# *Chapter 3*
# *Effects of Interaction*

ONE QUESTION which interests both the psychologist and the layman is what role personality plays as a determinant of performance in a Prisoner's Dilemma game. On the face of it, this question seems eminently sensible. The choice in Prisoner's Dilemma appears to be the choice between competing and cooperating, between conflict and conflict resolution, between trust and suspicion, between loyalty and betrayal. To be sure, everyone recognizes that a contrived laboratory experiment of the sort discussed here is far removed from life and cannot be expected to serve as a reliable test of people's deep attitudes. Still, one might argue, if personality factors are going to emerge in any laboratory experiments, perhaps the ones described here are as good as any, precisely because they do not simulate life in any content-pegged context. One might argue that attempts to simulate life in the laboratory only emphasize the artificiality of the laboratory. The more "realistic" the simulation, the more the participants may be aware of the artificiality and consequently behave in ways unrelated to the ways they would behave in real life. On the other hand, a situation such as the apparently trivial game, Prisoner's Dilemma, because of its very triviality may tap the player's psychological propensities more thoroughly than attempted reproductions of real life.

Another thing in favor of a trivial game repeated hundreds of times is the difficulty of "presenting a front." For example, a subject who knows that others attribute to him a tendency toward hostility in argu-

ment or discussion may deliberately act in a friendly manner in a group dynamics experiment. In a Prisoner's Dilemma experiment it is not immediately apparent to the subject what he is being tested for, and even if he gets the idea that the choice between $C$ and $D$ is essentially a choice between cooperation and pursuit of self-interest, the costs and consequences of each choice are not clear-cut. What happens depends not on him alone but also on what the other does. Under these circumstances it is difficult to decide once and for all how "one will act" (what "role" one will play). Finally the quickness with which the decisions must be made (every few seconds) also militates against a thought-out policy. It can be assumed, therefore, that the patterns of responses contain a good deal of spontaneity and so the situation may, after all, tap some basic attitudes and propensities in a significant way.

It would seem, therefore, that a search for personality correlates of performance in a Prisoner's Dilemma game might reveal something. There have been a few attempts to find such correlates (Deutsch, 1960; Lutzker, 1960).*

We can expect the emergence of personality effects if (1) performances of different but homogeneous populations are compared, or (2) the subjects play the game only a very few times. The first condition was fulfilled in Lutzker's experiments, the second in Deutsch's. However, if the game is played many times and if the pairs are randomly matched, the effect of personality on performance may well be masked by the interaction between the paired players. To take an example, suppose the interaction effects are so strong that the two

* Lutzker's game was, strictly speaking, not Prisoner's Dilemma, since the payoffs were ordered $T > R > S > P$ instead of $T > R > P > S$, as in Prisoner's Dilemma.

players become exact copies of each other—what one does in many repeated plays, the other is sure to do. If this happens the personalities of the two can no longer be reflected in the performance in the sense that some distinguishing features of the individual performances can serve as indices of underlying personality characteristics. Under these conditions attempts to find correlations between individual performances and personality correlates in experiments with repeated games are doomed to failure.

There is a quick way to find out whether the strength of the interaction effects is such that the search for individual correlates in experiments with repeated games is unwarranted. The members of pairs playing Prisoner's Dilemma are randomly matched. Therefore if each is assigned some quantitative index of some personality trait, the covariance or the product moment correlation of the two indices of paired subjects, taken over a large population of pairs, ought to be zero. Further, if the performance were completely determined by the magnitude of this personality index, then the correlation of the performance indices ought to be zero also. In particular the product moment correlation of the frequencies of cooperative choices of two pair members $C_1$ and $C_2$ ought to be near zero when taken over a population of pairs playing the game under the same conditions.

In our experiments just ten pairs played under the same conditions (i.e., the same game) in all cases except the Block Matrix Condition. Here seventy pairs played under the same conditions except for the order in which the games were played. We assume for the moment that the effect of order is not large compared to other effects (such as payoffs, personal propensities, and some chance effects to be discussed below). Under this assumption, if individual personality correlates

play an important part in determining $C$, then the individual indices $C_1$ and $C_2$ ought not to be significantly correlated in the Block Matrix Condition, and any spurious correlation is likely to be inhibited by the comparatively large number of pairs.

The correlations $\rho_{C_1C_2}$ are given in Table 3. Although a population of ten pairs is not large enough to establish confidence in a correlation, we have twenty-three replications represented in Table 3, since each ten pairs is an independent sample.

TABLE 3

The product moment correlations $\rho_{C_1C_2}$ between the frequencies of cooperative responses of paired players. The correlations were not computed for each game separately in the Mixed Conditions.

| Game | Pure Matrix Condition | Block Matrix Condition | Pure No Matrix Condition | Mixed Matrix Condition | Mixed No Matrix Condition |
|---|---|---|---|---|---|
| I | 1.00 | .90 | .96 | | |
| II | .99 | .82 | .83 | | |
| III | .98 | .89 | .62 | | |
| IV | .98 | .87 | .89 | | |
| V | .96 | .95 | .90 | | |
| XI | .92 | .86 | .98 | | |
| XII | .91 | .93 | .23 | | |
| Average | .96 | .89 | .77 | .87 | .96 |

We see from Table 3 that $\rho_{C_1C_2}$ are all positive and for the most part very large. In the Mixed Matrix Condition this could be attributed to the similar effect of each game, i.e., to the fact that some games elicit more cooperation and some less cooperation in all players. However, as we have seen, the games are only weakly differentiated in the Mixed Matrix Condition and hardly at all in the Mixed No Matrix Condition; so this effect cannot be very important. We surmise, therefore, that the consistently high corre-

lations are results of strong interaction effects between paired players. What one does the other is also likely to do.

*Interaction within Individual Sessions*

There is another way of detecting interaction effects in Prisoner's Dilemma. A given string of plays yields a "protocol," i.e., a sequence of "states" labeled *CC*, *CD*, *DC*, or *DD*. Suppose there were a strong positive interaction—one player tended to behave like the other. Then the matched responses, *CC* and *DD*, would predominate, while the unilateral ones, *CD* and *DC*, would be rare. Suppose, on the contrary, there were a strong negative interaction—each player would tend to do the opposite of what the other did. In this case the "unilateral" states *CD* and *DC* would predominate, while the matched states would be rare. From our examination of the correlation of $C_1$ vs. $C_2$, we already know that the interaction tends to be strongly positive. This correlation, however, reflects only the gross interaction effect—how dependent is the overall degree of cooperation shown by one player on that shown by the other. We wish to examine the effects of the interaction more closely. We wish to see whether it operates in the sequence of individual plays. To do this, we assign arbitrarily value 1 to *C* and value 0 to *D* and examine the product moment correlation coefficients of the random variables $C_1$ and $C_2$ related to the two players, which take on values 1 and 0. Actually any other two values will yield exactly the same result, the product moment correlation being independent under a linear transformation of variables. The formula for the correlation of two random variables taking on either of two given values is

$$\rho_0 = \frac{(CC)(DD) - (CD)(DC)}{\sqrt{(CC + CD)(CC + DC)(DD + CD)(DD + DC)}}. \qquad (10)$$

This index can be calculated for each pair of players. The Pure Matrix Conditions are the most suitable for this purpose, since when the players play the same game all the way through, the value of $\rho_0$ can be attributed entirely to the interaction between them and not to some extraneous variables such as payoffs, which may be affecting both players in the same way and thus contributing to a positive bias for $\rho_0$.

The values of $\rho_0$ among the seventy pairs of the Pure Matrix Condition are overwhelmingly positive (sixty-two out of seventy pairs). There is also a positive bias in the Pure No Matrix Condition, although a weaker one. The positive bias in the Block Matrix Condition is even stronger than in the Pure Matrix Condition.

Table 4 shows the average values of $\rho_0$ in the several conditions.

TABLE 4

| Condition | Pure Matrix | Block Matrix | Mixed Matrix | Pure No Matrix | Mixed No Matrix |
|---|---|---|---|---|---|
| Average $\rho_0$ | .46 | .56 | .47 | .28 | .34 |

Note that the average values of $\rho_0$ are smaller in both No Matrix Conditions: in the absence of the matrix the interactions are apparently weaker. The correlation is highest in the Block Matrix Condition, where the same pair plays all seven games and where one can therefore expect the two players to vary together from game to game.

The coefficient $\rho_0$ does not actually reflect play-to-play interactions precisely. An interaction of this sort is a response of a player to what the other player did on the immediately *preceding* play. In order to get a measure of this interaction, we should define the states *CC*, *CD*, *DC*, and *DD* in such a way that one of

the pair of responses represents the preceding play by the other player. This can be easily done if the protocols are "shifted" by one play. In fact this can be done in two ways by having the first player or the second player lag by one play. Each protocol then yields two "shifted" protocols, both representing a correlation coefficient which we designate by $\rho_1$.

We expect $\rho_1$ to show a larger positive bias than $\rho_0$, because the positive interaction will be more precisely reflected in the former. Generalizing the idea of shifting the protocol, we can define analogously $\rho_2$, $\rho_3$, $\rho_4$, etc. Each of these will measure the degree of interaction of a player's response to the responses of the other 2, 3, 4, etc., plays ago. We would expect the strength of the interaction to decay with the interval, and, if it does, we can get an idea of the "memory" component of the interaction. The values of the $\rho$'s in all the conditions are shown in Table 5.

TABLE 5

| Condition | $\rho_0$ | $\rho_1$ | $\rho_2$ | $\rho_3$ | $\rho_4$ | $\rho_5$ | $\rho_6$ |
|---|---|---|---|---|---|---|---|
| Pure Matrix | .46 | .51 | .46 | .42 | .40 | .38 | .36 |
| Block Matrix | .56 | .59 | .56 | .52 | .51 | .49 | .46 |
| Mixed Matrix | .47 | .34 | .31 | .30 | .31 | .30 | .29 |
| Pure No Matrix | .37 | .47 | .40 | .33 | .32 | .32 | .32 |
| Mixed No Matrix | .34 | .22 | .22 | .23 | .21 | .21 | .20 |

Note that the value of $\rho_1$ is largest, as expected, except in both mixed conditions. These exceptions are also understandable, since in the Mixed Matrix Condition, the next play in general involves a different game. Evidently "not only what the other did last" but also the game payoffs play a part in influencing the response, and so the effect of imitating the other's last response is obscured in these conditions.

### *The Lock-in Effect*

The strength of interaction can be very clearly seen in the "lock-in" effect. Consider a fifty-play block as a unit of analysis. Each such unit can be characterized by a fraction of times a given state occurs in it, for example *CC*. Thus a number is assigned to each fifty-play block—the fraction of *CC* responses contained in it. We can now examine the distribution of fifty-play blocks with regard to this index. If the distribution were of a type resembling the normal distribution, the mode (i.e., the index represented by the largest number of blocks) would be near the mean value of the index. The plot would have a peak somewhere near the middle and would taper off at the ends, i.e., would exhibit a "bell-shaped" curve. The plots of the *CC* distributions for the Pure and the Block Matrix Conditions are shown in Figures 4 and 5. We note that the picture is exactly the opposite of the one expected from a normal-type distribution. The concentrations are at the ends rather than in the middle.

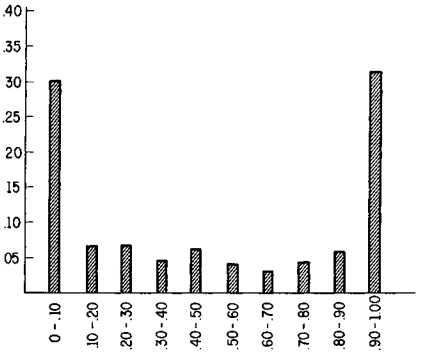


Figure 4. Horizontal: fraction of *CC* responses in a block of fifty plays. Vertical: fraction of fifty-play blocks corresponding to each fraction of *CC* responses (Pure Matrix Condition).

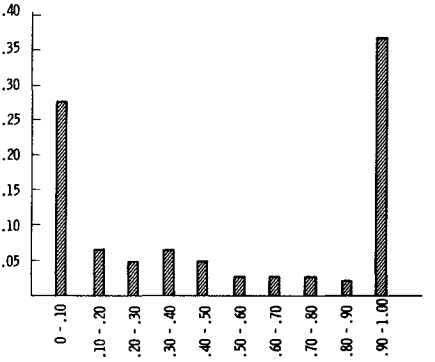


Figure 5. Horizontal: fraction of *CC* responses in a block of fifty plays. Vertical: fraction of fifty-play blocks corresponding to each fraction of *CC* responses (Block Condition).

This effect is not as apparent when the *DD* fractions are plotted, because of the relative paucity of almost total *DD* blocks in the Pure and in the Block Matrix Conditions. However, in the No Matrix Conditions, where the *DD* blocks are plentiful, a similar bi-modal distribution is observed with respect to the fraction of *DD* responses (cf. Figure 6).

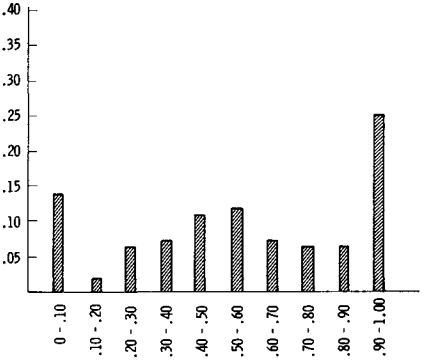


Figure 6. Horizontal: fraction of *DD* responses in a block of fifty plays. Vertical: fraction of fifty-play blocks corresponding to each fraction of *DD* responses (Pure No Matrix Condition).

The study of interaction effects indicates that these effects are very strong. They tend to make the members of a pair behave like each other. Later, when we examine the time course of these effects, we shall find that they become progressively stronger as the session continues. Moreover, the interaction effects tend to throw the performance toward one or the other extreme. The sessions tend to become either predominantly cooperative or predominately uncooperative. Of course the payoff matrix contributes to the outcome: the "mild" games (Games I, II, and IV) tend toward the cooperative extreme; the "severe" games (Games III and V) tend to the uncooperative extreme. Also the condition contributes to the outcome. When the matrix is displayed, more performances tend toward the cooperative extreme. In the No Matrix Conditions, the opposite is the case. However, both the extremes are found in all games and in both conditions. This leads to the conclusion of the inherent *instability* of the Prisoner's Dilemma situation. "Compromises" are comparatively rare. The pair will be thrown either toward a tacit collusion, which remains rather stable, or toward the *DD* trap, also rather stable.

There is no a priori reason why a pair should go to the one extreme rather than to the other since both are "rationalizable." The *DD* trap appears as an eminently reasonable "solution," since neither player can justify a departure toward cooperation on rational grounds. "If I cooperate," he can well say, "the other will simply take advantage of it and secure the big gain for himself." But also the *CC* tacit collusion is justifiable. If the pair have locked-in on it, either can justify refraining from defecting by pointing out that a defection will only break up the collusion: the other is also sure to defect, and the end result will be a lock-in in the *DD* trap to the disadvantage of both. Although

this argument also sounds eminently reasonable, it is well to keep in mind that it is demolished by the game-theoretical prescription of the totally uncooperative strategy $D^{(n)}$, when $n$ is known and finite.

Table 6 shows the fractions of pairs in each condition who have locked-in on *CC* and on *DD* respectively. Our criterion of a lock-in is 23 or more *CC* (*DD*) responses out of the last 25.

TABLE 6

| Condition | $L_{CC}$ | $L_{DD}$ | Total Fraction Locked in |
|---|---|---|---|
| Pure Matrix | .53 | .17 | .70 |
| Block Matrix | .56 | .07 | .63 |
| Mixed Matrix | .43 | .43 | .86 |
| Pure No Matrix | .16 | .43 | .59 |
| Mixed No Matrix | .20 | .40 | .60 |

$L_{CC}$: Fraction of pairs locked-in on *CC* in the last twenty-five plays.
$L_{DD}$: Fraction of pairs locked-in on *DD* in the last twenty-five plays.

The predominance of the *CC* lock-ins in the favorable conditions shows that our subjects are not sufficiently sophisticated game-theoreticians to have figured out that $D^{(n)}$ is the only strategically defensible strategy. Apparently this lack of strategic sophistication allows many of them to find the commonsense solution, namely the tacit collusion, and so to win money instead of losing it.

### *Summary of Chapter 3*

The interaction effect in repeated plays of Prisoner's Dilemma is strong and positive. In the single sessions, a pronounced tendency is observed of each player to imitate the other. The product moment correlation of the frequencies of cooperative responses of paired subjects is in many cases very nearly plus one.