UTILITY COMPARISON AND THE THEORY OF GAMES

L. S. Shapley

April 1967

P-3582

UTILITY COMPARISON AND THE THEORY OF GAMES

Lloyd S. Shapley*

The RAND Corporation, Santa Monica, California

1. Interpersonal comparability of utility is generally regarded as an unsound basis on which to erect theories of multipersonal behavior. Nevertheless, it enters naturally--and, I believe, properly--as a nonbasic, derivative concept playing an important if sometimes hidden role in the theories of bargaining, group decisionmaking, and social welfare. The formal and conceptual framework of game theory is well adapted for a broad and unified approach to this group of theories, though it tends to slight the psychological aspects of group interaction in favor of the structural aspects--e.g., complementary physical resources, the channels of information and control, the threats and other strategic options open to the participants, etc. In this note I shall discuss two related topics in which game theory becomes creatively involved with questions of interpersonal utility comparison.

The first topic concerns the nature of the utility functions that are admissible in a bargaining theory that satisfies certain minimal requirements. I shall show, by a simple argument, that while cardinal utilities are admissible, purely ordinal utilities are not. Some intriguing intermediate systems are not excluded. The argument does not depend on the injection of probabilities or uncertainty into the theory.

The second topic concerns a method of solving general n-person games by making use of the interpersonal comparisons of utility that are implicit in the solution. After two complementary modes of comparison have been distinguished, a "principle of equivalence" points the way to an attractively direct extension of the definition of the value of a game, from the "transferable" case to the "nontransferable" case.

2. A number of qualitatively different definitions of "solution" exist in the literature of n-person game theory. An explanation for this multiplicity may be found in the essential ambiguity of decision-making in the presence of several independent "free wills." Simple

---

This paper was prepared for advance distribution to participants in an international colloquium: "La Décision: Agrégation et Dynamique des Ordres de Préférence", organized by the (French) Centre National de la Recherche Scientifique, to be held at Aix-en-Provence, July 3-7, 1967.

rationality (utility maximization) is not a sufficient determinant of behavior when one ventures beyond simple cases like the one-person decision problem or the two-person game with directly opposed interests. Determinateness (i.e., uniqueness of outcome) may be desirable in a solution, but it can generally be obtained only at the price of oversimplifying or ignoring the observed tendencies toward organized cooperation (e.g., markets, political parties, cartels, etc.) when real people cope with the indeterminacy of real-life multilateral competition. Such social, political, or economic institutions, often highly abstracted, are at the heart of many of the best-known solution concepts.

In this note, however, we shall be dealing only with the simplest kind of solution concept, which seeks to reduce each game to a single vector of payoffs, known as the value of the game. From what has just been said, it should be clear that a "value" concept is not the only possible capstone to a well-built theory of games. Hence conclusions (such as those in this paper) that are based on the assumed existence of a self-consistent method of game valuation can be rejected without necessarily demolishing the entire theory. From some plausible standpoints, in fact, it can be demonstrated that a valuation theory free from contradiction is unattainable.

But there are other standpoints from which it can be argued that a valuation theory is virtually indispensible. The most compelling argument, perhaps, is one that takes us back to the basics of utility theory. Let an individual be confronted with the prospect of entering some sort of multilateral game situation, e.g., a partnership, an oligopolistic industry, a political office. What is the utility to him of that prospect?

3. For our first topic, we consider a two-sided negotiation. The "rules of the game" are simple: if the parties agree, they can have any outcome in an "agreement set" A, but if they fail to agree, they must take the "disagreement point" D. A and D are construed in the cartesian product of the two bargainers' utility spaces, which we take to be real numbers scales, and A is assumed to be a continuous, strictly monotonic curve with endpoints aligned with D, as shown in Figure 1.

Let us postulate a valuation theory for cooperative games that is powerful enough to resolve at least this elementary kind of bargaining game (hopefully more), and that gives some interior point of A, say V, as the solution. Let us further postulate:

> (I) The solution depends only on the configuration (A, D) in the joint utility space, and not on any non-utilitarian attributes (e.g., physical symmetry, numerical quantities) of the actual subject of negotiation.

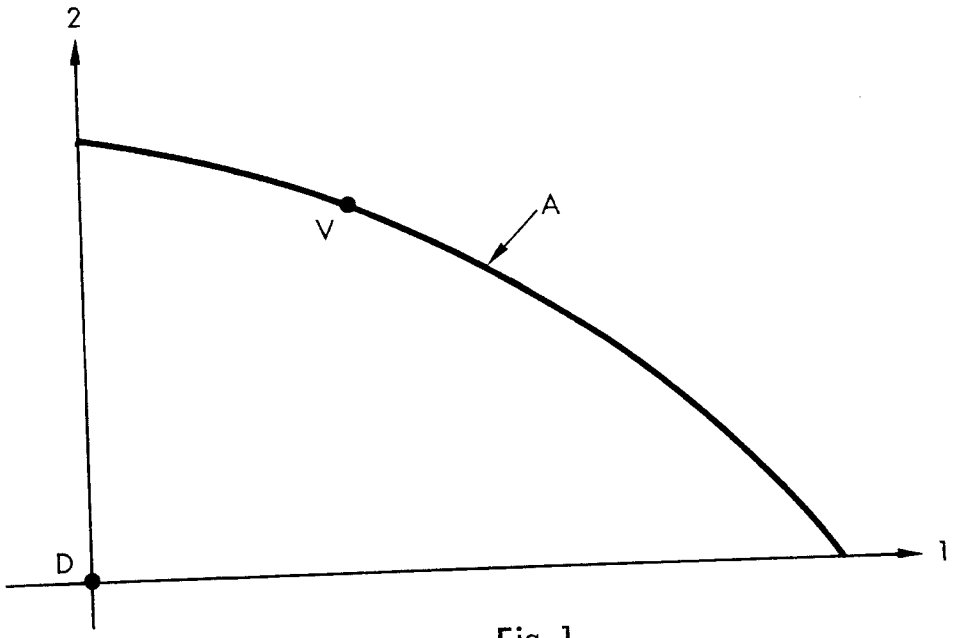

Fig.1

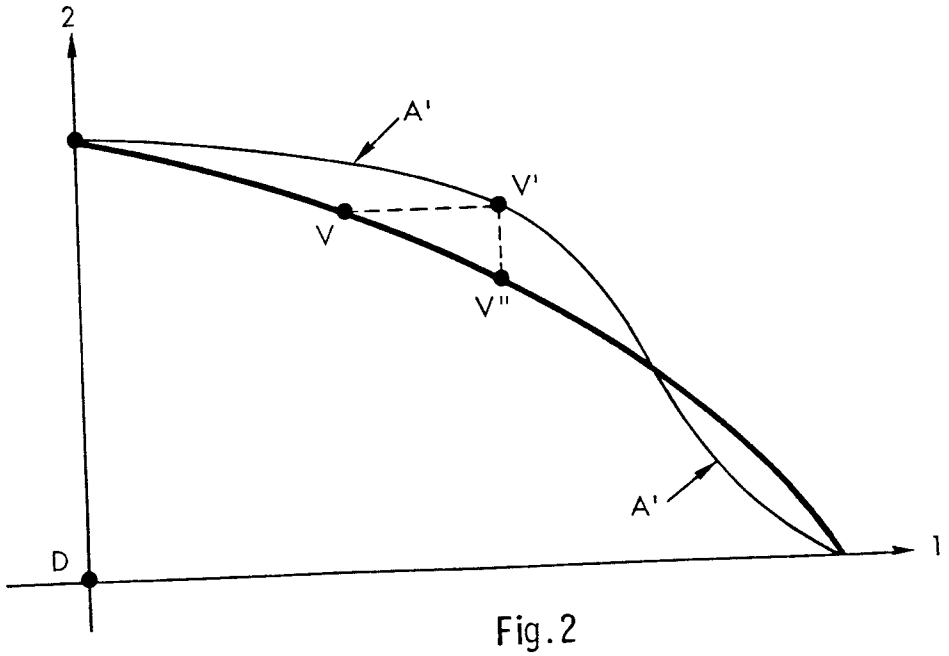

Fig.2

(II) The solution is covariant (i.e., the physical outcome is invariant) under some group G of order-preserving transformations, applied separately to the two utility scales.

What do these postulates imply about the nature of the group G?

First, let us test the ordinalist assumption. Let G be the group of all continuous order-preserving transformations on the real line R. We soon run into trouble. Let us apply to the first player's utility scale any continuous order-preserving transformation that leaves the endpoints of A fixed, but moves V. A curve like A' might result (Figure 2), the points of A having been displaced horizontally. But now we can immediately construct a transformation of the second player's scale that restores the curve A, by displacing the points of A' vertically. Since the point D has not moved, our first postulate plainly requires that V be the solution of the twice-transformed problem. But since the outcome originally at V is now at V", our second postulate just as plainly requires that V" be the solution. Since $V \neq V''$, we are forced to conclude that purely ordinal utilities are inadmissible.

4. This kind of contradiction is always present so long as there is any element of G that has two fixed points with a nonfixed point in between. Indeed, if g is such an element, we can arrange (A, D) so that g^{-1} applied to the second scale restores the displacement of A caused by g applied to the first scale, and in such a way that every interior point of A is shifted in the process. Conversely, if G contains no such element, then postulates (I) and (II) can never be brought into conflict. Let us call such a group--characterized by the property that the set of fixed points of each element is convex--an unwavering group.

Of course, a complete valuation theory will have other postulates, which may further restrict the admissible utilities. Nevertheless, it is of some interest to explore the class of unwavering groups, to discover just how far (I) and (II) force us to go in our retreat (advance?) from purely ordinal utility. The group of all positive linear transformations, which characterizes the usual "cardinal" or "linear" utility, is certainly unwavering, as are its subgroups. But there are various other possibilities, qualitatively different; the least strange among them is the group that leaves some point, say 0, fixed (the status quo?) while operating separately in a linear fashion on the positive and negative half lines. Our most interesting result, in the context of the present discussion, is the following:

THEOREM: If G is an unwavering group of order-preserving transformations of the real line R, and if G is <u>transitive</u>,* then there exists a continuous order-preserving recoordinatization of R such that, in the new coordinate system, every element of G is a positive linear transformation.

Our main conclusion is that if we want a value theory for bargaining games that is based on utility considerations alone, then we cannot use ordinal utilities, but are driven to cardinal utilities or some even more stringent system. This can perhaps be made more palatable to the intuition if we reflect that bargaining, by its very nature, tests the <u>intensities</u> of the desires of the contending parties. In other words, <u>utility differences</u> become comparable, between persons. By repetition, utility differences may also become comparable between different parts of the same person's scale of values. Thus, speaking intuitively, experience in bargaining forces an individual to "straighten out" his value system--makes him decide not only what he wants but how badly he wants it. A nonlinear transformation of such a man's utility scale would represent a real change in his bargaining attitudes.

---

*Transitivity is equivalent to the assertion that the orbit of any point of R is R itself. The theorem remains valid if this is weakened to the assertion that the orbit of each point of R is dense in R.

5. We now turn to our second topic. The interpersonal utility comparisons that figure in negotiatory processes can be divided into two classes. The distinction is one of relative direction. At times, a person may compare his projected gain against another's loss, or his loss against another's gain. Thus: "Do me a favor! It would only be a little bother for you, and would help me a lot." The comparison is implicit in the words "a little ... a lot"; and the point of the comparison is that society as a whole would be better off if the request were granted.

At other times, a person may compare his gain against another's gain, or loss against loss. "This is going to hurt me more than it hurts you!", the classic slogan of parental discipline, has its counterpart in the language of negotiation. The criterion in this form of comparison is not total welfare, as above, but "fair division", or equity. "My demand is more reasonable than yours," the bargainer may plead, "therefore you should give in!" It may be observed that the utility comparison implicit in "more reasonable" is not of an absolute or universal nature, but is relative to the realities of the particular bargaining positions, i.e., to the competitive rules of the game. In the previous type of comparison, only the cooperative rules, which determine feasibility, were relevant.

Given a game and its outcome, we can make separate estimates regarding the two types of interpersonal comparison we have described. On the one hand, we can measure the given outcome against other possible outcomes--for example, nearby points on the Pareto surface, and base a utility comparison on the exchanges that could have occurred, but did not. On the other hand, we can measure the given outcome against the initial prospects and opportunities of the players, and base a comparison on the presumption that each player "got what he deserved"--i.e., that the outcome was in some sense <u>equitable</u>. (In contrast, we might say that the first comparison was based on the presumption that the given outcome was <u>efficient</u>.) If the comparisons are expressed as sets of weights, or scaling factors, to apply to the players' (cardinal) utility scales, then the first set of weights becomes a guide to the <u>maximization of social welfare</u>, the second to the <u>sharing of social profit</u>.

6. Having thus pedantically, if informally, distinguished between two modes of interpersonal comparison, we now proceed to declare their <u>equivalence</u>, as a central prerequisite for a value theory.

> (III) An outcome is acceptable as a "value of the game" only if there exist scaling factors for the individual (cardinal) utilities under which the outcome is <u>both</u> equitable <u>and</u> efficient.

As it stands, (III) is a guiding principle rather than an exact postulate, since there remain several questions of interpretation,

particularly regarding the meaning of "equitable". Before continuing, however, it may be helpful to show how this principle works in a special case: the familiar two-person pure bargaining game.

In Figure 3, the point X implies "efficiency" weights in the ratio 1:3, since the slope of the tangent at X is -1/3. (That is, in order for X to maximize the sum of utilities, the second player's nominal, numerical payoffs must be tripled.) At the same time, X implies "equity" weights in the ratio 2:1, since the line joining D to X has slope 2. (That is, in order for the players to be sharing equally at X, the first player's payoffs must be doubled.) Since the two sets of comparison weights are not proportional, X cannot be the value of the game, according to (III).

It is well known that a pure bargaining situation of this kind has a unique solution satisfying (III), namely the "Zeuthen-Nash" outcome that maximizes the product of the utility gains (the point V in the figure). It can be given a much more solid derivation than we have given here; the only virtue we claim for our present method is that it can be generalized, as we shall now proceed to demonstrate, to the more complex models of general game theory, where strategies and coalitions play essential roles.

7. We must first become more explicit about the notion of "equitable" (with respect to a given set of comparison weights), when used in a multiperson, strategically rich context. The simple idea of "sharing equally" no longer works; it may not even be well-defined.

Our procedure, in the present note, will be to assume that we have a notion of equitable value for games where utility is transferable, at stated rates of exchange, so as to concentrate on the extension to the more general, nontransferable case. Our nominee for this "transfer value" is the value for games with side payments, discovered by Harsanyi* and axiomatized by Selten,** but our method of extension will be independent of this particular choice. We shall require merely that the transfer value be unique, Pareto optimal, individually rational, and continuous as a function of the game's payoffs.

Consider now an n-person game in which there is no vehicle for the direct, unrestricted transfer of utility, but in which utility is nevertheless assumed to be interpersonally comparable according to some definite system of weighting factors. For convenience, assume that the weights are all equal. In this case, the efficient outcomes are simply those that maximize the sum of utilities. But which are the equitable outcomes?

---

*Annals of Math Study 40 (1959), pp 325-355.

**Ann. Math. Study 52 (1964), pp 577-626.

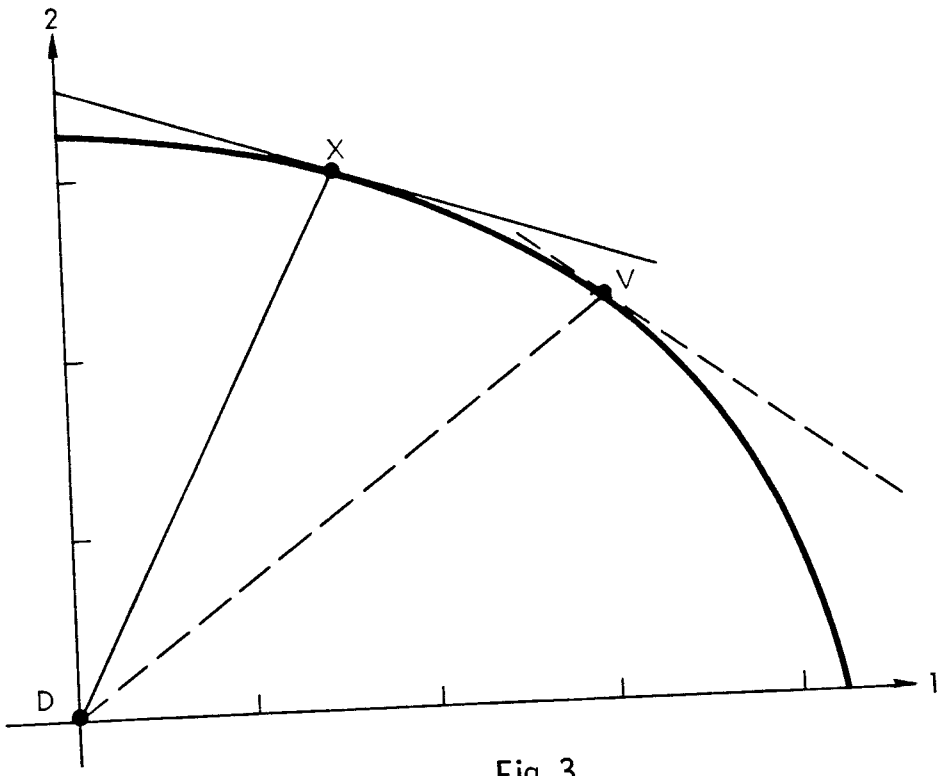
2
X
V
D
1

Fig.3

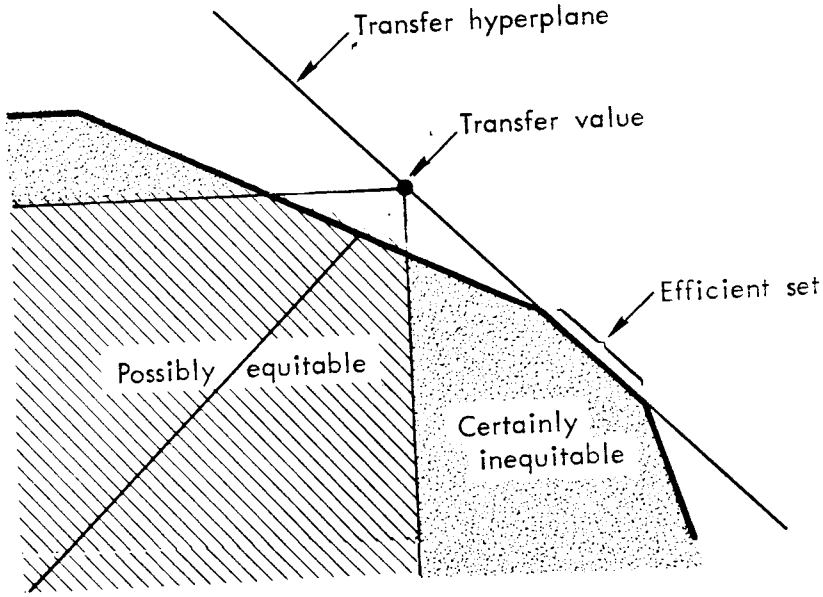
Transfer hyperplane
Transfer value
Efficient set
Possibly equitable
Certainly inequitable

Fig.4

A first candidate would be the transfer value, which by assumption is equitable if utility side payments are admitted. In general, however, the transfer value will not be a feasible outcome in the nontransferable game. To be sure, one might propose as equitable all outcomes of the form

$$(\varphi_1 - x, \varphi_2 - x, \ldots, \varphi_n - x),$$

where φ is the transfer value and x is large enough to make the result feasible (or subfeasible). (This rule would pick out the outcomes along the line DX in Figure 3, if the first player's payoff is doubled.) However, "equal taxation to overcome the deficit" is a dubious principle of fair division, and we do not adopt it. Instead, we shall make only the more modest claim that any outcome giving some player more than his transfer value, while giving some other player less, is certainly inequitable. This is illustrated schematically in Figure 4.

As already intimated, we shall also claim that the transfer value itself, if it is feasible, is certainly equitable. This may be regarded as an application of the "principle of irrelevant alternatives": If restricting the feasible set by eliminating side payments does not eliminate some solution point, then that point remains a solution.*

A glance at Figure 4 may now convince the reader that an outcome can be both efficient and equitable, as required by the principle of equivalence (III), if and only if it is the transfer value and the transfer value is feasible. To prove this, we merely note that the transfer value and all the efficient points lie on the hyperplane representing the maximum feasible utility sum. Hence, if the transfer value is feasible, then it is efficient as well as equitable, while if it is not feasible then the efficient points all lie in the "certainly inequitable" zone.

8. To complete the picture, we must of course allow for different sets of weights. Varying the weights will in general move the set of efficient points, since it changes the slope of the transfer hyperplane. Also, it will in general move the transfer value. (Under our assumptions the second motion will be continuous, the first semi-continuous.) We shall therefore speak of λ-efficiency and the "λ-transfer value," where $\lambda = (\lambda_1, \ldots, \lambda_n)$ is an arbitrary vector of nonnegative weights, not all zero.

---

*In thinking about these claims concerning "fair division," it is well to remember that they are being made in reference to a never-never land in which utility has been assumed to be extrinsically comparable. This assumption is a technical expedient with no standing in the final theory, which has been introduced to aid in discovering an intrinsic comparability.

In view of the preceding discussion, linking feasibility and efficiency, our goal is to choose λ in such a way that the λ-transfer value is feasible and hence λ-efficient. By the principle of equivalence (III), no other outcomes are acceptable as value solutions of the nontransferable game.

It may be noted that we can put $\Sigma\lambda_i = 1$ without loss of generality, since only ratios matter in the end. Hence there are really only n-1 "degrees of freedom" in the choice of λ. On the other hand, there are essentially n-1 conditions involved in requiring the λ-transfer value to be feasible, since the transfer hyperplane is n-1 dimensional. This naive "equation counting" suggests that there may be a unique λ such that the λ-transfer value is feasible, or at least, a 0-dimensional set of such λ. Thus encouraged, we propose (III) as a sufficient as well as necessary condition, and define a value of the nontransferable game to be λ-transfer value that is feasible. As a companion to each value obtained under this definition, there will be a set of intrinsic utility-comparison weights.

9. Our fundamental existence theorem states that every game (of a suitably wide class of games) has a value. Since our account has so far been free from complicated mathematical argumentation, we have put the details in an appendix. However, an outline of the proof may be of interest.

For each λ we consider the possible side-payment vectors that could take us from the λ-efficient set to the λ-transfer value, which we denote $\varphi(\lambda)$. The set of all such vectors, denoted $P(\lambda)$, is nonempty, convex, and compact under the hypotheses of the theorem, and it varies upper semi-continuously with λ. If $P(\lambda)$ contains the zero vector, then $\varphi(\lambda)$ is feasible and we are through. Restricting λ to the hyperplane $\Sigma\lambda_i = 1$, we define a point-to-convex set mapping $\lambda \to \lambda + P(\lambda)$. After an extension, which is needed to make this mapping go from points of a simplex to subsets of the same simplex, we apply the Kakutani fixed point theorem to show that the mapping has a fixed point. The individual rationality of the λ-transfer value is then invoked, to ensure that the fixed point belongs to the original mapping rather than the extension, and we conclude that for at least one λ, $0 \in P(\lambda)$.

10. The value definition developed here was first contrived in an attempt to approximate Harsanyi's 1963 bargaining value* by something that might prove analytically more tractable in dealing with

*J. C. Harsanyi, "A simplified bargaining model for the n-person cooperative game," International Economic Review 4 (1963), pp. 194-220.

economic models having large numbers of participants.* We then perceived that the "approximation" had virtues of its own. Despite the rather different approach we have adopted (deductive rather than constructive), the influence of Harsanyi's work remains considerable -- particularly his key idea of using intrinsically-defined utility weights.

The two values have the following common features: (1) Zero weights can occur, and must be allowed if the existence theorem is to hold in general. (2) Nonunique solutions can occur, even in games (with three or more players) that are not noticeably exceptional. But different solutions never have the same comparison weights. (3) If utility is transferable, then the solution is unique and agrees with the "Harsanyi-Selten" value (or "modified Shapley" value). In fact, uniqueness holds whenever the Pareto surface is a hyperplane, or ** coincides with a hyperplane over a sufficiently large compact set. (4) In the two-person case the solution is "almost always" unique, even without transferable utility, and agrees with the Nash cooperative solution for such games.***

We have no results indicating how well or poorly the two solutions approximate to each other, nor have we discovered any concrete example where the numerical contrast between them seems especially significant. But we can state two general properties of our present solution that are not satisfied by the other; either one, in fact, could be used (with suitable technical adjustments) in place of the principle of equivalence (III) in the derivation of our definition:

---

*The author and Martin Shubik have calculated explicit values, as functions of n, for an "Edgeworth" market game (i.e., bilateral exchange economy) with 2·n players. Interestingly, as n tends to ∞ the value payoffs (which are covariant only under positive linear transformations) and the competitive equilibrium payoffs (covariant under arbitarary order-preserving transformations) converge to the same limit. See "Pure competition, coalitional power, and fair division," International Economic Review (to appear); also The RAND Corporation, Memorandum RM-4917-1.

**For our present value, which can easily be shown to be individually rational, it is sufficient that the Pareto surface coincide with a hyperplane within the individually-rational zone. I do not know of any proof that Harsanyi's 1963 value is necessarily individually rational (but see Harsanyi, op. cit., p. 194, footnote 4).

***J. F. Nash, "Two-person cooperative games," Econometrica 21 (1953), pp. 128-140.

(IV) If two games have the same solution (payoffs and comparison weights), then any probability mixture of the two games -- i.e., a game in which the first move decides by chance which of the two given games is to be played---also has that solution.

(V) If a game is modified by allowing side payments (restricted or unrestricted in amount), at exchange ratios corresponding to the comparison weights associated with a solution of the original game, then that solution remains a solution of the modified game.

(The latter will be recognized as a converse of the "irrelevant alternatives" condition that we invoked earlier.)

Finally, for those familiar with the Harsanyi model, we might briefly describe the modification that would be required in order to obtain the present value. In effect, one must permit each syndicate, after maximizing its potential as measured by the weighted sum of the members' utilities, to pay dividends in "utility scrip," written in the utility units of any or all of its members, without regard to who receives it. One would then require, as an additional equilibrium condition, that all the scrip so issued be redeemed in the end--i.e., traded back to the players named on the face of the scrip, at the rates of exchange corresponding to the weights of the game. While the immediate effect of this modification (which could be formulated in other ways) is to complicate the bargaining model, there is indirect compensation in the new ability of the syndicates to transfer utility internally, which removes some of the difficulties associated with variable threats.

APPENDIX: THE EXISTENCE THEOREM

The mathematical formulation of the game Γ in normal, extensive, or characteristic-function form is not relevant for our purpose; our only assumption is that the set F of payoff vectors that are feasible for the all-player coalition without the use of side payments is compact and convex. Let $\Gamma(\lambda)$ denote the same game with the payoffs to players 1, 2, ..., n multiplied by $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, where λ is a point in the simplex $\Lambda = \{\lambda \geq 0 | \sum \lambda_i = 1\}$. Let $F(\lambda)$ denote the feasible set for $\Gamma(\lambda)$, and let $\varphi(\lambda)$ denote the side-payment value of $\Gamma(\lambda)$—i.e., the λ-transfer value of Γ. We assume that $\varphi(\lambda)$ is continuous in λ, Pareto optimal, and individually rational. From the latter it follows that $\lambda_i = 0$ implies $\varphi_i(\lambda) \geq 0$; this is the only use we make of individual rationality.

THEOREM Under the stated assumptions, there exists $\lambda \in \Lambda$ such that $\varphi(\lambda) \in F(\lambda)$.

Proof. Let $P(\lambda)$ be the set of vectors π such that $\Sigma\pi_i = 0$ and $\varphi(\lambda) - \pi \in F(\lambda)$. $P(\lambda)$ is nonempty, convex, and compact for each $\lambda \in \Lambda$ and is an upper-semi-continuous function of λ. Define the function T by

$$T(\lambda) = \lambda + P(\lambda) = \{\lambda + \pi | \pi \in P(\lambda)\}.$$

Let A be a simplex in the hyperplane $\{\alpha | \Sigma\alpha_i = 1\}$, large enough to contain all sets $T(\lambda)$, $\lambda \in \Lambda$, as well as Λ itself;

the upper-semi-continuity of T makes this possible. Extend the definition of T to A by

$$T(\alpha) = T(f(\alpha)), \text{ where } f_i(\alpha) = \frac{\max(0, \alpha_i)}{\Sigma_j \max(0, \alpha_j)} .$$

According to Kakutani's theorem, there is a "fixed" point α^* satisfying $\alpha^* \in T(\alpha^*)$. Denote $f(\alpha^*)$ by λ^*. Suppose first that $\alpha^* \neq \lambda^*$. Then $\alpha^* \in A - \Lambda$, and for some i $\lambda_i^* = 0 > \alpha_i^*$. But $\alpha^* \in T(\lambda^*) = \lambda^* + P(\lambda^*)$, hence $\pi_i^* < 0$ for some $\pi^* \in P(\lambda^*)$. Since $\varphi_i(\lambda^*) \geq 0$ by individual rationality, the feasible payoff vector $\varphi(\lambda^*) - \pi^* \in F(\lambda^*)$ gives player i a positive amount. But this is impossible without side payments, since all his payoffs are zero. We conclude that $\alpha^* = \lambda^*$; hence that $0 \in P(\lambda^*)$; hence that $\varphi(\lambda^*) \in F(\lambda^*)$.