ECONOMETRICA

VOLUME 22 January, 1954 NUMBER 1

INTRANSITIVITY, UTILITY, AND THE AGGREGATION OF PREFERENCE PATTERNS

BY KENNETH O. MAY[1]

During the first half of the twentieth century, utility theory has been subjected to considerable criticism and refinement. The present paper is concerned with certain experimental evidence and theoretical arguments suggesting that utility theory, whether cardinal or ordinal, cannot reflect, even to an approximation, the existing preferences of individuals in many economic situations. The argument is based on the necessity of transitivity for utility, observed circularities of preferences, and a theoretical framework that predicts circularities in the presence of preferences based on numerous criteria.

THEORIES of choice may be built to describe behavior as it is or as it "ought to be." In the belief that the former should precede the latter, this paper is concerned solely with descriptive theory and, in particular, with the intransitivity of preferences. The first section indicates that transitivity is not necessary to either the usual axiomatic characterization of the preference relation or to plausible empirical definitions. The second section shows the necessity of transitivity for utility theory. The third section points to various examples of the violation of transitivity and suggests how experiments may be designed to bring out the conditions under which transitivity does not hold. The final section shows how intransitivity is a natural result of the necessity of choosing among alternatives according to conflicting criteria. It appears from the discussion that neither cardinal nor ordinal utility is adequate to deal with choices under all conditions, and that we need a more general theory based on empirical knowledge of preference patterns.

I. CHOICE AND PREFERENCE

There are various ways in which an empirical reference may be given to statements about preference, but all those known to the writer consider choice as its observable evidence. The statement "x is preferred to y" is generally understood to mean that some organism chooses x "over," "before," or "in preference to" y. In order to make precise the meaning of choice it is necessary to specify the experimental conditions.

Suppose that an individual (or group) is confronted with precisely two possibilities, x and y, one and only one of which must be chosen, verbally or by some

[1] The author is indebted for assistance, comments, or suggestions to R. E. Adamson, W. Edwards, M. Friedman, D. Hawkins, H. S. Houthakker, T. Koopmans, D. McGarvey, A. G. Papandreou, S. W. Rousseas, and various participants in the December, 1952 session of the Econometric Society at which an earlier version of this paper was presented.

other overt act, under certain conditions. If we are to judge the preference by a single observation of this kind, we are inclined to say that whichever is chosen is the preferred alternative. However, if the experiment is repeated, it may happen that the individual does not always choose the same. It then seems reasonable to identify preference with a greater frequency of choice and to write yPx, xIy, or xPy according as $p(x) <, =,$ or $> p(y)$. Here P and I stand for preference and indifference, and $p(z)$ refers to the probability of the occurrence of the choice of z.[2] This notation is not entirely satisfactory since it fails to indicate explicitly the field of choice. Accordingly, we write $p(x \mid x, y, E)$ to mean the probability that x is chosen from x and y under experimental conditions E.

It easily follows from the above definition of preference that P and I have the following properties.

(1) One and only one of yPx, xIy, xPy holds. (Trichotomy.)
(2) xIx. (Reflexivity for I.)
(3) xIy implies yIx. (Symmetry for I.)
(4) xPy implies not-(yPx). (Anti-symmetry for P.)

However, it does *not* follow from the definition that either P or I is transitive, i.e. that

(5) xPy and yPz imply xPz. (Transitivity for P.) (?)
(6) xIy and YIz imply xIz. (Transitivity for I.) (?)

This is easily seen if we translate these assertions into probability statements. For example, (5) becomes

(5′) $p(x \mid x, y) > p(y \mid x, y)$ and $p(y \mid y, z) > p(z \mid y, z)$ imply $p(x \mid x, z) > p(z \mid x, z)$. (?)

Since the probabilities are calculated with respect to different fundamental probability sets, we cannot assert the equality of probabilities with the same first argument.[3] Accordingly, the transitivity of $>$ and $=$ does not imply (5) or (6). Of course, this does not mean that preferences are not transitive. It means simply that transitivity does not follow from this empirical interpretation of preference, but must be established, if at all, by empirical observation.[4]

In economics we are usually concerned with a set of alternatives such that an individual is able to choose between *any pair* in such a way that (1)–(4) hold for every pair.[5] We call such a set with such relations a *binary preference pattern.*[6]

[2] Here probability is meant in the strict relative frequency sense. This kind of empirical reference has been used by psychologists and economists. An interesting special case in which 0, ½, and 1 are the only possible probabilities is treated by Georgescu-Roegen in "The Theory of Choice and the Constancy of Economic Laws," *Quarterly Journal of Economics*, Vol. LXIV, No. 1, February 1950, pp. 125–138.

[3] J. Neyman, *First Course in Probability and Statistics*, Henry Holt and Co., New York, 1950, Ch. 2.

[4] Essentially this point was made by Georgescu-Roegen in his "The Pure Theory of Consumer Behavior," *Quarterly Journal of Economics*, Vol. L, No. 4, August 1936, pp. 579–580.

[5] See for example Paul A. Samuelson, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, 1947, p. 94. It may be that this way of formulating the problem of

If in addition (5) and (6) hold, we call the pattern an *ordering*. If I is excluded we call the pattern *strong*. Since our discussion is of binary preference patterns, whether interpreted in terms of our particular empirical reference or some other, what follows applies to any theory of preference in which (1)–(4) hold, and this includes, so far as this writer knows, all theories of economic choice.

II. TRANSITIVITY AND UTILITY

The formulation of the problems of economic choice and preference in terms of the algebra of binary relations is a recent phenomenon.[7] The traditional approach has been, rather, to assume the existence of a utility function that represents in some fashion the desirability of the alternatives. In the crude Benthamite version, the utility function is defined over all alternatives and takes numerical values that represent the cardinal utilities assumed to exist. The more sophisticated modern version assumes a function that represents the utilities only ordinally, i.e. if we represent the utility function by $\phi(z)$, then $\phi(x) > \phi(y)$ if and only if xPy. The cardinal utility function has this property also.[8] The assumption that values can be treated like numbers, at least ordinally, is consistent with linguistic expressions of the form "A likes x *more* than y" and with the pricing propensities of a pecuniary society. It does not follow, however, that it is necessary to rationality.

It is easy to show that the existence of a utility function, either cardinal or ordinal, implies the transitivity of preferences. In fact, xPy and yPz imply $\phi(x) > \phi(y)$ and $\phi(y) > \phi(z)$. By the transitivity of $>$, these imply $\phi(x) > \phi(z)$ and hence xPz. A similar argument holds for I.[9] This means that transitivity is a logical consequence of the assumption of a utility function. Of course, (1)–(4) follow also, so that the assumption of utility implies a preference ordering. This

consumer choice is artificial and that consumer decisions are really made without pairwise comparisons. But only careful experimentation can answer this question, and, in any case, we are concerned here with a critique of the currently accepted binary theory.

[6] The reason for including the word "binary" in the above definition is that we can conceive of other ways of empirically defining preferences that involve more than two elements. Suppose for example that we confront an individual with n alternatives $x_1 \cdots x_n$ and require him to choose one and only one. A simple example is an experiment in which the subject is confronted with two possibilities and asked to indicate verbally preference for one of them or else indifference. This is not a binary choice since here $n = 3$. In general, if we identify, x_iPx_j with $p(x_i \mid x_1 , \cdots , x_n , E) > p(x_j \mid x_1 , \cdots , x_n , E)$, it is easy to see that (1) – (4) hold and that in addition (5) and (6) hold for any triple in the set $x_1 , \cdots , x_n$. Thus intransitivity can arise only from comparison of preferences defined over different sets of alternatives. In a given set of alternatives the relation xPy may be referred to any subset including x and y, and there is no reason why we might not have xPy, xIy, and yPx all present for some subsets. This does not imply "inconsistency" of the individual but merely that he may make difference choices in the presence of different possibilities.

[7] Initiated by K. J. Arrow in his *Social Choice and Individual Values*, New York, 1951. Arrow starts from the relation R which is equal to (P or I).

[8] See Samuelson, *op. cit.*, Ch. V.

[9] Transitivity for P is independent of transitivity for I as can be shown easily by examples. For instance, if we define over the set of the first n positive integers yPx to mean "$y > x$ and $y - x$ is even," we find that P but not I is transitive.

result means that transitivity is necessary for utility. It makes it possible to test the hypothesis of the existence of a utility function by experimental tests of transitivity. Where transitivity does not hold, the use of utility is not justified.

A recent tendency is to derive utility from plausible axioms instead of assuming it.[10] Transitivity is always included, since it is necessary as indicated above. It is easy to see that (1)–(6) over a *finite* set of alternatives imply a utility function. For transitive I is an equivalence relation and so partitions the alternatives into mutually exclusive equivalence classes, such that indifferent elements belong to the same class and the classes are ordered in the same way as any set made up of an arbitrary element from each class. We call these *indifference classes*.[11] Now if the set of alternatives is finite, the set of indifference classes is also, and hence they may be arranged in a preferential sequence and numerical utilities assigned in the same numerical order.

If the set of indifference classes is infinite, however, we cannot state that they can be arranged in a sequence or even that they can be assigned numerical utilities.[12] The existence of a utility function, i.e., of a function mapping the set of indifference classes onto a subset of the real numbers so as to preserve order, is equivalent to the set of indifference classes having the same "order type" as some subset of the real numbers in their natural order.[13] Now it is a well-known fact of set theory that ordered sets are by no means all of the same order type as some subset of the real numbers in natural order. In the first place two ordered sets of the same type must have the same cardinal, so that a strong preference ordering of, say, the set of all real-valued functions on the unit interval, which has a cardinal greater than the continuum, could not be represented by a utility function. There would not be enough numbers "to go round" in assigning utilities. Also, not every ordered set of cardinality less than or equal to that of the continuum is of the same order type as a subset of the real numbers in natural order.[14]

The above discussion means that transitivity is not sufficient for utility. However, for practical purposes it may be considered to be so. In the first place we have seen that it is sufficient for finite sets, and these are the only kind with which people are actually confronted. Secondly, the sufficiency holds for a very large class of sets that includes those ordinarily used to idealize sets of alternatives. This class includes all those sets of indifference classes that can be parti-

[10] See, for example, Jacob Marschak, "Rational Behavior, Uncertain Prospects, and Measurable Utility," ECONOMETRICA, Vol. 18, No. 2, April 1950, pp. 111–141.

[11] Garrett Birkhoff and Saunders MacLane, *A Survey of Modern Algebra*, The Macmillan Company, New York, 1953, pp. 155–156. The indifference classes are, of course, the familiar loci of indifference or, in the continuous case, the indifference curves. Transitivity of I is necessary and sufficient for their existence, regardless of the transitivity of P.

[12] Professor Marschak states the contrary in the article cited in footnote 10, and hence some further assumption about the alternatives must be implicit in his thinking.

[13] See E. Kamke, *Theory of Sets*, Dover Publications, New York, 1950, Ch. 3, and R. L. Wilder, *Introduction to the Foundations of Mathematics*, John Wiley and Sons, New York, 1952, Ch. V.

[14] Gerard Debreu cites the so-called lexicographical ordering of the points in the plane by $(x, y)P(x', y')$ if and only if either $x > x'$ or else $x = x'$ and $y > y'$. (Cowles Commission Discussion Paper: Econometrics No. 2040.)

tioned into a finite or denumerable number of sets each of which is either finite, denumerable and discrete, denumerable and dense, or continuous and separable.[15]

We have shown that transitivity plays a key role in the utility theory of choice. Its necessity gives us the possibility of making critical empirical tests of the validity of utility arguments. Its practical sufficiency means that as long as we assume it we may as well use utility functions.

III. EVIDENCE OF INTRANSITIVITY

It is a familiar fact that some preference patterns are transitive. For example, the preference pattern of an individual confronted with different amounts of cash will usually be an ordering according to the amounts. In fact, the discussion of the previous sections suggests that transitivity holds just when a money price (a utility expressed in money terms) can fully express the preference pattern, at least ordinally. If it were true that "everything has a price" reflecting its preference status, intransitivity of value judgments could hardly arise.

We shall cite examples of triples that are circular, i.e. such that xPy, yPz, and zPx. It is by no means asserted that there are any larger sets in which every triple is circular. In fact, a little experimentation will convince the reader that a pattern of four or more alternatives must contain at least two transitive triples.[16] What is asserted is that circular triples actually do occur in individual preference patterns. Moreover, the evidence suggests that the typical pattern contains what we shall call *cycles*, i.e. sets $x_1, \cdots, x_n$ in which x_iPx_j for all $i < j$ except that x_nPx_1. We call n the *order* of the cycle. Circular triples are cycles of order three.

The cycle of order three that arises from voting by three individuals is well known as the "paradox of voting." Letting x, y, and z be the three candidates, the individual orderings are represented by xyz, yzx, and zxy. It is easy to see that majority vote gives xPy, yPz, and zPx. Now suppose that we have a fourth individual who is indifferent to the candidates but thinks that majority vote should decide. Then his binary preference will be circular. Since an arbitrary binary preference pattern is the group pattern of some set of individuals by the method of simple majority decision, an individual addicted to majority voting as a method of decision may have any arbitrary preference pattern.[17] More gen-

[15] In order to save space we omit a discussion of the way in which the order preserving mapping is established in each case. The appropriate theorems and definitions of terms will be found in Kamke, *op. cit.*, Ch. III and in Wilder, *op. cit.*, Ch. VI.

[16] An enumeration of all possible patterns on 4 or 5 alternatives supports the conjecture that not more than half the triples in any pattern can be circular. If we abstract from the interpretation in terms of choice, a preference pattern is simply a set subject to a relation satisfying (1) – (4). Binary relations and their geometric representation by means of line diagrams (under the unfortunate term "graph theory") have been studied to some extent by mathematicians, but there is available in English very little material. A useful introduction that includes a bibliography is *Graph Theory as a Mathematical Model in Social Science*, by Frank Harary and Robert Z. Norman, Institute of Social Research, University of Michigan, Ann Arbor, 1953.

[17] David McGarvey, "A Theorem on the Construction of Voting Paradoxes," ECONOMETRICA, Vol. 21, No. 4, pp. 608–10.

erally, if we have any relations P_o and I_o satisfying (1)–(4), and if an individual has incentives to choose x over y if and only if xP_oy, then his preference pattern will be of the same structure as P_o and I_o and hence not necessarily transitive.[18] In this way we can create arbitrary individual patterns simply by constructing an arbitrary relation and giving an individual an incentive to make his choices accordingly.

The previous discussion was intended to indicate two ways in which intransitive preference patterns may arise, namely from an intransitive relation connected with choice and from the aggregation of transitive patterns. We shall be concerned here with the latter, which appears to be of greatest interest to economists. Returning to the paradox of voting, suppose we think of the fourth individual as comparing the candidates on the basis of three characteristics, i.e. their standing with each of the three voters. Considered in this way the paradox of voting throws light on the nature of individual action in the presence of numerous components. Might not circularities arise if alternatives were ordered in conflicting ways according to different criteria? It was such considerations that led to the following simple pilot experiment.

The subjects were 62 college students. The alternatives were three hypothetical marriage partners, x, y, and z. In intelligence they ranked xyz, in looks yzx, in wealth zxy. The structure of the experiment was not explained, but subjects were confronted at different times with pairs labelled with randomly chosen letters. On each occasion x was described as very intelligent, plain looking, and well off; y as intelligent, very good looking, and poor; z as fairly intelligent, good looking, and rich. All prospects were described as acceptable in every way, none being so poor, plain, or stupid as to be automatically eliminated. Each individual's responses were kept together. During the experiment proper, the subjects were never confronted with all three alternatives at once. Later they were asked to order all three. They showed a lively interest in the choices, which were connected, if only in a symbolic fashion, with preferences involved in their own real decision making. Parts of the experiment were repeated to test for consistency and possible capriciousness. The results, as well as the behavior of the subjects, indicated practically no random element in the choices. In terms of the probability definition of preference given in the first section, it was evident that 0 and 1 were the only possible probabilities and that repeated trials were not necessary.

Since indifference is ruled out, there are six possible orderings and two circular patterns designated by $xyzx$ and $xzyx$. If group preferences be defined by majority vote, the results indicate a circular pattern, since x beat y by 39 to 23, y beat z by 57 to 5, and z beat x by 33 to 29. The number of individuals having each of the possible patterns was xyz: 21; $xyzx$: 17; yzx: 12; yxz: 7; zyx: 4; xzy: 1;

[18] Among observed non-transitive relations linked with preferences (choices) may be mentioned the dominance concept in game theory, numerous examples of combat superiority (battleship sinks destroyer sinks submarine sinks battleship; mongoose kills cobra kills cat kills mongoose; tank beats machine gunner beats bazookaman beats tank, etc.) and winning ability in such games as "rock breaks scissors cuts paper wraps rock."

zxy: 0; and *xzyx*: 0. The intransitive pattern is easily explained as the result of choosing the alternative that is superior in two out of three criteria. The orderings *xyz* and *yzx* seem to have resulted from giving heavier weight to intelligence and looks respectively. The four who chose inversely with respect to intelligence (*zyx*) were men and may indicate the extent of male fear of intelligent women. The seven who chose inversely with respect to wealth (*yxz*) must not be considered to have a wanton disregard for money. They may well have preferred *y* over *x* because of a wide disparity in looks, *x* over *z* because of a wide disparity in intelligence, and *y* over *z* because of a combination of looks and intelligence. When required to rank all three alternatives, those with intransitive patterns scattered, most choosing *yzx* (9) and *yxz* (4). Those with transitive orderings for binary choices for the most part made the obvious orderings.

What is the significance of this experiment? Of course it does not prove that individual patterns are always intransitive. It does, however, suggest that where choice depends on conflicting criteria, preference patterns *may* be intransitive unless one criterion dominates. It suggests how we might construct further experiments displaying circularities, and also how we might organize experiments that would display only transitivity! It also throws some light on examples of intransitivity that have been observed in various fields.

A preference cycle said to have been observed during the war concerns the behavior of pilots in burning planes. When confronted with desperate choices between pairs from the set {flames, red hot metal, falling}, pilots most often made the choices (flames) *P* (red hot metal), (red hot metal) *P* (falling), and (falling) *P* (flames). The previous discussion suggests the following intentionally naive explanation. Suppose that the pilot has been conditioned to recoil from hot objects, the reaction to red hot metal being even stronger than to fire. Suppose also that he is accustomed to support himself as solidly as possible and to react against lack of visible support. Finally suppose that he knows that his life is endangered most by falling, somewhat less by putting up with red hot metal, and least by flames. The first two suppositions seem to conform to experience; the third does not seem unreasonable. On these assumptions, the pilot "ranks" the alternatives in a way exactly similar to the patterns in the paradox of voting and the marriage partner experiment. Of course this explanation is a caricature of the complexities of human reactions, but it may hint at the true explanation in terms of multiple components involved in alternatives and the corresponding reactions.

Experiments have shown that when rats are sufficiently hungry they will prefer food to sex, sex to avoidance of pain, and avoidance of pain to food.[19] While only a minority of human beings behave like rats, it is still true that such experiments throw light on human choice, since the neurological mechanisms may be similar. In fact, a rather simple nervous network with the proper topology is easily seen to determine an intransitive preference pattern of responses to

[19] Warren S. McCulloch, "A Recapitulation of the Theory," in *Teleological Mechanisms, Annals of the New York Academy of Science*, Vol. 50, Art. 4, October 1948, p. 263.

pairs of stimuli.[20] More complicated reaction mechanisms of a similar kind may underlie the fact that choice patterns arising out of attitudes are not always found to be "scalable" by social psychologists.[21]

The alternatives of interest to the economist are typically commodity bundles, i.e. vectors whose components are quantities of goods and services of various kinds. Even an individual commodity is really a vector of its specifications and other attributes such as its price.[22] The previous discussion suggests that where the components are ranked in certain conflicting ways, we might expect circularities. The experiments conducted by M. M. Flood at The Rand Corporation support this conjecture.[23] Subjects were presented with alternatives consisting of physical objects of household utility and quantities of money. They were asked to indicate binary preferences and also to rank the alternatives. In one experiment, each of three subjects showed a cycle of order four over a set of ten alternatives. In another, of 21 subjects 11 showed circularities, typically of order greater than three and in one case of order six.

No doubt there is other evidence not known to this writer, but what is available indicates that the question is no longer "Are preferences transitive?" but rather "Under what conditions does transitivity fail?"[24] Of course the whole issue may be avoided by simply asserting transitivity as part of the definition of "rational behavior." The question then is whether rational behavior as so defined has very much importance, either descriptive or normative. Still another way of avoiding inconvenient circularities is to define troublesome alternatives as "not comparable." But it is just these "non-comparable" cases that are of interest. Comparison only of alternatives in which one is superior to the other in every respect makes for a simple but rather trivial theory. Incidently, the relation of non-comparability is itself intransitive! There seems no way to avoid considering intransitivity as a natural phenomenon.

[20] Warren S. McCulloch, "A Heterarchy of Values Determined by the Topology of Nervous Nets," *Bulletin of Mathematical Biophysics*, Vol. 7, 1945, pp. 89–93. See also p. 227 in the same volume. McCulloch, on pages 92–93 of this article, states the case against the transitivity axiom as follows: "Experimental aesthetics, economics, and conditioned reflexology have produced instances in which, under constant conditions, preference was circular. One such instance would have been sufficient basis for categorical denial of the subsumption that values were magnitudes of any one kind. Thus for values there can be no common scale."

[21] See *Measurement and Prediction*, Vol. IV of *Studies in Social Psychology in World War II*, Princeton, 1950.

[22] In *Consumer Reports* for November, 1952 ten electric shavers are ranked according to six different criteria: closeness of shave, absence of irritation, speed, ability to trim, effectiveness on long hair, and ease of cleaning. With the price we have seven components. The rankings were by no means consistent, e.g. the best shaver by one criterion was the worst by another!

[23] M. M. Flood, "A Preference Experiment," P-256, P-256, P-258, and P-263, November 1951–January, 1952, The Rand Corporation, Santa Monica, California.

[24] See the comments by Ward Edwards in the report of the session on "Individual Preference Functions" in ECONOMETRICA, Vol. 21, No. 3, July 1953, pp. 476–477.

IV. AGGREGATION OF PREFERENCE PATTERNS

The purpose of this section is to show how intransitive preference patterns may arise from the aggregation of preference orderings. We consider n alternatives that have been ordered according to $m \geqslant 3$ criteria. The i-th alternative may be characterized by a vector $X_i = (x_{i1}, \cdots, x_{im})$, whose components are real numbers such that $X_iP_jX_k(X_i$ is preferred to X_k by the j-th criterion) if and only if $x_{ij} > x_{kj}$. We may interpret the components as aspects of alternatives among which an individual must choose, or as reflecting judgments of different individuals in a group. In either case, the problem of aggregation is to determine a preference pattern of the vectors from a knowledge of the patterns of the components. If the components represent the individual orderings of "social states," we have Arrow's problem of constructing a "social welfare function." We shall think in terms of an individual confronted with conflicting criteria applied to a set of alternatives. If Q_j represents the pattern according to the j-th criterion, the problem is concerned with a function of the form

$$Q = F(Q_1, \cdots, Q_m) \tag{7}$$

that gives a pattern Q corresponding to each set of component patterns $\{Q_i\}$. We call it a *preference aggregating function.*

Since there is a one-to-one correspondence between functions F and methods of aggregation, we shall proceed by making plausible and weak assumptions about F. This will be simplified by introducing a more precise characterization of a binary preference pattern. It will be recalled that a pattern is uniquely determined by indicating which of three possible relations exists for each pair of alternatives. Moreover, unless transitivity is assumed, the relations among different pairs are independent, so that it is necessary as well as sufficient to specify $\binom{n}{2}$ relations in order to determine a pattern on n alternatives. Let $D(x, y) = -1, 0, 1$ according as yPx, xIy, xPy. Then the values of $D(x, y)$ over all pairs uniquely determine the pattern. We may think of Q's in (7) as functions of this kind, so that the aggregation maps a set of m such functions into a function. Since the range of both x and y is the set of n alternatives, we can designate each alternative by one of the first n integers. Then the values of $D(x, y)$ may be thought of as a matrix in which $D(x, y)$ is the element in the x-th row and y-th column. Since $D(y, x) = -D(x, y)$, the matrix is skew-symmetric. Then (7) is a function aggregating matrices.

Our first assumption is that the aggregate preference relation between a pair is determined completely by the component relations between the pair, i.e. that the aggregation, like the patterns themselves, is binary. This is *not* to say that aggregate preference is independent of the component preferences among other pairs (the component preferences of different pairs may be related), but simply that component preferences between other pairs are involved in the aggregation only via the components of the pair in question. There are, of course, situations where aggregate choices are not binary, as we indicated in the first section, but we are limiting ourselves here to the binary case. Empirically we have in mind an experiment in which an organism is required to make a choice between just two alternatives. We are assuming then that this choice reaction is determined completely by component reactions to the various aspects of the alternatives, and by other preferences and factors only in so far as these may affect the component reactions to these two alternatives. To claim otherwise means in effect either that not all relevant components have been included or that other alternatives may be chosen.[25] The assumption may be stated formally as follows:

$$\text{(8)} \qquad \text{If } D_j(x, y) = D'_j(x, y) \text{ for all } j, \text{ then } D(x, y) = D'(x, y).$$

Here the D are the aggregates corresponding to the components $\{D_j\}$. The idea could also be stated in terms of the Q's by saying that if two sets of patterns $\{Q_j\}$ and $\{Q'_j\}$ are the same for a pair of alternatives (x, y), then the corresponding Q and Q' are the same.

This assumption is formally equivalent to Arrow's condition 3.[26] He states it for any subset of alternatives, instead of just for any pair. However, if it holds for any subset, it certainly holds for any pair. Also, if it holds for any pair, it holds for any subset since a binary pattern is uniquely determined by the relations among pairs.[27] However, we prefer not to give it the interpretation suggested by the name that Arrow assigned to it, "the independence of irrelevant alternatives." As we remarked above, the relations between different pairs may be linked in many ways in particular cases. A person's preferences are all the outcome of his whole experience and so no one can be really independent of the rest. Also the preference relations in a transitive pattern are not independent since some of them cannot be changed without changing others. What (8) asserts is something quite different, namely that a person's total reaction, when confronted with two and only two alternatives of which one must be chosen, is determined completely by *all* his various reactions to *those* two alternatives, since by definition of the experimental situation they are the only stimuli offered. Of course his other experiences are involved as conditioners, but they are determinative only in so far as they affect his reactions to the alternatives before him. Accordingly (8) means simply that all aspects of the alternatives are included in the components and that we are dealing with binary choices.

The problem is now reduced to determining for each pair of alternatives a function that gives the preference corresponding to each set of component preferences. Thus for an ordered pair (x, y) we wish a function that gives a $D(x, y)$ corresponding to each set $\{D_j(x, y)\}$. For each (x, y) we then have a function

$$\text{(9)} \qquad D = f(D_1, \cdots, D_m) = f(\{D_j\}).$$

[25] The situation in which a voting body decides on one of several alternatives by voting on pairs or on one at a time is not a binary situation. The voters know that they are deciding among more than two alternatives. Their preferences are not the same as if they were presented with two and only two alternatives.

[26] Arrow, *op. cit.*, p. 27.

[27] *Ibid.*, p. 28.

Here all variables have the range $\{-1, 0, 1\}$. The functions may of course be different for different ordered pairs. However, it should make no difference if we let the values corresponding to (x, y) be elements in the y-th row and x-th column of matrices. In this representation each term of a matrix will be the negative of the term in the corresponding matrix of the original representation. Hence the aggregating function for (x, y), f_{xy}, must yield the negative of the aggregating function for (y, x), f_{yx}, when the arguments of one are the negatives of the arguments of the other, i.e.

$$(10) \qquad f_{xy}(\{D_i\}) = -f_{yx}(\{-D_i\}).^{28}$$

We now state four very weak conditions on the decision functions f_{xy}. We usually omit the subscripts, since the conditions are assumed to hold for all pairs, even though the functions may be different. First we wish the decision functions to be defined for any set of $\{D_i\}$. This is certainly desirable since the components may have quite varied rankings.

Condition A: The decision function is defined and single valued for every set $\{D_i\}$ where each D_i may take the values -1, 0, or 1.

This rules out, of course, the consideration of decision as a stochastic process, even though we have defined preference in statistical terms. A more general theory would replace Condition A by the assertion that each set of components determined a probability distribution for the aggregate.

Our second condition is merely that f be an averaging function in the weak sense that if all components are the same, then the aggregate takes this same value.[29]

Condition B: $D_o = f(D_o, \cdots, D_o)$ for $D_o = -1, 0, 1$.

This implies, of course, that $Q_o = F(Q_o, \cdots, Q_o)$, where Q_o is any pattern. It is a very weak condition, since without it "unanimity" for an alternative would not achieve it.

Next we wish to state a condition that guarantees that the aggregation is at least not negatively responsive to the components. In short, we wish f to be positively monotonic, i.e.

Condition C: If $D_i' \geqslant D_i$ for all i, then $D' \geqslant D$.

This guarantees, for example, that if everything remains the same except that one component changes favorably toward an alternative, the result will not be less favorable to that alternative.

Our final condition is that one component does not completely dominate.

[28] This is a matter of notation, since it means merely that if we rename everything, including the functions, we get the same results. It must not be confused with the assumption of symmetry or oddness for f_{xy}. Oddness for f_{xy} follows from the assumption that $f_{xy} = f_{yx}$, but we are *not* assuming this. Compare Condition III, Kenneth O. May, "A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision," ECONOMETRICA, Vol. 20, No. 4, October 1952, pp. 680–684.

[29] See Edward L. Dodd, *Lectures on Probability and Statistics*, University of Texas Press, 1945, pp. 20 and 35.

Condition D: For every i_o there is a set of components including $D_{i_o} = -1$ and such that $D > -1$.
If this were false, it would mean that there was an i_o such that $D_{i_o} = D$ regardless of the other components. This would mean that there would be no aggregation problem, since only one component would matter. Now in actual practice organisms often do determine aggregate preferences by considering only one component. Condition D merely limits us to the cases where a balancing problem really exists and is to be solved rather than avoided.

Anyone familiar with Arrow's basic work will recognize these four conditions as very similar to his Conditions 1, 2, 4, and 5. Actually, (8) together with Conditions A–D are *almost* equivalent to Arrow's Conditions 1–5, and the exact relation is important because of the use we wish to make of his results. First it should be noted that Arrow from the beginning makes the assumption that both components and aggregates are orderings. We prefer not to do this on the general heuristic principle that properties of aggregates should be derived rather than assumed to be the same as those of micro-elements. Since transitivity is not formally stated as one of Arrow's conditions, we shall think of them without this requirement, as though transitivity were assumed in a sixth Arrow condition.[30]

We have already noted that our (8) is equivalent to Arrow's Condition 3. Our Condition A implies his Condition 1, since Condition A asserts that the aggregating function is defined for all sets of orderings while his asserts that this is so for at least three alternatives. Arrow wished to state the weakest possible conditions under which his principal result could be proved. Our stronger Condition A means that we are trying to construct an aggregation theory that will be able to deal with all the problems that might be posed to it.

Our Condition B implies Arrow's Condition 4 that the aggregate pattern not be "imposed." His Definition 4 of imposition means that if and only if the aggregate is not imposed there will be a set of components yielding yPx, i.e. $D = 1$, for every pair of alternatives. But our Condition B exhibits this set of components. However, Arrow's Condition 4 does not imply Condition B because the former does not rule out the possibility that aggregate indifference may not be accessible, while the latter guarantees that it can be achieved.[31]

Condition C is equivalent to Arrow's Condition 2 provided we accept his Condition 3, which is equivalent to our (8). For by (8) the aggregate relation between two alternatives depends only on the component relations for those alternatives. The hypothesis of Arrow's Condition 2, in so far as it concerns the two alternatives in the conclusion, is that xR_iy implies xR'_iy, xP_iy implies xP'_iy,

[30] His definition 4 of the social welfare function includes this sixth requirement. All the conditions discussed here are to be found in Chapter III of Arrow's book previously cited.

[31] Arrow apparently meant Condition 4 to imply Condition B since he wrote ". . . we certainly wish all choices to be possible if unanimously desired by the group." (p. 29). Since his conditions can be satisfied by an aggregating function that never yields indifference, he may have meant to exclude indifference from the meaning of "choice." Certainly Condition 4 does imply Condition B for $D = \pm 1$, provided we take into account his Condition 2, which we show below is equivalent to our Condition C. For if $f(1, 1, \cdots, 1) < 1$, then $D < 1$ for all sets of components, since $D_i \nless 1$ for all i.

and xPy. Since his xRy means xPy or xIy, this hypothesis is that $D_i \geqslant 0$ implies $D'_i \geqslant 0$, $D_i = 1$ implies $D'_i = 1$, and $D = 1$. Since his conclusion is $xP'y$ or $D' = 1$, his Condition 2 can be restated: If $D'_i \geqslant D_i$ and $D = 1$, then $D' = 1$. It is easy to see that this is implied by Condition C. The converse is also true. Obviously this is so in the case $D = 1$. It holds trivially in the case $D = -1$, since the conclusion of Condition C is always true in this case. Suppose now that under the hypotheses of Condition C we have $D = 0$ but $D' = -1$. Consider the aggregating function for the pair (y, x). Since $D(y, x) = -D(x, y)$ and $D_i(y, x) = -D_i(x, y)$, the same situation may be described in terms of f_{yx} as one in which $D = 0$, $D' = 1$, yet $D_i \geqslant D'_i$. Since this is incompatible with Arrow's Condition 2, the conclusion of Condition C follows in this case also, and the equivalence is proved.

Finally, Condition D is equivalent to Arrow's Condition 5, which asserts that there is no "individual" whose strong preference determines the aggregate regardless of other preferences. In fact, Condition D is just a restatement in our terms of Arrow's Condition 5, taking into account his Definition 6.

This discussion shows that (8) with Conditions A–D imply Arrow's Conditions 1–5 if we omit from them the implicit transitivity requirements. The converse almost holds, being denied by slightly stronger requirements as to the range of definition and unanimity. Now Arrow's General Possibility Theorem asserts that his conditions are inconsistent for $n \geqslant 3$ when the transitivity of the components and aggregate is included.[32] Another way of stating this result is then that our Conditions A–D imply that there is some set of component orderings that yields a non-transitive aggregate pattern. We do not include (8) explicitly in this statement since, as we pointed out, it really amounts to no more than an assertion that we are dealing with a binary situation. In still other words, we may conclude that among binary preference patterns (8) arising from the aggregation (B) of three or more binary patterns (8) there are non-transitive patterns arising from transitive components (A) unless the method of aggregation fails to be non-negatively responsive (C) or unless one component dominates (D).

Arrow's work showed that we cannot count on transitivity of group preferences even if individual preferences are transitive. The present discussion shows that we cannot expect individual preferences to be always transitive. The expectation of intransitive group preferences is, of course, increased by these considerations. We intentionally did not include in the statement of conditions on the aggregating function any requirements on the component patterns, so that the theory here applies to aggregation of any binary patterns. It appears plausible, however, that any observed preference pattern may be analyzed eventually into transitive components. Accordingly, if we wish to explain the complex behavior patterns of groups, we might hope to resolve them into the behavior patterns of their members and finally into the transitive component preferences of individuals.

Carleton College

[32] Arrow, *op. cit.*, p. 59.