Chapter 4
The Contingent Propensities

WE HAVE SEEN that in repeated plays of Prisoner's Dilemma positive interaction effects are operating. The strongest manifestation of those effects is in the correlations between the frequencies of cooperative choices of the two players. It is fairly clear, therefore, that a search of individual personality correlates of this frequency will not be rewarding. How can we get at indices of performance which are less dependent on interaction effects and therefore more suitable as possible indicators of personal propensities? We shall now describe our attempts to get at such indices.

The Response-Conditioned Propensities

Consider the four conditional probabilities defined as follows:

ξ_1: the probability that player 1 responds cooperatively following player 2's cooperative response on the preceding play.

η_1: the probability that player 1 responds cooperatively following his own cooperative response on the preceding play.

ζ_1: the probability that player 1 responds cooperatively following his own defecting response on the preceding play.

ω_1: the probability that player 1 responds cooperatively following player 2's defecting response on the preceding play.

The conditional probabilities ξ_2, η_2, ζ_2, and ω_2 are defined analogously with the roles of players 1 and 2 reversed.

It is clear that these probabilities are components of the C's. Thus, by definition of the conditional probabilities, we must have the following equations satisfied:

$$\xi_1 C_2 + \omega_1(1 - C_2) = C_1; \tag{11}$$

$$\eta_1 C_1 + \zeta_1(1 - C_1) = C_1; \tag{12}$$

$$\eta_2 C_2 + \zeta_2(1 - C_2) = C_2; \tag{13}$$

$$\xi_2 C_1 + \omega_2(1 - C_1) = C_2. \tag{14}$$

If we solve Equations (12) and (13) for C_1 and C_2 respectively, we obtain

$$C_1 = \frac{\zeta_1}{1 + \zeta_1 - \eta_1}; \tag{15}$$

$$C_2 = \frac{\zeta_2}{1 + \zeta_2 - \eta_2}. \tag{16}$$

If we now substitute these values into Equations (11) and (14), we obtain

$$\frac{\xi_1\zeta_2 + \omega_1(1 - \eta_2)}{1 + \zeta_2 - \eta_2} = \frac{\zeta_1}{1 + \zeta_1 - \eta_1}; \tag{17}$$

$$\frac{\xi_2\zeta_1 + \omega_2(1 - \eta_1)}{1 + \zeta_1 - \eta_1} = \frac{\zeta_2}{1 + \zeta_2 - \eta_2}. \tag{18}$$

Thus two constraints are imposed upon our eight conditional probabilities. Consequently, only six of them are mathematically independent. In the symmetrical case, when $\xi_1 = \xi_2$; $\eta_1 = \eta_2$, etc., relations (17) and (18) reduce to

$$\frac{(1 - \eta)}{(1 - \xi)} = \frac{\zeta}{\omega} \tag{19}$$

and so instead of four independent parameters, we have only three. In the C's, we had two independent parameters in the general case and only one in the symmetric case (since $D = 1 - C$). Thus in either case, we have trebled the number of independent parameters and so have refined our indices to that extent.

Let us now see how these conditional probabilities are correlated across pairs. Tables 7 and 8 show the values of the paired conditional probabilities in the Pure Matrix Condition and in the Pure No Matrix Condition, as well as the correlations between the paired indices. We see that the interaction effect on the conditional probabilities is still large, but it is not as large as the effect on C.

Let us now interpret ξ, η, ζ, and ω psychologically and make appropriate conjectures. We shall call these indices *response-conditioned propensities*. From its definition ξ appears to be a propensity to respond cooperatively to the other's cooperative response. The conditional probability η, on the other hand, is the propensity to "respond" cooperatively to one's own cooperative choice, in other words to *continue* to respond cooperatively. We note from Tables 7 and 8 that

TABLE 7

Mean values of response-conditioned cooperative propensities and correlations between paired values in the Pure Matrix Condition. The last column shows correlations between paired C frequencies for comparison.

Game	ξ	η	ζ	ω	ρ_ξ	ρ_η	ρ_ζ	ρ_ω	ρ_C
I	.88	.89	.36	.36	.93	.98	.61	.84	1.00
II	.89	.87	.41	.34	.86	.98	.17	.01	.99
III	.65	.67	.19	.19	.82	.68	.83	.71	.98
IV	.84	.84	.24	.26	.81	.84	.83	.69	.98
V	.57	.59	.12	.15	.85	.83	.91	.75	.96
XI	.82	.88	.22	.25	.93	.70	.93	.36	.92
XII	.68	.74	.21	.24	.85	.93	.90	.72	.91
Mean	.76	.78	.25	.26	.87	.85	.74	.58	.96

the values of ξ and η are quite close to each other. Since the bulk of cooperative responses, at least in the Pure Matrix Condition comes from the locked-in runs CC, a great many cooperative responses following the other's cooperative responses coincide with the coop-

TABLE 8

Mean values of response-conditioned cooperative propensities and correlations between paired values in the Pure No Matrix Condition. The last column shows correlations between paired *C* frequencies for comparison.

Game	ξ	η	ζ	ω	ρ_ξ	ρ_η	ρ_ζ	ρ_ω	ρ_C
I	.55	.58	.21	.23	.87	.81	.94	.96	.96
II	.52	.57	.29	.31	.87	.89	.27	.23	.83
III	.46	.56	.12	.15	.77	.79	−.09	−.37	.62
IV	.64	.65	.22	.22	.50	.64	.82	.76	.89
V	.40	.47	.12	.14	.54	.63	.63	.16	.90
XI	.56	.63	.11	.12	.99	.77	.96	.94	.98
XII	.45	.55	.18	.22	.63	.65	.60	.26	.23
Mean	.51	.57	.18	.20	.74	.74	.59	.42	.77

erative responses following one's own cooperative responses. There is thus a large overlap in the two sets of responses with respect to which ξ and η are calculated. To a certain extent this is true also of noncooperative responses, since the bulk of these comes from the locked-in *DD* runs. Therefore we would expect the values of ζ and ω also to be close to each other, which they are.

Let us now interpret ζ and ω. It is more instructive to look at their complements $1 - \zeta$ and $1 - \omega$. The former appears from its definition to be a measure of the persistence in the *D* response; the latter is related to the propensity to respond noncooperatively to the other's defecting response (i.e., a measure of "vengefulness"). Now on the basis of the high correlations of ξ and η and the comparatively lower correlations of ζ and ω observed in both conditions we can make the following conjectures with respect to the response-conditioned propensities.

1. The two players tend to become like each other with respect to a propensity to respond cooperatively to self's and other's cooperative responses.

2. To a lesser degree the two players tend to become like each other with respect to a propensity to respond noncooperatively to self's and other's defecting responses.

The second statement is a consequence of the fact that the correlations between $1 - \zeta_1$ and $1 - \zeta_2$ and between $1 - \omega_1$ and $1 - \omega_2$ must be equal respectively to those between ζ_1 and ζ_2 and between ω_1 and ω_2.[13] Incidentally, it is clear from this argument that the high correlations between ξ_1 and ξ_2 and between η_1 and η_2 are not consequences of the persistently high values of these variables, since the values of $1 - \zeta$ and $1 - \omega$ are also persistently high, while the correlations between those paired variables are considerably lower. A conclusion is therefore warranted that high interaction effects are operating on the propensities to respond cooperatively to self's and other's cooperative responses but not as much on the propensities to respond noncooperatively to self's and other's defecting responses. The latter propensities are in themselves high but are apparently not subjected to quite as strong interaction effects as the former.

The State-Conditioned Propensities

Next we introduce four other conditional probabilities, which we shall call the *state-conditioned propensities*. These are:

x: the probability that a player will choose cooperatively, following a play in which he chose cooperatively and received R (i.e., following a play in which both players chose cooperatively).

y: the probability that a player will choose cooperatively following a play in which he chose cooperatively and received the sucker's payoff S (i.e., following a play in which he was the lone cooperator).

z: the probability that a player will choose coop-

eratively following a play on which he defected and received T (i.e., following a play on which he was the lone defector).

w: the probability that a player will choose cooperatively following a play on which he defected and received P (i.e., following a play on which both defected).

Like the response-conditioned propensities, the state-conditioned propensities are defined for player 1 and for player 2. Unlike the former, however, all eight of the latter are mathematically independent (four in the symmetric case). Thus we have in the state-conditioned propensities a still more refined set of indices.

The propensity x indicates a willingness to continue the tacit collusion (implied, by definition, to have been achieved on the previous play). This willingness is associated with a willingness to resist the temptation to defect, which is always present. It therefore suggests something like "trustworthiness."

The propensity y indicates a willingness to persist in cooperating, even though one has been "betrayed." It therefore suggests either "forgiveness" or "martyrdom" or a strong faith in teaching by example, or, perhaps, stupidity, depending on the ethical views of whoever evaluates this behavior.

The propensity z indicates a willingness to stop defecting in response to the other's cooperative choice. It may indicate "repentance" or "responsiveness."

Finally w indicates a willingness to *try* cooperating as a way of breaking out of the DD trap. Clearly this action is justifiable only if a certain amount of trust in the responsiveness of the other exists in the initiator of cooperation. Hence w suggests "trust."

As we have said, the conditional propensities x, y, z, and w are refinements of the gross cooperative index C, and so one reason for examining them is to

look at the data with greater resolving power, as it were. The other reason, as already stated, is derived from our search for indices relatively unaffected by the interaction process, since such independence makes an index a more suitable variable which one might try to associate with personality traits or with experimental conditions.

We shall give a mathematical argument for the conjecture that x, y, z, and w may be less affected by interaction than C. If this is true, it will be reflected in weaker correlations of x_1 vs. x_2, y_1 vs. y_2, etc., than we have observed in C_1 vs. C_2. We shall not attempt to derive general conditions under which this situation obtains, but will confine ourselves to a very special case, where it is shown to obtain and hence establishes such a possibility.

The Two Simpletons

Imagine Prisoner's Dilemma played by two simpletons, or, if you prefer, by two automata with exceedingly simple reactions to reinforcements. By and large, these simpletons or automata are characterized by an aversion to negative payoffs. Thus, whenever one of them has made an unreciprocated cooperative choice, he is certain to defect the next time. Also, when both have defected (and therefore received negative payoffs), both change their responses, so that the next response is inevitably *CC*. If the simpletons were equally consistent with regard to positive payoffs, i.e., if they always repeated their choice whenever they got a positive payoff, then, it can be easily seen that the two would "lock-in" on the *CC* response after at most two plays. For if the first play happened to be *CC*, it would remain so thereafter. If the first play happened to be *DD*, it would change to *CC* and remain so. If the first play happened to be *CD* or *DC*, the cooperator

would defect on the following play, while the defector would continue to defect (having got a positive payoff, T). But this would lead to DD and therefore to CC next. For these two simpletons, Prisoner's Dilemma is not a dilemma. They always achieve tacit collusion almost immediately.

We shall, however, introduce a temptation into our simpletons' psyches. When they are in the CC state, each, we shall assume, is tempted to defect. He defects only at times, namely with probability $1 - x$. It follows that he sticks to the tacit collusion with probability x, the same conditional probability, which we have already defined as one of our conditional propensities, namely "trustworthiness." For the rest, we see that the other conditional propensities of our simpletons must be assigned the following values: $y = 0$, $z = 0$, $w = 1$. This is a consequence of the way we have defined their reactions to the payoffs. To summarize, the only variable in the psychological makeups of our simpletons is described by a single parameter, x—the propensity for repeating a rewarded cooperative choice. We shall suppose that this propensity is distributed in some way throughout the population of simpletons, from which we recruit our subjects, whom we then pair randomly and let play Prisoner's Dilemma some large number of times.

Since the pairs are matched randomly, it follows that the correlation of x_1 vs. x_2 in the population will not vary significantly from zero. Let us see what will be the case with C_1 and C_2.

The cooperative response frequencies C_1 and C_2 can be computed from the x's. In fact, if we suppose that x_1 and x_2 are probabilities of independent events, the resulting process can be viewed as a four-state Markov chain, whose transition probabilities are compounded from x_1 and x_2 and their complements (some transition

probabilities being zero or one). If we solve for the steady-state distribution of states in this stochastic process, we obtain the following values of the asymptotic state probabilities:[14]

$$C_1C_2 = \frac{1}{2 + x_1 + x_2 - 3x_1x_2}; \tag{20}$$

$$C_1D_2 = \frac{x_1(1 - x_2)}{2 + x_1 + x_2 - 3x_1x_2}; \tag{21}$$

$$D_1C_2 = \frac{x_2(1 - x_1)}{2 + x_1 + x_2 - 3x_1x_2}; \tag{22}$$

$$D_1D_2 = \frac{1 - x_1x_2}{2 + x_1 + x_2 - 3x_1x_2}. \tag{23}$$

Now $C_1 = C_1C_2 + C_1D_2$, while $C_2 = C_1C_2 + D_1C_2$. We can therefore, write

$$C_1 = \frac{1 + x_1 - x_1x_2}{2 + x_1 + x_2 - 3x_1x_2}; \tag{24}$$

$$C_2 = \frac{1 + x_2 - x_1x_2}{2 + x_1 + x_2 - 3x_1x_2}. \tag{25}$$

We wish to exhibit a situation in which x_1 and x_2 are not correlated, while C_1 and C_2 are positively correlated. Consider a population of pairs in which the members designated as "player 1" all have the same x_1, while those designated as "player 2" have different values of x_2. Clearly, in this population sample, x_1 and x_2 will be uncorrelated. However, under certain conditions, C_1 and C_2 will be positively correlated. To see this, consider the partial derivatives[15] of C_1 and C_2 with respect to x_2. We have

$$\frac{\partial C_1}{\partial x_2} = \frac{2x^2 - 1}{(2 + x_1 + x_2 - 3x_1x_2)^2}; \tag{26}$$

$$\frac{\partial C_2}{\partial x_2} = \frac{1 + 2x_1 - x_1^2}{(2 + x_1 + x_2 - 3x_1x_2)^2}. \tag{27}$$

From these equations, we see that $\partial C_1/\partial x_2$ is positive when $x_1 > .707$, while $\partial C_2/\partial x_2$ is always positive.

Therefore if $x_1 > .707$, both C_1 and C_2 increase if x_1 remains constant and x_2 increases, and both decrease if x_1 remains constant and x_2 decreases. But this is precisely the situation in our hypothetical sample, where the x_1's are all equal and the x_2's vary. If, therefore, the x_1's in our sample are sufficiently large, C_1 and C_2 will be positively correlated even though the correlation of x_1 and x_2 is zero.

This admittedly very special case was offered as an illustration of how a positive correlation can be found among the gross cooperative frequencies even though the underlying conditional propensities which give rise to these frequencies may be uncorrelated.

Generalizing the argument, it is conceivable to have weak positive correlations between the paired state-conditioned propensities and stronger positive correlations among the paired C frequencies. Turning the argument around, we can expect under similar circumstances to find weaker correlations among the propensities than we observe among the C frequencies. But this is precisely what we are seeking—indices of performance which lie "deeper" and so are less subjected to interaction effects than the gross indices of overt performance.

We turn to our data to examine the values of the state-conditioned propensities and of the correlations among them in the Pure Matrix Conditions. These are shown in Tables 9 and 10.

From Tables 9 and 10 we see that although the correlations ρ_x are predominantly positive, their mean is smaller than that of ρ_ξ (cf. Tables 7 and 8) which in turn are lower than that of ρ_C. The ρ_w are also positive and are still smaller. As for ρ_y and ρ_z, they oscillate rather wildly, and so their significance is open to question. We shall not pursue a rigorous investigation of

TABLE 9

Mean values of state-conditioned cooperative propensities and correlations between paired values in the Pure Matrix Condition. The last column shows correlations between paired *C* frequencies for comparison.

Game	x	y	z	w	ρ_x	ρ_y	ρ_z	ρ_w	ρ_C
I	.92	.45	.45	.30	.99	.61	−.61	.96	1.00
II	.93	.42	.53	.32	.66	−.01	.08	.16	.99
III	.83	.35	.34	.15	.91	.02	−.44	.91	.98
IV	.91	.42	.40	.18	.68	−.15	−.45	.48	.98
V	.71	.28	.25	.05	.96	−.39	−.34	.14	.96
XI	.85	.43	.39	.20	.98	−.35	.63	.59	.92
XII	.79	.44	.33	.20	.36	−.46	−.61	.15	.91
Mean	.84	.40	.38	.20	.79	−.10	−.25	.48	.96

TABLE 10

Mean values of state-conditioned cooperative propensities and correlations between paired values in the Pure No Matrix Condition. The last column shows correlations between paired *C* frequencies for comparison.

Game	x	y	z	w	ρ_x	ρ_y	ρ_z	ρ_w	ρ_C
I	.66	.36	.25	.14	.47	−.04	.51	.69	.96
II	.70	.34	.28	.25	.76	−.26	.26	.29	.83
III	.68	.37	.20	.09	.88	.24	.13	−.18	.62
IV	.75	.33	.32	.22	.54	−.22	.02	.60	.89
V	.67	.33	.24	.08	.44	.29	.11	.26	.90
XI	.72	.34	.34	.16	−.09	−.74	.38	.64	.98
XII	.69	.40	.22	.11	.52	−.72	−.25	.54	.23
Mean	.70	.35	.26	.15	.50	−.21	.17	.41	.77

the level of significance of our correlations but will rely instead on a rough argument.

Consider the complete table of (mean) correlations among all the state-conditioned propensities, x_1, x_2, y_1 . . . w_2, as shown in Table 11.

Note that of the 28 correlations in Table 11, 24 can be matched into 12 pairs, where one member of the pair can be obtained from the other by interchanging

TABLE 11

The correlation matrix of the state-conditioned propensities in the Pure Matrix Condition.

	x_1	y_1	z_1	w_1	x_2	y_2	z_2	w_2
x_1	1.000	.201	.236	.194	.745	.034	.261	.271
y_1	.201	1.000	−.120	.083	.051	−.146	.329	.026
z_1	.236	−.120	1.000	.444	.028	.150	−.189	.537
w_1	.194	.083	.444	1.000	.197	.272	.099	.556
x_2	.745	.051	.028	.197	1.000	.116	.382	.162
y_2	.034	−.146	.150	.272	.116	1.000	.078	.125
z_2	.261	.329	−.189	.099	.382	.078	1.000	.052
w_2	.271	.026	.537	.556	.162	.125	.052	1.000

subscripts 1 and 2. Thus designating the correlations by ρ with appropriate subscripts, we can match them as follows: $\rho_{x_1y_2}$ with $\rho_{x_2y_1}$; $\rho_{x_1z_2}$ with $\rho_{x_2z_1}$, etc. The significance of the matching is that in any reasonably large population the two members of each pair ought to be nearly equal. This follows from the random assignment of designation 1 or 2 to the players, i.e., from the fact that we ought to consider our population of players 1 identical with our population of players 2, if the labeling of the players makes no difference. Hence the correlation $\rho_{x_1y_2}$ ought to be nearly equal (except for statistical fluctuations) to correlation $\rho_{x_2y_1}$, $\rho_{x_1z_2}$ to $\rho_{x_2z_1}$, etc.

From the actual differences between these pairs we can estimate (very roughly) the limits of the statistical fluctuations of the ρ's.

Another way to obtain an idea about the significance of these correlations is to treat the several conditions as replications. If a given correlation varies widely among the conditions, this may be due to either statistical fluctuations or to the differences in the conditions. But if a correlation has a consistently large value in all or in most of the conditions, this is an indication of its "reality" and relative independence of the varied conditions.

Table 12 shows the correlations in all the conditions. The correlations appear twice in each condition showing the values when i and j are interchanged. The differences between these paired correlations are the strongest indicators of the expected magnitude of their statistical fluctuations.

TABLE 12

The entries shown for $\rho_{x_iy_i}$ designate $\rho_{x_1y_1}$ and ρ_{x_2,y_2}; the entries shown for $\rho_{x_iy_j}$ designate ρ_{x_1,y_2} and ρ_{x_2,y_1}, etc. Apparently anomalous values (those which show rather large discrepancies from their paired values) are boxed.

Condition	Pure Matrix	Pure No Matrix	Block Matrix	Mixed Matrix	Mixed No Matrix
$\rho_{x_1x_2}$	.745	.430	.563	.232	.635
$\rho_{y_1y_2}$	−.146	−.156	−.360	−.372	.687
$\rho_{z_1z_2}$	−.189	.148	.161	.002	.434
$\rho_{w_1w_2}$	.556	.630	.181	.966	.909
$\rho_{x_iy_i}$	.201; .116	.230; .336	.264; .125	.702; .666	.373; .338
$\rho_{x_iz_i}$	.236; .382	.236; .013	.102; −.123	.218; .493	−.180; [.442]
$\rho_{x_iw_i}$	.194; .162	.046; .125	−.013; .048	.179; .177	.088; .117
$\rho_{y_iz_i}$	−.120; .078	.124; −.051	−.111; [.331]	−.085; [.734]	.432; .249
$\rho_{y_iw_i}$	.083; .125	.115; .112	.085; .135	−.160; .231	.785; .592
$\rho_{z_iw_i}$	.444; .052	.475; .234	.276; .100	.849; .366	.826; .743
$\rho_{x_iy_j}$	.034; .051	.030; −.039	−.320; −.105	−.254; −.312	.330; .068
$\rho_{x_iz_j}$	.261; .028	.201; −.019	.025; −.123	.241; .224	.405; .193
$\rho_{x_iw_j}$	.271; .197	.204; −.034	−.094; −.187	.323; .237	.028; .145
$\rho_{y_iz_j}$	.329; .150	−.051; −.135	.045; .194	.190; .125	.364; .597
$\rho_{y_iw_j}$	.026; .272	.129; −.107	.057; .087	.052; .230	.629; .841
$\rho_{z_iw_j}$	[.537]; .099	.321; .254	−.085; .232	.781; .209	.809; .518

From the table we can make reasoned guesses about the nature of the correlations.

From the entries for $\rho_{x_1x_2}$ and $\rho_{w_1w_2}$ we can assume that the x's and the w's are still positively correlated. That is to say, the tendency of one player to lock-in on *CC* or on *DD* elicits a similar tendency in the other.

Next, we may assume that the "true" correlations $\rho_{y_iz_i}$, $\rho_{x_iy_i}$ and possibly $\rho_{x_iw_i}$ are probably near zero, because of the more or less symmetric distributions of these correlations around zero. The remaining correlations

show a weak positive bias, which shows that they represent overall linked tendencies to cooperate or not to cooperate in the two players. The one exception is $\rho_{y_1y_2}$, which is negative in four of the conditions. If it were not for the large positive value of this correlation in the Mixed No Matrix Condition, we could surmise that a forgiving (martyr) inclination on the part of one player tends to inhibit a similar inclination in the other, which would be an interesting finding. Unfortunately, this result does not hold up in the Mixed No Matrix Condition. On the other hand, the correlations are on the whole more strongly positive in this condition than in any other.

A comparison of the means of all the correlations in each of the conditions is shown in Table 13.

TABLE 13

Condition	Pure Matrix	Pure No Matrix	Block	Mixed Matrix	Mixed No Matrix
$\bar{\rho}$	.18	.14	.05	.26	.44

The differences are very likely larger than can be accounted for by statistical fluctuations. However, an explanation of the differences does not easily occur. We cannot surmise that more "mixing" tends either to increase or to decrease the correlations, because the "mixing" is medium in the Block Condition, while the mean value of the correlations is lowest. Nor can we say that the displayed matrix tends to increase or to decrease the correlations, for these effects are opposite in the Pure and in the Mixed Matrix Conditions.

Only one conclusion seems justified: the state-conditioned propensities are still positively correlated, like the *C* frequencies, but not nearly as strongly as the latter. The most pronounced correlations are those

between the x's and the w's. Doubtless these reflect the lock-in effect. The remaining correlations are rather low, except for several in the Mixed No Matrix Condition, an effect which we have not attempted to explain.

An interesting negative result should be mentioned. On a priori grounds, we might expect the correlations $\rho_{y_i z_i}$ to show a negative bias. Such a bias would indicate a tendency on the part of the unilaterally defecting player to exploit the unilaterally cooperative player. However, the bias, if any, of this correlation is in the positive direction, as is the case with almost all the others. That is to say, a "forgiving" attitude of one elicits (if anything) a relenting attitude in the other. The effect is admittedly slight.

In summary, we have in the state-conditioned propensities indices of performance which show more promise as indices of individual characteristics, because they are less subject to interaction effects than the gross frequencies of cooperative choices.

Comparison of the Propensities

We see that in passing from the unconditional cooperative propensities (C) to the response-conditioned propensities (ξ, η, ζ, ω), to the state-conditioned propensities (x, y, z, w) we get indices of performance progressively more "immune" to interaction effects, hence presumably more stable propensities, possibly reflecting deeper psychological traits. We see also that in the case of at least two state-conditioned propensities, namely x and w, we have not yet attained our goal of getting an index independent of interaction. We could go to still "higher" (or deeper) indices by defining conditional probabilities of still higher order (e.g., the probability of responding cooperatively after exactly one, exactly two, etc., CC responses).

There is, of course, the usual price to pay for this, namely a greater complexity of the appropriate analytical apparatus and the need for more massive data to cut down the fluctuations. We shall therefore not pursue this road at this time. Instead we shall examine the indices ξ, η, ζ, ω, and x, y, z, w for their own sake. We have already proposed psychological interpretations for these indices. Now we shall draw some inferences from their magnitudes and the relations among them.

We compare the magnitudes of ξ, η, ζ, and ω in each of the seven games (cf. Tables 7 and 8). We find that in the Pure Matrix Condition, the inequality

$$\eta \geq \xi \geq \omega \geq \zeta \tag{28}$$

is satisfied in all cases with two exceptions: in Game II $\zeta > \omega$ and, $\xi > \eta$. In the No Matrix Condition, there are no exceptions.

Our interpretation of inequality (28) is the following. One's own cooperative choices have a slightly greater tendency to make one's subsequent responses cooperative than the other's cooperative responses and similarly for noncooperative responses. In other words, while it is true that one is more likely to respond cooperatively to the other's cooperative response than to the other's defecting response ($\xi > \omega$), one is still more likely to continue one's own cooperative responses ($\eta > \xi$) and also one's own noncooperative responses ($\zeta < \omega$).

We now introduce the complements of ζ and ω, namely $1 - \zeta$ and $1 - \omega$. These are, of course, the tendencies to respond noncooperatively to one's own and to the other's defecting response respectively. In the Pure Matrix Condition, these satisfy the following inequalities in each of the games:

From these we see that in the games where cooperation is lowest (III, XII and V) the tendency to respond

TABLE 14

Game I	$\eta > \xi > 1 - \omega \geq 1 - \zeta$
Game II	$\xi > \eta > 1 - \omega > 1 - \zeta$
Game III	$1 - \zeta \geq 1 - \omega > \eta > \xi$
Game IV	$\eta \geq \xi > 1 - \zeta > 1 - \omega$
Game V	$1 - \zeta > 1 - \omega > \eta > \xi$
Game XI	$\eta > \xi > 1 - \zeta > 1 - \omega$
Game XII	$1 - \zeta > 1 - \omega > \eta > \xi$

noncooperatively to the other's noncooperative response $(1 - \omega)$ is greater than the tendency to respond cooperatively to the other's cooperative response (ξ). In the games with the most cooperation (I, II, IV and XI) the tendency to respond cooperatively to the other's cooperative response is greater than the tendency to retaliate.

In the No Matrix Condition, the corresponding inequalities are as shown on Table 15.

TABLE 15

Game I	$1 - \zeta > 1 - \omega > \eta > \xi$
Game II	$1 - \zeta > 1 - \omega > \eta > \xi$
Game III	$1 - \zeta > 1 - \omega > \eta > \xi$
Game IV	$1 - \zeta \geq 1 - \omega > \eta > \xi$
Game V	$1 - \zeta > 1 - \omega > \eta > \xi$
Game XI	$1 - \zeta > 1 - \omega > \eta > \xi$
Game XII	$1 - \zeta > 1 - \omega > \eta > \xi$

Here in all cases without exception the tendency to retaliate and to persist in defection is stronger than the tendency to respond cooperatively and to persist in cooperation.

We turn to the corresponding results on the state-conditioned propensities x, y, z, and w (cf. Table 9).

With only a few exceptions, the following inequality is satisfied:

$$x > y > z > w. \tag{29}$$

When we compare x, y, $1 - z$, and $1 - w$ in the Pure Matrix Condition, we have the inequalities as shown in Table 16:

TABLE 16

Game I	$x > 1 - w > 1 - z > y$
Game II	$x > 1 - w > 1 - z > y$
Game III	$1 - w > x > 1 - z > y$
Game IV	$x > 1 - w > 1 - z > y$
Game V	$1 - w > 1 - z > x > y$
Game XI	$x > 1 - w > 1 - z > y$
Game XII	$1 - w > x > 1 - z > y$

Note that the propensities x, y, $1 - z$, and $1 - w$ represent the tendencies to repeat the previous response in each of the four states respectively, *CC*, *CD*, *DC*, and *DD* (taking the point of view of player 1). Let us see how these tendencies relate to the payoffs associated with the four states. These are (to player 1) R, S, T, and P respectively. Taking into account the inequality $T > R > P > S$, we see that according to certain assumptions made in stochastic learning theory,[16] we could expect the corresponding inequality

$$1 - z > x > 1 - w > y. \tag{30}$$

Inequality (30) implies six paired comparison inequalities, namely (a) $1 - z > x$; (b) $1 - z > 1 - w$; (c) $1 - z > y$; (d) $x > 1 - w$; (e) $x > y$; (f) $1 - w > y$. Now violation of (a) indicates a greater propensity to cooperate than one would expect from the payoffs, while violation of (d) indicates a greater propensity to defect than one would expect from the payoffs. Violation of (b) is ambivalent, because both $1 - z$ and $1 - w$ are propensities to defect. The interpretation of the remaining violations is unnecessary, since they have never been observed. In Table 17 we observe the inequalities violated in the Pure Matrix Condition.

The picture is clear. The "cooperative bias" vio-

TABLE 17

Game	Violation
Game I	a, b
Game II	a, b
Game III	a, b, d
Game IV	a, b
Game V	b, d
Game XI	a, b
Game XII	a, b, d

lation of (a) is present in all except the most uncooperative Game V. The "defecting bias" violation of (d) is present only in the three least cooperative Games XII, III, and V. The ambivalent violation of (b) is present in all games. It indicates that fear of receiving S rather than hope of receiving T is the more important factor in persisting D responses. In short, the mild games are mild because the players tend to respond cooperatively to the other's initiation of cooperation (i.e., they do not persist in unilateral defection). The severe games are severe because the players tend to persist in the DD state—the trap set by Prisoner's Dilemma.

The same analysis of the Pure No Matrix Condition yields the inequalities shown in Table 18.

TABLE 18

Game I	$1 - w > 1 - z > x > y$ (violation of b, d)
Game II	$1 - z > 1 - w > x > y$ (violation of d)
Game III	$1 - w > 1 - z > x > y$ (violation of b, d)
Game IV	$1 - w > x > 1 - z > y$ (violation of a, b, d)
Game V	$1 - w > 1 - z > x > y$ (violation of b, d)
Game XI	$1 - w > 1 - z > x > y$ (violation of b, d)
Game XII	$1 - w > 1 - z > x > y$ (violation of b, d)

Here the picture is even simpler. The reversal $1 - w > 1 - z$ (violation of b) is present in all games (except II). The defecting bias violation of (d) is also

present in all games. Only in Game IV (which shows the most cooperation in this condition) do we see the cooperative bias violation of (a). In short, without the matrix displayed, the tendency to persist in the defecting response is greater than would be expected on the basis of a stochastic theory which predicts a rank order of response propensities corresponding to the rank order of the associated reinforcements.

Summary of Chapter 4

The response-conditioned cooperative propensities ξ, η, ζ, and ω and the state-conditioned propensities, x, y, z, and w reveal a more detailed picture of the pressures operating in the Prisoner's Dilemma game. The correlations among the paired indices ξ_i, η_i, ζ_i, and ω_i are lower than those among the C_i, and those among the x_i, y_i, z_i, and w_i are still lower, indicating that these indices are less subject to interaction effects. From the latter, moreover, we gain a picture of how the tendencies to persist in a given response depart from those expected on the basis of comparing the associated payoffs. Specifically, in the Pure Matrix Condition two biases are operating, namely a tendency not to persist in the rewarded defecting response (a cooperative bias) and a tendency to persist in the punished defecting response (a noncooperative bias). In the mildest games (where most cooperation is observed) only the cooperative bias seems to be operating in the Pure Matrix Condition. In the most severe game (where least cooperation is observed) only the noncooperative bias seems to be operating. In the intermediate games, both biases are observed.

In the Pure No Matrix Condition, the noncooperative bias is observed in all seven games, and the cooperative bias only in the mildest game.