S} \quad \text{and} \quad r_2 = \frac{R - S}{T - S}. \tag{3}$$

As an example of the dependence of the other ratios on r_1 and r_2, observe that the interval ratio $(P - S)/(T - S)$ is obtained as $r_2 - r_1$; $(T - R)/(T - S)$ is obtained as $1 - r_2$; $(P - S)/(T - R)$ is obtained as $(r_2 - r_1)/(1 - r_2)$, etc. In short, all the fifteen interval ratios and their fifteen reciprocals can be obtained either as linear functions or as bilinear functions (ratios of two polynomials of the first degree) of r_1 and r_2. Thus if r_1 and r_2 are given, all the thirty interval ratios upon which the rank-ordering index of the game can depend are also determined. It suffices, therefore, to consider such an index as a function of the basic pair r_1 and r_2 alone.

In choosing the basic pair given by (3), we were guided by the convenience of having $T - S$ in the denominator. Since $T - S$ is the largest of the six

differences [cf. inequality (1)], this guarantees against infinitely large values of r_1 and r_2, because the denominator cannot vanish without the numerator vanishing simultaneously. In fact, for all permissible values of T, R, P, and S, we must have $0 < r_1 < 1$ and $0 < r_2 < 1$.

Now let us examine the dependence of r_1 and r_2 on the four parameters. We have

$$\frac{\partial r_1}{\partial R} > 0; \quad \frac{\partial r_1}{\partial P} < 0; \quad \frac{\partial r_1}{\partial T} < 0; \quad \frac{\partial r_1}{\partial S} > 0. \tag{4}$$

Since Hypothesis 1 implies

$$\frac{\partial C}{\partial R} > 0; \quad \frac{\partial C}{\partial T} < 0; \quad \frac{\partial C}{\partial P} < 0; \quad \frac{\partial C}{\partial S} > 0, \tag{5}$$

it also implies that C increases with r_1. As we shall see, this determines the rank ordering of Games II and III. This rank ordering is already implied by Hypothesis 1 and so puts no new restriction on the theoretical rank ordering of the games. However, as we shall see, the dependence of C on r_1 also establishes the rank orders of Games IV, XI, and XII, in that order, which Hypothesis 1 does not do. Hence the assumption that C increases with r_1 is a stronger hypothesis than Hypothesis 1. We shall call it Hypothesis 2:

$$\frac{\partial C}{\partial r_1} > 0. \tag{6}$$

We turn to r_2. Taking the partial derivatives, we get

$$\frac{\partial r_2}{\partial R} > 0; \quad \frac{\partial r_2}{\partial T} < 0; \quad \frac{\partial r_2}{\partial S} < 0. \tag{7}$$

We see that while the partials of r_2, with respect to R and T have the same sign as the corresponding partials of r_1, the partial of r_2 with respect to S is negative, while the corresponding partial of r_1 is positive. Hence the dependence of C upon r_2 is not determined

by Hypothesis 1. If such dependence is to be assumed, an independent hypothesis must be proposed. There seems to be no a priori justifiable reason for supposing either $\partial C/\partial r_2 > 0$ or $\partial C/\partial r_2 < 0$. We therefore can have either of the following hypotheses:

Hypothesis 3:

$$\frac{\partial C}{\partial r_1} > 0; \quad \frac{\partial C}{\partial r_2} > 0; \tag{8}$$

Hypothesis 4:

$$\frac{\partial C}{\partial r_1} > 0; \quad \frac{\partial C}{\partial r_2} < 0. \tag{9}$$

Both Hypothesis 3 and Hypothesis 4 include Hypothesis 2 and hence a fortiori Hypothesis 1. But Hypotheses 3 and 4 are mutually exclusive. The data can therefore support the one or the other (if any) but not both.

Table 1 shows the values of r_1 and r_2 in each of our seven games.

TABLE 1

Game	r_1	r_2
I	1/2	19/20
II	1/2	11/20
III	1/10	11/20
IV	1/2	3/4
V	1/50	51/100
XI	3/10	3/4
XII	3/10	11/20

We see that Games II, XII, and III all have the same value of r_2 but the respective values of r_1 decrease in that order. Therefore the rank order should be II > XII > III according to our Hypothesis 2. But this was already implied by Hypothesis 1. Next, Games IV and XI have the same value of r_2 and therefore are rank ordered IV > XI according to the respective

values of r_1. This was not implied by Hypothesis 1 but is consistent with it.

Examining games which have the same value of r_1, we find that they are Games I, II, and IV. On the basis of Hypothesis 3, C increases with r_2; hence these games must be rank ordered I > IV > II. On the basis of Hypothesis 4, on the contrary, these games must have the opposite rank ordering, namely II > IV > I. Finally, XI and XII have the same value of r_1 and consequently must be rank ordered XI > XII if Hypothesis 3 is assumed, or XII > XI if Hypothesis 4 is assumed.

All these implications of the hypotheses can be expressed in the form of lattices shown in Figure 2. A link connecting two games shows that their rank order is implied by the corresponding hypothesis, the game with the higher value of C being on the higher level. The rank order is transitive. Where no link connects two games, no rank order is implied. Observe that only Hypotheses 3 and 4 imply a rank order of all seven games.

Let us now compare all of these conclusions with data. There are several sources of data. In addition to the experiment described on page 36, we have performed several other experiments under somewhat varying conditions. Our present aim will be to check the extent to which the conclusions concerning the rank ordering of the games with respect to C are valid. The variations of the first experiment will serve as quasi-replications. We shall at this point make no use of statistical analysis to test the significance of the results. In certain instances significance will be apparent almost to the naked eye. In other cases rather involved procedures would be required to test for significance, since, as we shall see, the distributions are not of the sort which allow an application of con-

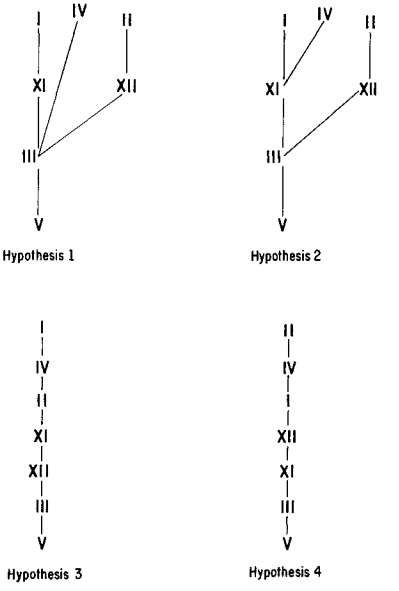

Figure 2. The rank orders of games implied by the various hypotheses. Links connect games whose rank order is implied. Games with greater frequency of cooperative responses are shown on higher levels.

ventional parametric tests. We shall on occasions use nonparametric tests (cf. Appendix II), but for the most part we shall forgo such analysis in order not to detract from the main line of thought. As has already been stated, the generation rather than the establishment of hypotheses has been the main purpose of the work described here.

Variants of the Experiment

We shall now describe the different variants of the experiment with the Prisoner's Dilemma game. The condition under which the experiment has been described will be called the *Pure Matrix Condition*, to indicate that a given pair of subjects plays the same

game (not mixed with other games) throughout the session (three hundred plays) and the subjects see the game matrix displayed. That is to say, their performance on that game is not "contaminated" by their performance on different games. As has already been said, ten pairs play each of the seven games in the Pure Matrix Condition, seventy pairs in all or twenty-one thousand responses.

In the *Block Matrix Condition* each pair of subjects plays all seven games. Since learning has an important effect on performance, the order in which the games are played must be varied in order to distinguish the effects of the payoff matrix from the effects of learning. Accordingly our design in this condition was a Latin Square—seven orders of blocks of fifty plays of each game of the seven. There were two such Latin Squares, the orders in the second being the exact reverse of the orders in the first, or a total of fourteen orders. There were five pairs playing the seven playing games in a given order or seventy pairs in all. Since there were fifty plays per game, each pair gave 350 responses.

In a third condition, called the *Mixed Matrix Condition*, each of ten pairs played seven hundred times or one hundred plays per game. Here the games were presented in random order slightly modified to allow each game to be represented exactly one hundred times. The matrices of the seven games were displayed.

The *Pure No Matrix Condition* was introduced by having pairs of subjects play the games without having the game matrix in front of them. However, the payoffs to both players were announced following each play. The *Mixed No Matrix Condition* is defined analogously.

Thus there were five different variants, all independent, in the sense that the corresponding subject

populations were nonoverlapping—each pair playing in only one condition. In Table 2 the frequencies of

TABLE 2

Game	Pure Matrix Condition	Block Matrix Condition	Mixed Matrix Condition	Combined Matrix Condition	Pure No Matrix Condition	Mixed No Matrix Condition	Combined No Matrix Condition
I	73.4	70.0	70.0	71.5	32.4	33.4	32.7
II	77.4	68.6	60.9	71.3	44.6	30.2	41.0
III	45.8	49.2	61.0	49.4	22.3	34.7	25.4
IV	66.2	67.5	71.6	67.5	45.8	29.2	41.4
V	26.8	39.5	40.4	34.2	22.6	19.1	21.7
XI	63.5	66.0	67.0	65.1	34.5	32.6	34.0
XII	59.4	59.1	60.3	59.4	33.4	26.0	31.5
Hypothesis 1	13/13	13/13	11/13	13/13	12/13	8/13	12/13
Hypothesis 2	14/14	14/14	12/14	14/14	13/14	8/14	13/14
Hypothesis 3	19/21	20/21	17/21	20/21	16/21	13/21	18/21
Hypothesis 4	19/21	18/21	15/21	17/21	16/21	11/21	18/21

the *C* responses (in percent) are shown for each of the seven games and five conditions. In addition, the conditions in which the matrix is displayed are combined (Combined Matrix Condition) and also the two conditions in which the matrix is not displayed (last column). In the last four rows of the table the entries show crude measures of agreement between the rank order of the games as observed and as implied by the four hypotheses. Recall that only Hypotheses 3 and 4 rank order all seven games. Hypotheses 1 and 2 imply only partial ordering. Thus Hypothesis 1 implies the ordering of only 13 of the 21 possible pairs, namely I > XI; I > III; I > V; XI > III; XI > V; III > V; IV > III; IV > V; II > XII; II > III; III > V; XII > III; XII > V. Of course, not all of these inequalities are independent. There are only six independent ones among them. In fact each of the four hypotheses implies an ordering of only six independent pairs. But

the strength of each hypothesis is reflected in the number of paired comparisons *implied* by the independent six pairs. In the case of Hypothesis 1, there are 13 such pairs; in the case of Hypothesis 2, there are 14; each of the last two hypotheses implies an ordering of all the 21 pairs (among seven games taken two at a time). Our crude measure of agreement is simply a fraction showing how many of the paired comparisons implied by each of the hypotheses in each condition are consistent with the corresponding hypotheses. The weak Hypotheses 1 and 2 are completely corroborated in the Pure and the Block Conditions and also in the combined conditions with matrix displayed. The strong Hypotheses 3 and 4 are nowhere completely corroborated. Of the two, Hypothesis 3 seems to be corroborated somewhat more than Hypothesis 4. No further attempt will be made at this point to estimate a confidence level for these hypotheses, which were proposed here not for the purpose of explaining the data but merely as points of departure for a theory. In Part II, we shall develop a mathematical model in which a rank ordering of the games will be derived as a consequence of certain dynamic considerations. At that time the questions raised here will be discussed more fully.

Summary of Chapter 1

If the payoffs are varied singly, common sense suggests that the frequency of cooperative responses ought to increase with the reward parameter R, ought to decrease with the temptation parameter T, and ought to increase as the magnitude of the (negative) punishment parameter P increases. These conjectures, embodied in Hypothesis 1, have been corroborated in the Block and Pure Matrix Conditions.

If the payoffs are equated to utilities and if the

theory of utility used in game theory is assumed, then the behavior of Prisoner's Dilemma ought to depend not on the payoff parameters themselves but on ratios of differences between them. It was shown that only two such ratios are independent.

Two hypotheses have been proposed in the form of directional dependence of cooperative frequencies in each of the two independent difference ratios. Neither of these hypotheses was corroborated perfectly in any of the conditions. One of them, however, was almost completely corroborated in one of the conditions and both were somewhat less completely corroborated in another condition. On the whole, one of the hypotheses was favored somewhat more than the other.