Prospects for experimental games[1]

ANATOL RAPOPORT
Mental Health Research Institute, University of Michigan

Two-person non-constant-sum games offer an opportunity to study mixed-motive conflict in the context of controlled laboratory experiment. As is typically the case, the advantages of tractability and control are offset by the artificiality of the laboratory environment. Extrapolation of laboratory results to conclusions about real life are always hazardous and perhaps especially so in the context of conflict. It is not only the lack of realism of the situation depicted in the laboratory which makes the extrapolation unwarranted but also the limited range of the results obtained. Let us suppose the best, namely that to every experimental game there corresponds some real life situation to which the conclusions obtained from the experiments apply. Even in the simplest case of the 2×2 game (two players with two strategies each), there are at least eight independent variables to vary (the eight payoffs associated with each game matrix). Giving each payoff just three distinct values (low, medium, and high), we must perform 6,561 experiments to get a body of data for a comprehensive description of how the choices are influenced by the payoffs. Each experiment, be it noted, must involve some tens of subjects to give any confidence in the results.

To be sure, some patterns may emerge from only a fraction of this corpus, but the relation of choices to the absolute payoffs may not be the only nor even the main relation of interest. The relation between choices and the *relative* magnitudes of payoffs (comparing both the payoffs in the different outcomes to the same subject and the payoffs in the different outcomes to the two subjects) may be of greater interest, not to speak of other possibly very important independent variables, such as the personal characteristics of the subjects and the effects of learning in iterated plays of a game.

The foregoing was meant to serve as a warning against placing much faith in "barefoot empiricism"; that is, a program aimed at acquiring some integrated knowledge about situations of this sort by examining large volumes of data, even if systematically collected. Clearly some theoretical guidance is needed to make sense of the data. It goes without saying that experimentalists understand this guiding role of theory and they give evidence of this awareness in their experimental design.

Examining the papers on experimental games published in this journal since the gaming section was established, we get an impression of the sort of theoretical orientation which suggested the design of the experiments. The orientation, as could be expected, was strongly colored by the central theme of this journal, namely conflict

---

[1] This review was completed before the author went on leave abroad, so that—with regard to gaming articles in this journal—it covers only those through Volume XI (1967).

resolution. A substantial majority of the experiments used the paradigm of the Prisoner's Dilemma game or games closely related to it. The reason is not far to seek. This game has been most widely publicized as a paradigm of a conflict situation in which individual interest dictates one choice while collective interest dictates another. Moreover, if both players are guided by collective interest, each of them does better in the outcome of the game than if both are guided by individual interest.

Prisoner's Dilemma, then, is the mixed-motive game *par excellence* in which a player faces the choice of "competing" or "cooperating" and where the decision to cooperate can be rationalized only on the basis of assuming some sort of mutual trust (in the absence of opportunities to make binding agreements). Frequencies of cooperative choices suggest a quantitative measure of "the amount of cooperation." Measures of this sort, which seem to capture an objective index of some previously only intuitively understood personal or behavioral characteristic, are always welcomed in experimental psychology.

This measure (the probability of choosing C, the cooperative strategy in Prisoner's Dilemma) overshadows all other dependent variables in the experiments reported. Practically all of the experiments published in this journal, then, seem to have been directed to answer questions of this sort: What has an effect on the frequency of the C choice and to what extent? That is, what factors tend to facilitate or to inhibit cooperation? As has been said, this is a natural question to ask in the context of research on conflict resolution, and, of course, it ought to be asked. Unfortunately, the intense preoccupation with this question has obscured other equally important potentialities of gaming experiments.

The question "What factors inhibit or facilitate cooperation?" has a direct appeal in the sense that the relevance of the question to real life situations involving conflict resolution is immediately apparent. Further, the circumstance that most of the investigators doing gaming experiments are psychologists is responsible for the salience of variables related to the "psychological set" and to the personality of the player as independent variables. Another favorite independent variable has been the (programmed) strategy of one of the subjects (who is then, naturally, a confederate of the experimenter). In some of the experiments, the experimental situation itself is manipulated. All of these are examples of what I shall call *convergent* designs. In these designs one dependent variable (in this case "the amount of cooperation") is held in a central position while the effects on it of various manipulated or naturally observed independent variables are studied.

An alternative to a convergent design is a *divergent* design. Here a set of conditions is held constant while a great many dependent variables are examined. Only two or three papers published in the gaming section so far have been based on the principle of the divergent design. Such a design reflects a "natural history" approach to a given experimental situation rather than a hypothesis-testing approach. Let us see what a divergent design actually involves.

The data obtained from a long iterated sequence of choices in an experimental game constitute a *protocol.* The protocol exhibits a large body of statistics in which the frequency of a particular response is only one example. From a sufficiently large protocol one can obtain a large variety of such statistics; for example, *conditional* frequencies of responses (i.e., frequencies following any of the four possible outcomes of a 2×2 game), distributions of the lengths of runs of any of the responses or of the joint responses (i.e., the outcomes), the interdependencies of the responses of the two players, the systematic shifts of all of these in iterated sequences (i.e., the learning effect of interactions between the players), etc.

Naturally, all of these statistics are not independent of each other. Indeed, if the iterated game could be described as a stochastic process, *all* of the statistics would be theoretically deducible from the mathematical model of the process. Clearly, the "divergent design" approach, i.e., the design of an experiment so as to offer the possibility of estimating several statistics of a population of protocols, lends itself best for a test of a *mathematical* theory, in which the various statistics would appear as derived consequences of a mathematical model. What would then be the connection between the mathematical theory and a psychological theory of the process studied? These connections would be suggested by interpreting the *parameters* of the mathematical model.

To take an example, suppose a finite state Markov chain were chosen to represent the stochastic process embodied in the iterated game. Then a subject would be assumed to be in one of a specified number of states. It is natural to choose the outcomes as the states (although in principle sequences of outcomes could also be taken). Next, the subject is assumed to make the one or the other response with certain probabilities which depend on the state in which he finds himself. Combining the choice probabilities of the two subjects, we obtain the transition matrix of the Markov chain. All the statistics of the process are now deducible. If the estimate of the statistics obtained from the protocols corroborate the model, we have the beginnings of a theory. That is because the parameters of the process, namely, the conditional choice probabilities, can be interpreted psychologically. They are the propensities of the subject to do one thing or another *in the several situations* in which he finds himself. The corroboration of the model involves the assignment of values to these parameters. *Now* the interesting problem arises of how these *parameters* change as independent variables are manipulated.

In principle, this approach is no more than a refinement of the single dependent variable studies. For the parameters, e.g., the conditional probabilities, are essentially breakdowns of a single variable (the frequency of a response) into component parts. However, the values of parameters of a mathematical model can be expected to reflect more faithfully whatever governs the process than the values of grossly observed variables, if only because the parameters are the relatively constant characteristics of the process, while the variables (such as response frequencies) are subjected to the dynamics of interaction. Thus, to take the Markov chain as an example, even though the transition probabilities are assumed to be constant in the iterated game, the response frequencies will be expected to change, approaching asymptotic values.

There are serious limitations on the mathematical model approach. Ordinarily a model of two interacting players will involve a very large number of independent parameters. The estimation of such parameters is a very laborious process. Nor does the labor end with the estimation. If a behavioral theory is to be built on the basis of a mathematical model of an iterated game, experiments must be repeated with different sets of independent (manipulable) variables in order to ascertain how the changes are reflected in the values of the parameters (always assuming that the basic mathematical model remains valid). In this way, even with the guidance of a mathematical model, the magnitude of a systematic experimental program quickly becomes unmanageable.

For this reason, one might conclude that an attempt to construct a serious mathematical theory of even the 2×2 game would be premature. Thus we are forced to fall back on the "natural history approach," i.e., designs where the independent variables are limited to a few basic ones (and consequently temptations to introduce "fancy" variations in the experimental conditions are resisted) but where the data corpuses are large and subjected to intensive study.

The following independent variables of a 2×2 game seem to me to be basic.

1. *The payoff matrix.* The psychologist is understandably interested not in the payoff matrix *per se* but in what the payoffs "mean" to the subjects; or, in game-theoretic terms, in the utilities of the payoffs. The problem of ascertaining utilities is a separate and fundamentally different problem. Little would be gained in pursuing this problem, or in raising questions of whether the payoffs can have any meaningful analogues in real life situations. If we decide to study the 2×2 game intensively for its own sake (the value of such study will be discussed below), then the payoffs should be taken simply as the givens of our study and results should be related to them. It is advisable, however, to standardize the range of the payoffs in order to make the results of various studies comparable.

Judging by results obtained so far, there is little question that payoffs as monetary rewards do act as incentives at least for the subject populations studied. The range of the payoffs could be roughly standardized if the expected gains or losses were made of the same order of magnitude as the fees the subjects receive as participants in the experiments. Fees are practically mandatory if some payoffs are negative, otherwise the subjects can suffer real losses, which can create problems. On the other hand, it is not advisable to make all payoffs positive: some outcomes should be genuinely "punishing." Fees and payoffs can be so calibrated that at worst the subject may leave empty-handed and on the average will come out with a modest supplement to the subject fee.

If the eight payoffs of a 2×2 game are allowed to vary independently and without restriction, different *types* of games will result. If the four payoffs accruing to each of the players are all distinct, and if we read them only on an ordinal scale, the number of types is finite. In fact, there are exactly 78 strategically nonequivalent 2×2 games. Analysis of the 78 matrices suggests a classification of these 78 games into 12–15 categories depending on the existence of one, two, or no equilibria, and on the "vulnerability" of these equilibria, i.e., on the pressures resulting from departures from the equilibria by one or the other of the players.

It is interesting to note that Prisoner's Dilemma is the only game in its category: a game with a single so-called strongly stable but Pareto-nonoptimal equilibrium. Perhaps this is one of the reasons why this game has received so much attention: its uniqueness was somehow intuited.

Like Prisoner's Dilemma, each of the categories brings out specific motivational pressures. For example, while in Prisoner's Dilemma there is a strong motivational pressure against cooperation (since from the point of view of the individual player noncooperation is more advantageous than cooperation regardless of the other's choice), in the closely related game of Chicken (a "preemption" game), the pressure is ambivalent. It is more advantageous not to cooperate if the other does, but to cooperate if the other does not. An important feature in Chicken is "preemption," that is, a firm commitment not to cooperate, which leaves the other with the choice between "submitting," as the lesser of two evils, or "resisting" at an exorbitant cost to both. Other games give players opportunities to communicate "threats." These threats are effected by shifting away from equilibria. They are calculated to induce the other player to shift his choice rather than suffer future losses due to such departures by the first player. Still other games have "force vulnerable" equilibria. A departure from an equilibrium by one of the players actually forces the other to shift (on the next play) if he wants to improve his payoff.

Some of the games are symmetric in the sense that the situations of the two players are exactly the same; others are asymmetric, permitting the study of behavior of players whose situations in the game are different.

Finally, in each of the strategically nonequivalent games, the magnitudes of the payoffs can be varied without disturbing the strategic structure of the game. Here it is not the nature of the motivations but their intensity which is the manipulated variable.

2. *Time.* Learning certainly takes place in iterated plays if the outcome of each play becomes known to the subject. A rather stable feature of the Prisoner's Dilemma game is the tendency of the players to "lock-in" either on the double-cooperative or on the double-defecting outcome. Unilaterally cooperative outcomes tend to become extinct in long iterations. The lock-in effect is most clearly observable if the game is played several hundred times.

In general, there is evidence that subjects do not really "get the hang" of the game until they have played it a hundred times or so, with an opportunity to compare the results of different "strategies," these being understood not simply as the two choices available in the 2×2 game but as different sequences of choices. However, the short-term behavior of the subjects is also interesting. In particular, it is interesting to see the distribution of responses in one-shot plays.

Experiments with one-shot (noniterated) plays present a problem of design efficiency. Since a decision in a 2×2 game takes only a short time, it is extremely inefficient to confine such an experiment to only one play, i.e., to recruit a pair of subjects, to give them instructions, etc., only to have them actually perform for a few seconds. A natural way out is to gather data from the same pair on several different games, each played once. This, however, raises another problem. If the results of each play are fed back to the subjects, they may develop some "principles of play" abstracted from the results. If these principles are applied across all the plays in the experiment, the responses may "contaminate" each other. This is particularly noticeable if all the games presented in an experimental session are of the same type. For example, if all the games are Prisoner's Dilemma with varying payoffs, the subjects may decide to cooperate or not to cooperate in *any* such game. In that case, the data will not differentiate between the games with different sets of payoffs, whereas this differentiation may have been the original aim of the experiment. Contamination of this sort can, perhaps, be prevented by (a) mixing games of widely different types in random sequence in a single experiment and (b) by giving no information about the outcome until all the games have been played. Thus, every play becomes a "first" play without the intervention of learning, and a "principle" is not easily transferable from one play to another.

If the above method is used, the subjects need not even be present at an experimental session. Each subject can be given a stack of cards, each representing a game matrix, to take home with the instruction to indicate his choice on each of the cards. It does not matter how much time the subject spends in making up his mind, nor whether he confers with others. The purpose of the experiment may well be to get a statistical distribution of the responses where the respondents are given the fullest opportunity to think about each problem. The principal feature of most non-constant-sum games is that they do not have an unambiguously rational solution in the form of noncooperative (nonnegotiable) games; therefore it does not matter how much time is spent in arriving at a decision. In fact, the more time is spent, the more the distribution of responses is likely to reflect a "cultural norm."

In short iterated runs, feedback about outcomes is essential; hence the subjects must be present in the experiment. In this case only the first stratagem (mixing games of widely disparate type) is available. The principal question of interest here is how the response pattern varies with the total length of the iteration, say from two to ten plays.

3. *Subjects.* The question of how choice behavior in 2×2 games is related to the personality of the subject was one of the first ones raised in the context of these experiments. The question is a legitimate one, of course. Still, attempts to answer it must be tempered by a number of caveats. In iterated plays, interaction effects between the two players play a major, sometimes a decisive role. For instance, in long iterations of Prisoner's Dilemma, it has been shown that the Pearson correlation coefficient of the frequency of choosing the C response (across a population of pairs of subjects) is in the high nineties. This high correlation is, no doubt, a result of the lock-in effect frequently observed in this game. If lock-ins occur, this means that the unilateral responses (C_1D_2) and (C_2D_1) are rare. In fact, they tend to be gradually eliminated in the course of the iteration. If there were no unilateral responses, the correlation would be 1.00. Because of these high correlations, it makes no sense to look for correlations between personality traits of subjects and their performance within pairs. If the performance of one subject is an excellent predictor of the performance of the other, no variance is left to be accounted for by the personalities. It does make sense to compare performances of independent populations of subjects recruited from different pools. Here, again, however, some approaches are more promising than others. Although some interesting correlations have been obtained between performances and personality profiles (notably in Prisoner's Dilemma, where the propensity to cooperate has been shown to correlate negatively with measures of rigidity or "authoritarianism"), still it would seem that attention would be turned with more profit to obvious differences among populations (rather than differences derived from paper and pencil personality tests), such as sex, age, occupation, and above all, cultural background. The spread of these experimental techniques to several countries (as is already happening) may yield highly interesting results; but this will come about only if a few standard experimental designs are adopted by investigators everywhere in order to make comparisons meaningful.

4. *Strategy of Other.* The realization of

the important role played by interaction in iterated plays raises the important question of how the pattern of responses of one subject depends on that of the other. The question can be answered most directly if the responses of one subject are subjected to manipulation and control. This is accomplished by the use of a "stooge," i.e., a programmed pseudo-subject. The data so far obtained are not always consistent. Some experimenters have reported distinct dependencies of the responses of the real subject on the pattern of choices assigned to the stooge; others have failed to find such dependencies. The inconsistencies can be attributed to a lack so far of massive, persistent, and systematically standardized investigations. The range of *patterns* of choices in an iterated game is enormous, and it is not surprising that little can be seen by comparing the effects on the subject of a few more or less haphazardly selected programmed strategies. This is particularly true if the programmed strategies are "noncontingent," for example, if the stooge is assigned a random sequence of C and D choices, in which only the relative frequency of C is the controlled variable. It may well be that within very wide limits the relative frequency of C has no appreciable effect, if the occurrence of the C responses is random. The subject cannot discern a pattern in the responses of his partner-opponent and so is not guided by it. On the other hand, it has been reported rather consistently that simple contingent strategies, particularly a tit-for-tat strategy on the part of the stooge, tend to push the subject toward cooperation. This suggests that variations in contingent strategies may be of greater importance than variations in noncontingent strategies.

An important exception is the use of the 100 percent cooperative strategy on the part of the stooge in the game of Prisoner's Dilemma. The distribution of C responses among the real subjects tends to be bimodal in this case, as if the subjects differentiated into two classes, those who exploit the cooperation of the other and those who respond to it by cooperation.

So far, practically the only results reported involving the use of a stooge have been on the Prisoner's Dilemma game. It is likely that the extension of this method to other 2×2 games will yield much richer results because of the greater variety of motivating factors built into the structure of those games. To take an example, compare Prisoner's Dilemma with the game of Chicken in the case where the stooge consistently uses the D (noncooperative) strategy. In Prisoner's Dilemma, there is only one rational response to such persistent use of D, namely D. If it is not likely that the other will cooperate, there is nothing to do but reciprocate with noncooperation if large losses are to be avoided. This is not so in the game of Chicken. The short-term rational response to D in that game is C, to avoid the even larger loss in (DD). However, this sort of response produces strong counter-pressures, because it is tantamount to submission. Thus a shift to D can be rationalized (if the punishment for double-defection is not too strong) as a means of "teaching him a lesson" even at the expense of a loss to oneself.

Still other games, in which departures from the "natural outcomes" can be interpreted as threats or the acceptance of temporary losses in order to induce the other to shift his strategy, offer additional opportunities for studying the possibly complex effects of different programmed strategies.

In summary, these four classes of manipulable variables seem to be important in the sense of a promise of interesting and theoretically interpretable results: (1) the payoff structure of the game matrix; (2) the number of plays; (3) the subject population; (4) a controlled (programmed) strategy of one subject.

As has been said at the outset, while it is desirable to keep the independent variables few and standardized, attention ought to be directed at a great variety of dependent variables, instead of confining it to the most obvious in the case of Prisoner's Dilemma, namely the frequency of cooperative responses. Since the distinction between cooperation and noncooperation is not nearly as clear in some other 2×2 games, the extension of the experiments to these games will also turn the attention of the investigators to other dependent variables.

Even in Prisoner's Dilemma, however, there is a great variety of dependent variables which should be studied in detail. Foremost among these are the conditional frequencies of response. In a more general context, these can be expressed as the "shift propensities" in each of the four outcomes of the game. The shift propensity is defined as the probability of shifting one's strategy given the outcome of a play of the game. In terms of the conditional propensities to cooperate in Prisoner's Dilemma, these can be defined in the following way:

$\tilde{x}$: the propensity to shift when the outcome is (CC), which is the complement of x, the propensity to cooperate when that outcome obtains.

$\tilde{y}$: the propensity of player 1 to shift when the outcome is (CD), which is the complement of y, the propensity to cooperate when that outcome obtains.

z: the propensity of player 1 to shift when the outcome is (DC), which is identical with the propensity to cooperate when that outcome obtains.

w: the propensity to shift when the outcome is (DD), which is identical with the propensity to cooperate when that outcome obtains.

The complements of the shift propensities, namely x, y, $\tilde{z}$, and $\tilde{w}$ respectively, are, of course, measures of the propensity to *persist* in the corresponding outcomes. We can now ask questions about what these propensities (to shift or to persist) depend on. The most simple-minded hypothesis would relate these propensities to the payoffs to the player in question associated with the outcomes. However, in Prisoner's Dilemma it has been definitely shown that the two variables are not directly related. For example, the propensity to persist in (DD) is as a rule greater than the propensity to persist in (DC), in spite of the fact that the payoff to the first player is larger in the latter outcome.

One is then led to the next simplest hypothesis, namely that the shift propensities are related to the change in payoff which can be effected by shifting. Here the actual magnitudes of the payoffs, not just the order of the magnitudes, is the decisive factor. As far as I know, this hypothesis has not yet been put to a test.

It should be noted that the search for correlations between the conditional probabilities of response and the payoff structure should be more rewarding than a search for correlations between the unconditional probabilities and the payoff structure, if only because the former are more stable during the course of the iterated game than the latter, which are subject to strong interaction effects. In general, it makes the more sense to seek correlations between indices of behavior and independent variables, the more stable these indices are. The construction of stochastic models of response protocols should be directed toward the discovery of such stable indices.

Another dependent variable of interest is the distribution of the lengths of runs of various outcomes. A cumulative distribution (the fraction of runs having at least a given length) which is exponential would corroborate a null hypothesis, namely that shifts occur entirely at random. An examination of these distributions in iterated Prisoner's Dilemma shows systematic departures from the exponential distribution. These departures are strongly suggestive. That is, if the distribution is "stronger" than the exponential (i.e., if the exponent turns out to be a convex downward function of time), this indicates that the longer a run lasts, the more likely is a shift away from it. On the contrary, if the distribution is weaker than the exponential (i.e., if the exponent turns out to be a convex upward function of time), a lock-in effect is evident: the more an outcome persists, the more it is likely to persist further. In estimating these distributions it is not permissible to lump data from several pairs of subjects, because this may confound the genuine effect with the effect of "natural selection." For example, the distributions of both the (CC) runs and of the (DD) runs in massed data on Prisoner's Dilemma show strong persistence effects. But these are partially accounted for by the fact that, as the runs become longer, the pairs which do not persist in those runs are eliminated from the sample, leaving the persistent pairs. This error can be circumvented by taking averages of individual distributions instead of distributions observed in massed data.

In conclusion, there is much to be learned about the dynamics of interaction in the playing of iterated 2×2 games. It seems desirable to obtain this knowledge by an intensive study of such games in their own right, postponing for the time being the question of how relevant this knowledge may be to an understanding of mixed-motive conflicts in real life. To raise this question in connection with every experiment may lead to unwarranted conclusions (if one makes rash extrapolations), or else to a premature discouragement concerning the value of the experimental method (if one constantly keeps in mind the dangers of extrapolation). What is worse, viewing the laboratory method in terms of simulation of real life conflicts leads to designs which are not guided by the inner logic of a systematic investigation. A laboratory game cannot be considered as a simulation of real life in the sense that a wind tunnel is a simulation of a real aerodynamic situation. In the latter, the same physical laws are operating as those of the simulated situation; in the former this can never be assumed. The value of the laboratory experiment in the simple context of the 2×2 game is in the opportunity it gives for building a systematic theory *of that situation.* What relation that theory will have to real life conflicts (if any) only time can tell—namely, when the concepts emerging from the descriptive theory of the 2×2 game have become stabilized as concepts of proven theoretical power. Such has been the history of every science which began life in the laboratory. The laboratory phase ought to be considered as the incubation period of a science. The conditions in the incubator are generally designed not to simulate realistically the conditions of the environment in which the adult organism will live, but rather to give the incubating organism the best chance to come to life.

Having indicated my personal bias in choosing research strategies for gaming experiments, I would like to dispel any impression that "realistic" simulations of conflict in the laboratory should be discouraged. The enlarged dilemma games (i.e., Pilisuk's disarmament games), "realistic" simulations of the game of Chicken and of nuclear deterrence (e.g., the games used by M. Deutsch and his associates), games which simulate "pacifist resistance" (e.g., those of Shure and Meeker) are all valuable vehicles for *demonstrating* situations of this sort. They also offer an opportunity for generating hypotheses concerning the effects on behavior of the psychological sets induced by the structure of the situations. The results of these experiments are often richly suggestive. The problem of how to pass from these suggestions to a defensible theory of conflict remains a difficult one, because the "realism" of the simulation precludes a reduction of the gaming experiment to a standardized, tractable format. As a consequence, results obtained in different simulations (which sometimes contradict each other) cannot be meaningfully compared. For this reason, a substantial portion of experimental gaming effort ought to be devoted to the simplest formats, which alone make possible a systematic build-up of *a* theory. The relevance of this theory to a general theory of conflict is a question which can be posed only after the theory has been constructed.