INDIVIDUAL CHOICE BEHAVIOR

A Theoretical Analysis

R. DUNCAN LUCE

PROFESSOR OF PSYCHOLOGY

UNIVERSITY OF PENNSYLVANIA

New York • John Wiley & Sons, INC.

LONDON • CHAPMAN & HALL, LIMITED

chapter 1

THE BASIC THEORY

A. INTRODUCTION

One large portion of psychology—including at least the topics of sensation, motivation, simple selective learning, and reaction time—has a common theme: choice. To be sure, in the study of sensation the choices are among stimuli, in learning they are among responses, and in motivation, among alternatives having different preference evaluations; and some psychologists hold that these distinctions, at least the one between stimulus and response, are basic to an understanding of behavior. This book attempts a partial mathematical description of individual choice behavior in which the distinction is not made except in the language used in different interpretations of the theory. Thus the more neutral word "alternative" is used to include the several cases.

In essence, the approach taken—in this respect, by no means novel—is orthogonal to that of *S-R* psychology, but not at variance with it. Rather than search for lawfulness between stimuli and responses and attempt to formulate a theory to describe those relationships, we shall be concerned with possible lawfulness found among different, but related, choice situations, whether these are choices among stimuli or among responses. Possibly the simplest prototype of this type of theory is the frequently assumed rule of transitivity among choices: given that a person chooses a over b and that he chooses b over c, then he chooses a over c when a and c

are offered. This assumption, were it true, would be a law relating a person's choice in one situation to those in two others, not a law relating responses to stimuli. It is evident that a sufficiently rich set of relations of this sort, coupled with a few simple *S-R* connections, will allow one to derive many more, and possibly quite complicated, *S-R* connections.

Such an approach seems to merit careful consideration, since several decades of pure *S-R* psychology have not resulted in notably simple laws of behavior. However, there seems little point in trying to discuss in detail its merits and demerits now, except to mention it in order to avoid confusion later. The results that follow—which seem to afford some insight into, and some integration of, psychological and psychophysical scaling, utility theory, and learning theory—will implicitly serve as the argument for the course taken.

1. Probabilistic vs. Algebraic Theories

A basic presupposition of this book is that choice behavior is best described as a probabilistic, not an algebraic, phenomenon. That is to say, at any instant when a person reaches a decision between, say, a and b we will assume that there is a probability $P(a, b)$ that the choice will be a rather than b. These probabilities will generally be different from 0 and 1, although these extreme (and important) cases will not be excluded. The alternative is to suppose that the probabilities are always 0 and 1 and that the observed choices tell us which it is; in this case the algebraic theory of relations seems to be the most appropriate mathematical tool.

The decision between these two approaches does not seem to be empirical in nature. Various sorts of data—intransitivities of choices and inconsistencies when the same choices are offered several times—suggest the probabilitistic model, but they are far from conclusive. Both of these phenomena can be explained within an algebraic framework provided that the choice pattern is allowed to change over time, either because of learning or because of other changes in the internal state of the organism. The presently unanswerable question is which approach will, in the long run, give a more parsimonious and complete explanation of the total range of phenomena.

The probabilistic philosophy is by now a commonplace in much of psychology, but it is a comparatively new and unproven point of view in utility theory. To be sure, economists when pressed will admit that the psychologist's assumption is probably the more accurate, but they have argued that the resulting simplicity warrants an algebraic idealization. Ironically, some of the following results suggest that, on the contrary, the idealization may actually have made the utility problem artificially difficult.

2. Multiple Alternative Choices

Once choice behavior is assumed to be probabilistic, a problem arises which does not exist in the algebraic models. Complete data concerning the choices that a person makes from each possible pair of alternatives taken from a set of three or more alternatives do not appear to determine what choice he will make when the whole set is presented. Because they cannot escape multiple alternative choice problems economists have been particularly sensitive to this feature of probabilistic models, and it has undoubtedly been one source of their resistance in admitting imperfect discrimination. Early psychologists, particularly learning theorists, studied multiple alternatives experimentally, but since the data seemed dreadfully complicated a trend set in toward fewer and fewer alternatives until now many studies employ only two. For the most part, present-day psychologists have been willing to ignore—or, to be more accurate, to bypass and postpone—the connections between pairwise choices and more general ones. And so the relations have remained obscure.

We shall center our attention on this problem. The method of attack is to introduce a single axiom relating the various probabilities of choices from different finite sets of alternatives. It is a simple and, I feel, intuitively compelling axiom that appears to illuminate many of the more traditional problems, in particular the question of whether or not a comparatively unique numerical scale exists which reflects choice behavior. Such a scale, unique except for its unit, is shown to exist very generally. It appears to be the formal counterpart of the intuitive idea of utility (or value) in economics, of incentive value in motivation, of subjective sensation in psychophysics, and of response strength in learning theory.

3. Well-Defined Sets of Alternatives

So far, there seems to have been an implicit assumption that no difficulty is encountered in deciding among what it is that an organism makes its choices. Actually, in practice, it is extremely difficult to know, and much experimental technique is devoted to arranging matters so that the organism and the experimenter are (thought to be) in agreement about what the alternatives are. All of our procedures for data collection and analysis require the experimenter to make explicit decisions about whether a certain action did or did not occur, and all of our choice theories—including this one—begin with the assumption that we have a mathematically well-defined set, the elements of which can be identified with the choice alternatives. How these sets come to be defined for organisms, how they may or may not change with experience, how to detect such changes, etc., are questions that have received but little illumination so far. There are limited experimental results on these topics, but nothing like a coherent theory. Indeed, the whole problem still seems to be floundering at a conceptual level, with us hardly able to talk about it much less to know what experiments to perform.

More than any other single thing, in my opinion, this Achilles' heel has limited the applicability of current theories of choice: it certainly has been a significant stumbling block in the use of information theory in psychology, it has limited learning theory applications to a rather special class of phenomena typified by T-maze experiments, etc. The present theory is no different in this respect from the others.

B. PROBABILITY AXIOMS

Throughout the book we shall suppose that a universal set U is given which is to be interpreted as the universe of possible alternatives (stimuli or responses). In practice U will have to possess a certain homogeneity: the decision maker will have to be able to evaluate the elements of U according to some comparative dimension and to be able to select from certain finite subsets of U the elements that he thinks are superior (or inferior or distinguished in some way) along that dimension. For example, in economics U may be taken to be a set of commodity bundles among which a person can express preferences; in psychophysics it may be the set of possible sound energies (at a fixed frequency) which a subject can be asked to evaluate according to loudness; or in learning theory U may be the set of alternative responses available to the organism. Note that U may be finite or infinite.

In general, a subject is not asked to make a choice from the whole of U but rather from some (small) finite subsets. In a great many experiments only two alternatives are presented to the subject at a time, and he is required to choose the one he prefers or the one he deems louder, etc. Of course, larger subsets could be used, although for the most part they have not been, and certainly most daily decisions are from larger subsets (e.g., the choice of a meal from a menu or the choice among several jobs, etc.).

Let T be a finite subset[1] of U and suppose that an element must be chosen from T. If x is an element of T (written $x \in T$), let $P_T(x)$ denote the probability that the selected element is x. Slightly more generally, if S is a subset of T (written $S \subset T$), let $P_T(S)$ denote the probability that the selected element lies in the subset S. These probabilities are the basic ingredients of the following theory.

[1] The restriction to finite subsets is not basic, but for most purposes it does not restrict the applicability of the theory (see, however, Appendix 2), and it introduces considerable simplicity.

In most choice models we would write $P(x)$ for $P_T(x)$ because the choice set T is held invariant throughout the discussion; in fact, we would let T and U be the same set. Here, however, several different choice sets are to be considered at once. Let us suppose that we are working with 1000 cps tones at different intensities measured in db above some reference level; let w, x, y, and z denote, respectively, the 50, 52, 54, and 56 db tones. Let $T = \{w, x, y\}$ and $T' = \{x, y, z\}$ and consider choices according to loudness. There is assumed to be some probability, denoted by $P_T(x)$, that x, the 52-db tone, will be called loudest when T is presented, and another, generally different, probability $P_{T'}(x)$ that x will be called loudest when T' is presented. There is no reason to expect these probabilities to be the same, and the purpose of the subscripts is to make the several probabilities identifiable.

It must not be forgotten, however, that all of the probabilities having the same subscript T form an ordinary probability measure on the subsets of T. This means, explicitly, that the following is assumed:

The ordinary probability axioms.

(i) *For* $S \subset T$, $0 \leqq P_T(S) \leqq 1$.
(ii) $P_T(T) = 1$.
(iii) *If* $R, S \subset T$ *and* $R \cap S = \phi$, *then* $P_T(R \cup S) = P_T(R) + P_T(S)$.

Repeated application of part iii implies that

$$P_T(S) = \sum_{x \in S} P_T(x);$$

therefore, it is always sufficient to state results just for $P_T(x)$.

Note that, given our interpretation of these probabilities, part ii means that the subject is forced to make a choice: the probability is 1 that his choice is in T when he must confine his choice to T.

For simplicity of notation, and to conform to standard usage, $P(x, y)$ is written to stand for $P_{\{x,y\}}(x)$ when $x \neq y$. It will be convenient to introduce the symbol $P(x, x) = \frac{1}{2}$ so that certain equations (e.g., $P(x, y) + P(y, x) = 1$) can be written without any restriction on the values assumed by x and y.

C. CHOICE AXIOM

1. Statement of Axiom

The axioms of ordinary probability theory establish certain restraints upon each of the measures P_T, but no connections are assumed among the several measures. However, one suspects that, at least for choice behavior, the several measures cannot be completely independent. The relationship we shall investigate can be stated as follows:

Axiom 1. *Let T be a finite subset of U such that, for every $S \subset T$, P_S is defined.*

(i) *If $P(x, y) \neq 0, 1$ for all $x, y \in T$, then for $R \subset S \subset T$*
$$P_T(R) = P_S(R)P_T(S);$$
(ii) *If $P(x, y) = 0$ for some $x, y \in T$, then for every $S \subset T$*
$$P_T(S) = P_{T-\{x\}}(S - \{x\}).$$

Throughout the book the expression "axiom 1 holds for the set T" is used to mean not only that it holds for T itself but also that it holds for every subset of T.

2. Discussion

There are a number of points, both technical and conceptual, that should be made about the axiom.

a. Interpretation. Part ii of the axiom simply states that if y is invariably chosen over x then x may be deleted from T when considering choices from T. This seems reasonable. If one never selects liver in preference to roast beef, then in choosing among liver, roast beef, and chicken one can immediately reduce the problem to consideration of roast beef and chicken.

Lemma 1. *If axiom 1 holds for T and if $P(x, y) = 0$ for some $y \in T$, then $P_T(x) = 0$.*

PROOF. For $z \in T$, $z \neq x$, part ii of axiom 1 implies
$$P_T(z) = P_{T-\{x\}}(z).$$
By parts ii and iii of the probability axioms,
$$\begin{aligned} 1 &= P_T(x) + \sum_{z \in T-\{x\}} P_T(z) \\ &= P_T(x) + \sum_{z \in T-\{x\}} P_{T-\{x\}}(z) \\ &= P_T(x) + 1, \end{aligned}$$
and the result follows.

By repeated applications of part ii of axiom 1, the choice set can be reduced to one in which only cases of imperfect discrimination ($P(x, y) \neq 0$ or 1) occur, and then part i becomes applicable. So let us consider that part.

To deal with complicated decisions, it is usual to subdivide them into two or more stages: the alternatives are grossly categorized in some fashion and a first decision is made among these categories; the one chosen is further categorized and a second decision is made, etc. It is commonly accepted, and it is probably true, that when such a multistage process is needed the over-all result depends significantly upon which intermediate partitionings are employed. One senses, however, that if the decision situation is quite simple—so that a two-stage process is not really needed—then the intermediate categorization, if used, will not matter. That is to say, the product $P_S(R)P_T(S)$ will not depend upon S. But, by taking $S = T$, we see that this product must be $P_T(R)$, which is part i of axiom 1.

These remarks make it clear that we cannot expect the axiom to be valid except for simple decisions, but this is no real limitation, since, as we shall see, our results really require only that it be correct for sets of three alternatives. The question of the range of validity of the axiom is raised again in section 5.B.

The axiom may be viewed in another way provided conditional probability is defined in the usual manner, i.e., if $P_T(S) > 0$, then
$$P_T(R|S) = \frac{P_T(R \cap S)}{P_T(S)}.$$

Lemma 2. *If $P(x, y) \neq 0, 1$ for all $x, y \in T$, then axiom 1 is equivalent to $P_S(R) = P_T(R|S)$, for $R \subset S \subset T$.*

PROOF. The result is obvious except for the condition $P_T(S) > 0$. It is clearly sufficient to show $P_T(x) > 0$ for all $x \in T$. Suppose this were not true for some x, then for any $y \in T$, $y \neq x$, axiom 1.i (p. 6) implies
$$\begin{aligned} 0 &= P_T(x) \\ &= P(x, y)[P_T(x) + P_T(y)] \\ &= P(x, y)P_T(y). \end{aligned}$$
Since $P(x, y) > 0$, it follows that $P_T(y) = 0$, and so $\sum_{y \in T} P_T(y) = 0$, which is impossible by the probability axioms.

Ignoring cases of perfect discrimination, this lemma says that the axiom requires that the measure P_S be identical to the conditional measure induced by P_T. As a concrete example, suppose that T is the set of entrees on a certain menu, S is some proper subset of T that includes roast beef, and R the single element set of roast beef. The heart of the axiom is the assumption that when, for whatever reason, the restaurant has only the entrees S the probability of selecting roast beef is the same as the conditional probability of selecting it from S when the whole menu is available.

When first examining part i of the axiom, some have felt that it is tautological; however, the foregoing example should make it clear that a substantive assumption is involved. This can be checked formally by writing out the sample space involved—it will not be done here—or, less formally, by just observing that two distinct experiments are required to verify the axiom. In one T is offered to the subject and P_T is estimated; in the other S is offered and P_S is estimated.

It has been implicit in the discussion, and is explicit in the title of the book, that this theory—axiom 1 in particular—applies to single organisms, not to averages over groups of them. It is not difficult to see that every organism in a group could satisfy the axiom, yet the average probabilities violate it, and vice versa. For example, consider two organisms, 1 and 2, with probabilities

$$P_T^{(1)}(R) = 0.72 \qquad P_S^{(1)}(R) = 0.80 \qquad P_T^{(1)}(S) = 0.90$$

$$P_T^{(2)}(R) = 0.02 \qquad P_S^{(2)}(R) = 0.20 \qquad P_T^{(2)}(S) = 0.10,$$

which satisfy axiom 1 individually. The group averages are 0.37, 0.50, and 0.50, which fail to satisfy the axiom, since $(0.50)(0.50) = 0.25 \neq 0.37$. This does not mean that group studies can never be used in connection with this theory, but they must be chosen with care so as not to do violence to the basic ideas.

b. An alternative axiom. As originally formulated in an unpublished manuscript, the second part of axiom 1 was not given; it was assumed that part i held without restriction. Several examples which are discussed in section 1.D.2 indicate that this is not reasonable. A simple calculation now will suffice to illustrate the difficulty. Suppose that part i held without restriction, that $P(x, y) = 0$, and that $P(x, z) > 0$. Let $T = \{x, y, z\}$. We would then have

$$P_T(x) = P(x, y)P_T(\{x, y\}) = 0,$$

and

$$P_T(x) = P(x, z)P_T(\{x, z\}) = P(x, z)[P_T(x) + P_T(z)].$$

Since $P_T(x) = 0$ and $P(x, z) > 0$, it follows that $P_T(z) = 0$. That is to say, unrestricted application of part i means that if y is always preferred to x and if, however infrequently, x is sometimes preferred to z then z is never chosen from the set of the three. Intuitively, this does not seem correct.

c. Independence from irrelevant alternatives.

Lemma 3. *If $P(x, y) \neq 0, 1$ for all $x, y \in T$, then axiom 1 implies that for any $S \subset T$ such that $x, y \in S$,*

$$\frac{P(x, y)}{P(y, x)} = \frac{P_S(x)}{P_S(y)}.$$

PROOF. By the axiom, we know

$$P_S(x) = P(x, y)[P_S(x) + P_S(y)],$$

so

$$P_S(x)[1 - P(x, y)] = P_S(x)P(y, x) = P(x, y)P_S(y).$$

From the proof of lemma 2 we know that none of the probabilities is 0, so cross-dividing gives the result.

The essential fact contained in lemma 3 is that when axiom 1 holds for T and its subsets the ratio $P_S(x)/P_S(y)$ is independent of S.

In decision theory (see, for example, Luce and Raiffa [1957]) one axiomatic idea, which may be termed "independence from irrelevant alternatives," is recurrent. The idea was brought to the fore by Arrow [1951] in a particular choice context, but the same basic notion appears in other contexts in which, of course, its axiomatic formulation differs somewhat. Arrow termed his axiomatization of the idea "independence of irrelevant alternatives," but, as Professor S. S. Stevens has pointed out to me, this phrase is unfortunately misleading, since it suggests that the irrelevant alternatives are independent of one another. The actual gist of the idea is that alternatives which *should be* irrelevant to the choice are in fact irrelevant, hence the present term. For example, the idea states that if one is comparing two alternatives according to some algebraic criterion, say preference, this comparison should be unaffected by the addition of new alternatives or the subtraction of old ones (different from the two under consideration). Exactly what should be taken to be the probabilistic analogue of this idea is not perfectly clear, but one reasonable possibility is the requirement that the ratio of the probability of choosing one alternative to the probability of choosing the other should not depend upon the total set of alternatives available, i.e., the assertion of lemma 3. In this sense, then, we can say that axiom 1 is *a* probabilistic version of the independence-from-irrelevant-alternatives idea.

It should be noted that it is only the ratio of the two probabilities, not the probabilities themselves, that is invariant with changes of the irrelevant alternatives; thus axiom 1 is not clearly at variance with *Gestalt* ideas, as it might first seem.

d. Transitivity. In choice work in which discrimination is assumed to be perfect it has been customary to assume that pairwise choices are transitive. It would be unfortunate if axiom 1 were at variance with this assumption; it is not.

Lemma 4. *If axiom* 1 *holds for* $T = \{x, y, z\}$ *and if* $P(x, y) = 1$ *and* $P(y, z) = 1$, *then* $P(x, z) = 1$.

PROOF. Since both $P(y, x) = 0$ and $P(z, y) = 0$, part ii of axiom 1 implies that $P_T(x) = P(x, z) = P(x, y)$, but, by assumption, $P(x, y) = 1$, hence the assertion.

Thus axiom 1 is a probabilistic version of two of the more important axioms in nonprobabilistic choice theory: independence from irrelevant alternatives and transitivity.

e. Alternative formulations of part i. Other ways of stating part i of axiom 1 are possible, but, as they seem to shed but little light on its meaning and they are not needed in the sequel, they have been relegated to Appendix 1.

3. Previous Work

So far as is known, no one has proposed and investigated an axiom exactly equivalent to axiom 1; however, in several places part i of the axiom has arisen.

a. Conditional probability theory. After the main ideas reported here were developed, Professor Patrick Suppes called to my attention papers by Császár [1955] and Rényi [1955] that are closely related to this work. Their problem was to axiomatize conditional probability. Without going into any of the niceties of probability theory, let me sketch their main idea. Suppose, first, that a probability measure p is given on a suitable class of subsets of a set U, then for the subsets S and T such that $p(T) > 0$ the conditional probability of S given T is defined as

$$p(S|T) = \frac{p(S \cap T)}{p(T)}.$$

Now, if $R \subset S \subset T$, then

$$\begin{aligned} p(R|S)p(S|T) &= \frac{p(R \cap S)}{p(S)} \frac{p(S \cap T)}{p(T)} \\ &= \frac{p(R)}{p(S)} \frac{p(S)}{p(T)} \\ &= \frac{p(R \cap T)}{p(T)} \\ &= p(R|T). \end{aligned}$$

This is, of course, the formal analogue of part i of axiom 1. By taking three arbitrary sets, instead of $R \subset S \subset T$, a somewhat more general condition can be shown to hold. They take this more general property as an axiom for conditional probability when no unconditional probability measure is given. In the traditionally general manner of abstract probability theory they establish the existence of a measure function such that their given probabilities are conditional measures relative to it. Because of certain empirically reasonable restrictions, a much simpler proof of this same result can be given (see theorem 3 below). The interpretation and use made of this theorem is considerably different from Rényi and Császár's work.

b. Axiomatic characterization of entropy in information theory. Shannon [1949], in his theory of information, has dealt with certain average properties of choices that are made from a finite set T of alternatives subject to a probability distribution P_T. A statistic of central importance in his theory—he called it the entropy of the distribution and others have called it the average amount of information transmitted—is

$$H = -\sum_{x \in T} P_T(x) \log_2 P_T(x).$$

Two a priori arguments for using this statistic have been given. One of these, due to Shannon [1949] and Fano [1949], considers recodings of the messages emanating from the source into "economical" strings of binary digits and shows that in the limit H binary digits are needed on the average for each selection from the source. This justification may be appropriate when information theory is applied to questions of language and coding, but it does not seem particularly relevant to most of the other uses of information theory in psychology.

The second justification, also due to Shannon, is axiomatic in nature, and it seems to have a reasonable interpretation in many nonlinguistic contexts (e.g., when measuring the amount of information that a subject can transmit about a display of lights). The most important of Shannon's axioms is the third one in which he assumes that the entropy of a distribution P_T, where T is finite, can always be expressed as the sum of two quantities:

(i) The (unknown) entropy of the distribution that results from the given distribution on T by treating an arbitrary subset S of T as a single element occurring with probability

$$P_T(S) = \sum_{x \in S} P_T(x),$$

plus

(ii) $P_T(S)$ times the (unknown) entropy of the distribution $P_T(x)/P_T(S)$ over the set S.

In other words, Shannon assumes that entropy can be decomposed in a nicely additive manner, using as the distribution over a subset S of T the one naturally induced by P_T. However, if we choose to apply information theory to behavior, as has been done, we must acknowledge that this induced distribution is not necessarily the one actually governing behavior when S rather than T is presented. Therefore, we are only really justified in applying that theory to problems of behavior if we are willing either to accept the recoding justification of the statistic H or to assume that

$$P_S(x) = \frac{P_T(x)}{P_T(S)},$$

which, of course, is part i of axiom 1.

This means that whenever the entropy statistic is used to describe animal or human behavior for which the recoding argument is inapplicable either Shannon's axiomatic defense of the statistic is implicitly rejected or axiom 1 is implicitly assumed. If the latter is true, then information theory implicitly presupposes the consequences of axiom 1, which are relatively strong—specifically, when discrimination is imperfect, it means that choice behavior can be scaled by a ratio scale. Many have believed that information theory could be applied with little regard to the laws satisfied by the organism making the choices, but this seems to be an error.

c. Constant-ratio rule for confusion matrices. Clarke [1957] reports studies in which subjects listened to sounds (monosyllables, digits, etc.) drawn from known finite sets of possible sounds but heavily masked by noise. If the noise level is appropriate, a considerable number of errors of identification occur which can be summarized by a square matrix $[P_{ij}]$ of the probabilities of confusion. P_{ij} is the probability that the subject reports sound j when i is actually transmitted. Clarke raises the question: if we know this matrix for a given set T, can we predict the one that will arise when a subset S of T is studied? He proposes using part i of axiom 1, which he has called the "constant-ratio rule" because of the property described in lemma 3. Although he does not explore the implications of his assumption, he does present the only direct empirical test of axiom 1 that has so far been published. His results are discussed presently.

4. Direct Empirical Testing of Axiom 1

a. The statistical problem. A discussion of the conditions under which axiom 1 may be expected to hold and something of the role that it might play in the study of choice behavior will not be taken up until a number of its consequences are known (see section 5.B). Since, however, it is clear that, at least in principle, choice data can be collected in specific situations to determine whether the axiom should be rejected there, it is worth considering a few of the statistical issues.

There are various forms in which part i of axiom 1 might be tested, but lemma 3 appears to lead to the simplest results. As shown in theorem 4, the axiom need only be assumed to hold for sets of three elements, so the hypothesis to be tested is that

$$\frac{P(x, y)}{P(y, x)} = \frac{P_{\{x,y,z\}}(x)}{P_{\{x,y,z\}}(y)}.$$

The problem is to get some idea of the number of observations that are needed to have anything like a sensitive test of this hypothesis. An (approximate) expression must be derived for the variance of estimates of these ratios. We know, of course, that if n independent Bernoulli trials are used to estimate each of the basic probabilities p for a single subject, then their variance is $\sigma_p^2 = pq/n$.

Suppose that $f(X, Y)$ is a function that can be expanded in a power series about the two variables X and Y, where X and Y are statistics having means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Of course, we will take $f(X, Y) = X/Y$. Using a linear approximation to f, we have

$$f(X, Y) \approx f(\mu_X, \mu_Y) + (X - \mu_X)f_X(\mu_X, \mu_Y) + (Y - \mu_Y)f_Y(\mu_X, \mu_Y),$$

where $f_X = \dfrac{\partial f}{\partial X}$ and $f_Y = \dfrac{\partial f}{\partial Y}$. Thus,

$$\mu_f = E[f(X, Y)] \approx f(\mu_X, \mu_Y)$$

and

$$\begin{aligned}\sigma_f^2 &= E\{[f(X, Y) - \mu_f]^2\}\\ &\approx \sigma_X^2 f_X(\mu_X, \mu_Y)^2 + \sigma_Y^2 f_Y(\mu_X, \mu_Y)^2 + 2\rho_{XY}\sigma_X\sigma_Y f_X(\mu_X, \mu_Y)f_Y(\mu_X, \mu_Y),\end{aligned}$$

where ρ_{XY} is the correlation between X and Y. A rigorous discussion of this result can be found in Cramér [1946], p. 353.

For our case, $f(X, Y) = X/Y$, so

$$f_X = 1/Y, \quad f_Y = -X/Y^2, \quad \text{and} \quad \rho_{XY} = -\mu_X\mu_Y/n\sigma_X\sigma_Y.$$

Substituting, we find

$$\begin{aligned}\mu_f &= \mu_X/\mu_Y\\ \sigma_f^2 &= (\mu_X/\mu_Y)^2[2 + (1 - \mu_X)/\mu_X + (1 - \mu_Y)/\mu_Y]/n.\end{aligned}$$

If, as in the left side of the equation that we want to test, $\mu_X = 1 - \mu_Y$, then σ_f^2 reduces to $\mu_X/n[1 - \mu_X]^3$.

To gain an idea of the sample sizes needed for the two-alternative case, consider the demand that the standard deviation be some fixed proportion k of the expected ratio, i.e.,

$$\sigma_f = k\mu_X/(1 - \mu_X),$$

then

$$n = 1/k^2\mu_X(1 - \mu_X).$$

The sample sizes for several values of k and μ_X are presented in Table 1; it is clear that rather large sample sizes are required from each subset to obtain reasonably sensitive direct tests of axiom 1.

TABLE 1. Sample Size n as a Function of μ_X for k = 0.10 and 0.05. See Text for Explanation of Symbols

k \ μ_X	0.1	0.2	0.3	0.4	0.5
0.10	1110	625	475	417	400
0.05	4450	2500	1900	1670	1600

b. Clarke's data. As mentioned earlier, the only published data which directly test axiom 1 are Clarke's [1957]. He used the average results from several subjects to estimate the probabilities. As he points out, this is appropriate only to the extent that the subject's probabilities have the same values (see section 1.C.2). Although separate estimates for each subject suggested that their probabilities were similar in this experiment, averaging undoubtedly increased the variance of his results. From data on confusion matrices of one size he predicted the results for smaller confusion matrices. His sample sizes (over subjects and repetitions) ranged from 400 to 1100 per stimulus in different situations. No statistical analysis of the data was presented, but his four scatter diagrams of predicted vs. observed results exhibit a nice linear relation, with apparently no systematic deviations from the 45-degree line. For example, in his first experiment he employed three master sets of six consonants followed by a vowel plus two subsets of three consonants from each master set. Each consonant-vowel pair in a master list was presented 150 times to a subject, and each pair from a three-element subset was presented 200 times. Four subjects and one talker were used. Using axiom 1 (constant-ratio rule), predicted values for the subsets were made from the data on the master sets. The scatter diagram is shown in Figure 1. The results from the other three experiments are similar, with possibly less variance. So, as a first approximation at least, axiom 1 seems to hold in an articulation context.

c. Time- and space-order errors. To test axiom 1 directly, the most reasonable studies appear to be psychophysical. Not only is it feasible to get the sample sizes needed, but the experimental techniques and controls are better worked out there than in other areas. There is only one problem: stimuli must be displayed successively either in time or in space, and on many continua there are corresponding time- or space-order errors. For example, let x and y be auditory stimuli that are to be judged according to loudness. If one presents x and then y and asks which is louder, the probability that x will be chosen is generally smaller than when the order

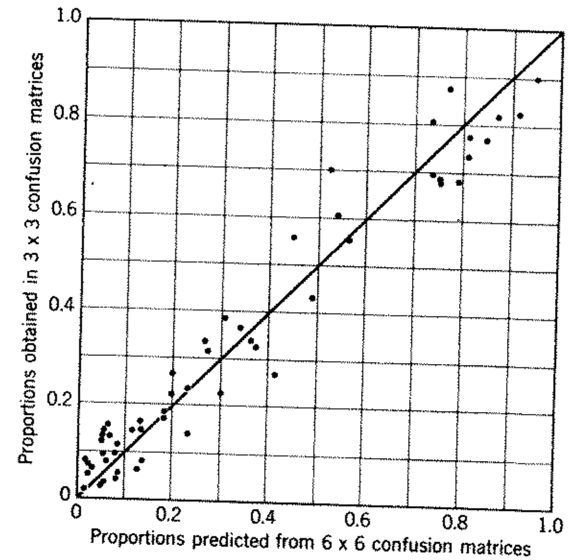

Figure 1. Scatter diagram of observed proportions of choices in 3 x 3 confusion matrices vs. the predicted proportions obtained by axiom 1 (constant-ratio rule) from the observed proportions in a 6 x 6 confusion matrix. (Adapted, with permission, from Figure 1, p. 718, Clarke [1957].)

of presentation is reversed. Space errors are similar but more complex. Neither phenomenon is well understood, and no techniques are known for eliminating them. Since axiom 1 is stated in terms of unordered sets, it is not immediately clear how one can possibly test it when ordering matters. This is not an idle problem, since the effects are large when compared with the deviations we should like to detect if axiom 1 is false, and it appears to suggest that we have omitted a basic phenomenon from our theory. Fortunately, this appearance is misleading, and we are able to encompass these effects in the present theory. The analysis must be postponed to section 1.F.

D. TWO CONSEQUENCES

1. Statement

The first theorem to be proved establishes formally that if axiom 1 holds all the probabilities are determined by the pairwise probabilities. It is clear that by repeated applications of part ii of axiom 1 we lose no generality in confining our attention to cases in which no discriminations are perfect.

Theorem 1. *If axiom 1 holds for T and if $P(x, y) \neq 0, 1$ for all $x, y \in T$, then*

$$P_T(x) = \frac{1}{\sum_{y \in T} \frac{P(y, x)}{P(x, y)}} = \frac{1}{1 + \sum_{y \in T - \{x\}} \frac{P(y, x)}{P(x, y)}}.$$

PROOF. By lemma 3,

$$\sum_{y \in T} \frac{P(y, x)}{P(x, y)} = 1 + \sum_{y \in T - \{x\}} \frac{P(y, x)}{P(x, y)}$$
$$= \frac{P_T(x)}{P_T(x)} + \sum_{y \in T - \{x\}} \frac{P_T(y)}{P_T(x)}$$
$$= \frac{1}{P_T(x)} \sum_{y \in T} P_T(y).$$

But, by parts ii and iii of the probability axioms,

$$\sum_{y \in T} P_T(y) = 1,$$

hence the assertion.

The next theorem shows that axiom 1 also demands that certain constraints be met by the pairwise probabilities.

Theorem 2. *If axiom 1 holds for $\{x, y, z\}$ and if none of the pairwise discriminations is perfect, then*

$$P(x, y)P(y, z)P(z, x) = P(x, z)P(z, y)P(y, x).$$

PROOF. Observe that if $T = \{x, y, z\}$,

$$\frac{P_T(x)}{P_T(y)} \frac{P_T(y)}{P_T(z)} \frac{P_T(z)}{P_T(x)} = 1.$$

Thus, by lemma 3,

$$\frac{P(x, y)}{P(y, x)} \frac{P(y, z)}{P(z, y)} \frac{P(z, x)}{P(x, z)} = 1.$$

Corollary. *Under the conditions of the theorem,*

$$P(x, z) = \frac{P(x, y)P(y, z)}{P(x, y)P(y, z) + P(z, y)P(y, x)}.$$

PROOF. Substitute $P(z, x) = 1 - P(x, z)$ in the theorem and solve for $P(x, z)$.

2. Discussion

The major significance of these two theorems will become apparent in what follows; however, one comment upon the second is in order. If each of the pairs from $\{x, y, z\}$ is offered to a subject just once, if his choices are governed by the given probabilities, and if they are statistically independent, then $P(x, y)P(y, z)P(z, x)$ is the probability that his reports form the intransitivity $x > y > z > x$. The second theorem asserts that if axiom 1 holds this probability must be the same as the probability of the reverse intransitivity, namely, $x > z > y > x$.

The primary reason for stating axiom 1 in the form given can now be presented. If it is assumed that part i holds whether or not any pairwise discriminations are perfect, then it is possible to show that theorem 1 also holds without any restrictions. The only problem in doing this is to handle divisions by 0, but with a little care the theorem can be shown to hold. Now, suppose $P(x, z) = 1$ and consider any $y \in U$ such that the modified axiom holds for $T = \{x, y, z\}$. We show, then, that either $P(x, y) = 1$ or $P(y, z) = 1$. Since $P(x, z) = 1$, theorem 1 gives

$$P_T(x) = P(x, y); \qquad P_T(y) = \frac{1}{1 + \frac{P(x, y)}{P(y, x)} + \frac{P(z, y)}{P(y, z)}}; \qquad P_T(z) = 0.$$

The sum of these three quantities must be 1, which by simple algebra leads to the condition

$$\frac{P(z, y)}{P(y, z)} P(y, x) = 0,$$

provided that $P(x, y) < 1$ is assumed. Therefore, $P(y, z) = 1$.

The essential point, then, is that if we uninhibitedly assume part i of axiom 1 we find that a single case of perfect pairwise discrimination implies that any third alternative is perfectly discriminated relative to at least one of the two original alternatives. To those familiar with empirical data, such a result will not seem reasonable. It is difficult to think of, say, a psychophysical continuum in which, by sufficient subdivision, it is not possible ultimately to produce imperfect pairwise discriminations.

One of two tacks can be taken: either modify the axiom to take care of perfect discrimination explicitly or attempt to argue that there are really no cases of perfect discrimination, only apparent ones. The latter argument can proceed as follows. Suppose that, with the psychophysicists, we say that alternative y is one just noticeable difference (jnd) "larger" than alternative x when $P(y, x) = \frac{3}{4}$. Similarly, z is one jnd larger than y, and so two jnds larger than x, when $P(z, y) = \frac{3}{4}$. And so on. Consider, now, alternatives a and b where a is n jnds larger than b. If we assume that all discriminations are imperfect, then by repeatedly applying the corollary of theorem 2 to the alternatives spaced at one-jnd intervals between a and b it is easy to show that $P(a, b) = 1/[1 + (\frac{1}{3})^n]$. So, for $n = 2$ there is one chance in ten of describing b as larger than a; for $n = 5$ it is already one chance in 244; and for $n = 10$ it is one in 59,050. Since laboratory estimates of such probabilities are rarely based upon samples larger than several hundred observations, it is highly likely that separations of more than three or four jnds will appear to be perfectly discriminated, even if mathematically they are not. Indeed, such rare events are not likely to be recorded when observed, the unexpected reversal being attributed to the apparatus or to the experimenter rather than to the subject.

Although all of this has a ring of plausibility, it is far from certain. There are examples, two of which are given below, in which it seems reasonable to suppose that mixed perfect and imperfect discriminations occur. More important, it is possible in some problems to show that certain reasonable conditions require, as a mathematical matter, that some discriminations be perfect; an example of this may be found in section 3.B. Here we shall be content with the examples. The first of these, except for minor modifications, is due entirely to Professor John Chipman.[2] Suppose that we have three urns each containing 100 balls, colored black and white. An event will be the occurrence of a black ball in a random selection from a specified urn, and we will suppose that a subject chooses between urns on the basis of his judgment of the "likelihood" of the events occurring. He has the following information:

Urn	Information
α	a random sample (replaced) of 10 balls yielded 6 black ones
β	contains 55 black balls
γ	contains 65 black balls

[2] Personal communication.

Because of the different numbers of black balls in these urns it seems plausible that for some subjects $P(\gamma, \beta) = 1$. Equally well, since all that is known about urn α is based upon a single sample of ten, a person is well justified in fearing that it has fewer black balls than β as well as hoping that it has more than γ. If so, we may well find that both

$$P(\gamma, \alpha) < 1 \quad \text{and} \quad P(\alpha, \beta) < 1,$$

which violates the above conclusion, thus casting doubt upon the unrestricted application of part i of axiom 1.

The second example was suggested by Professor William Vickrey.[3] Consider commodity bundles, each of which consists of two components, x and y, which are both perfectly ordered by preference (denoted by $>$). Suppose $x > x'$ and $y > y'$, then it is plausible that $P[(x, y), (x', y')] = 1$. Now choose (x'', y'') such that $x'' > x, x'$ and $y'' < y, y'$ then, at least for some choices, it is plausible that the resulting conflict leads to

$$P[(x, y), (x'', y'')] < 1 \quad \text{and} \quad P[(x'', y''), (x', y')] < 1.$$

Again, if this can actually happen, part i of axiom 1 cannot be assumed when pairwise discriminations are perfect.

3. Coombs' Data

In addition to direct tests of axiom 1, a number of indirect ones are also possible. The first of these arises from the property known as "monotonicity" (Coombs [1958]) or "strong stochastic transitivity" (Davidson and Marschak [1957]), which follows from the corollary to theorem 2. The property is that if $\frac{1}{2} \leqslant P(x, y) < 1$ and $\frac{1}{2} \leqslant P(y, z) < 1$, then $P(x, z) \geqslant P(x, y)$ and $P(x, z) \geqslant P(y, z)$. It is clearly met if $P(x, z) = 1$, so we assume $P(x, z) < 1$. By the corollary to theorem 2,

$$P(x, z) = \frac{P(x, y)P(y, z)}{P(x, y)P(y, z) + P(y, x)P(z, y)}$$

$$= \frac{P(x, y)}{P(x, y) + P(y, x)P(z, y)/P(y, z)}.$$

But since $P(y, z) \geqq \frac{1}{2}$, $P(z, y)/P(y, z) \leqq 1$, and so

$$P(x, z) \geqq \frac{P(x, y)}{P(x, y) + P(y, x)}$$

$$= P(x, y).$$

In a similar manner $P(x, z) \geqq P(y, z)$.

[3] At the September 1957 meetings of the Econometric Society.

Coombs [1958] presents preference (for shades of gray) data that appear, at first glance, to reject strong stochastic transitivity: of 120 triples $\{x, y, z\}$ of stimuli, the four subjects exhibited 19, 26, 31, and 58 violations, respectively. There are, however, two questions of interpretation that must be raised. First, the proportions actually compared are, of course, only estimates of the underlying probabilities, and thus, even if the probabilities satisfy strong stochastic transitivity, not all of the proportions can be expected to satisfy it, especially not if, for example, $P(x, y)$ is only slightly less than $P(x, z)$. Fortunately, Coombs reports the proportions, thus making it possible to estimate which violations are significant. A sufficient number seem to be, so this will not explain away his results. Second, the data actually collected were not paired comparisions but the subject's rankings of subsets of four stimuli. The probabilities $P(x, y)$ were then estimated by the number of sets in which x was ranked above y, divided by the total number of sets in which x and y both appeared. As is discussed more fully in section 2.F.2, this proportion need not necessarily be an estimate of $P(x, y)$. For some models that relate the ranking probabilities to the choice probabilities, it is; for others, it is not. It is, therefore, not entirely clear whether or not strong stochastic transitivity has been tested.

Although these observations cast some doubt upon the importance that should be attached to this study, one feature of Coomb's work somewhat tends to undercut these doubts. He presents an a priori argument which leads to the prediction that most of the violations of strong stochastic transitivity should lie within a particular class of triples, and this is rather well sustained by his data. It is clear that the study should be repeated in some fashion using paired comparisions.

E. RATIO SCALE

1. Background

As the study of choice behavior has developed, both in psychology and in economics, one of the central issues that a formal characterization must face are conditions that ensure the existence of a relatively unique numerical scale which in some sense represents the choice behavior of the subjects. Mathematically, the problem is simply one of imposing sufficient axiomatic structure to prove the existence of a scale that is unique up to some group of transformations—the group of positive linear transformations (zero and unit unspecified) has usually been deemed to be just acceptable. These are what Stevens [1951] terms interval scales. But the empirical side-condition that these mathematical assumptions must form a more or less plausible description of human and animal choice

behavior has rendered the problem difficult. There appear to have been three main approaches.

a. Economics. Preference among bundles of goods has been taken to be the underlying primitive in economics, and, as an idealization, it has been assumed to be an algebraic ordering of the commodity bundles. In such models, if any numerical order preserving scale exists, many do. In fact, they are unique only up to monotonic transformations, which renders the numerical character of the scales almost superfluous. That being so, some economists arrived at the position that it is safer to work only with orderings—as they say, with ordinal utilities in contrast to cardinal[4] ones—and for many of the traditional theorems of economics this is sufficient. Nonetheless, some work, particularly in modern decision theory, requires cardinal utility scales. Some extension of the traditional formulation was needed, and a little more than a decade ago it was affected by von Neumann and Morgenstern [1947]. (Actually, Ramsey [1931] suggested some of the same ideas a good deal earlier, but the importance of his work was not recognized until recently.) Roughly, they continue to suppose that preferences are algebraic, but the domain of choice is extended from a set of "pure alternatives" to the set of all possible gambles that can be generated from the alternatives and an infinite set of chance events. Preference over these gambles is assumed to meet certain fairly restrictive axioms which, although normatively compelling, seem at best to lack detailed descriptive realism. Under these conditions, a scale is shown to exist which is unique up to positive linear transformations and which has the important property that the utility of a gamble is equal to the expected utility of its components.

b. Psychophysics. The psychologist has been largely unwilling to make the economist's algebraic idealization, for in some measure the substance of his problem resides in the fact that people are unable to make consistent discriminations. The early psychophysicists proposed to use these data as a means of scaling subjective sensation. Ultimately, this question is discussed more fully, mainly because recent workers have tended to reject the earlier ideas, but here it suffices to mention the fact that the attempt was made and that analytical methods were presented to calculate an interval scale whenever certain consistencies are exhibited by the data. Mathematically, the uniqueness of these scales results in large part from the assumption that the set being scaled is a continuum—a reasonable assumption for such dimensions as sound energy, weight, length, etc. For a modern discussion of this mathematics, see Luce and Edwards [1958].

[4] Numerical.

c. Psychometrics. In the remainder of psychology a small group of workers, often referred to as psychometricians, have been concerned with scaling objects other than the traditional sensory stimuli. In particular, such concepts as attitude, preference, intelligence, and interest have concerned them. Their problem has in some ways been similar to that confronted by the economists in that scales with appropriate uniqueness properties are hard to come by. The continuous approximation of the psychophysicist was out, and the gambles of the utility theorist—which, in any event, are of dubious realism in many psychological contexts—were not thought of. The resolution arrived at during the second and third decades of this century, largely through the efforts of Thurstone and his students, was roughly this. The, by then, somewhat tarnished psychophysical assumption was taken over that the underlying scale has the property that discrimination between two objects depends upon the numerical difference of their scale values. Since the continuum assumption could not be transferred, this was quite insufficient to lead to a unique scale. Other assumptions had to be added. At the time, statistics was rapidly becoming the somewhat overworked handmaiden of psychology, and normality and independence assumptions were in the wind. With little real justification beyond convenience and need, these were freely introduced until finally adequate uniqueness was achieved. The result: an extensive literature that has been largely ignored by outsiders, who have been uneasy over the strong and none too compelling assumptions employed.

It is true, as Adams and Messick [1957] have recently re-emphasized and spelled out in detail, that the Thurstonian assumptions do lead to testable restrictions on the observables. Nonetheless, it does seem odd first to postulate this rather complex, normally distributed, but unobservable subworld and only then to determine the relations among observables. Are we to believe that our intuitions about the substrata of choice behavior are really as precise as this?

As we shall see, axiom 1 can serve as an alternate foundation for the analysis of choice behavior which, it turns out, imposes substantially the same restrictions upon paired comparisons data as the most widely used of Thurstone's models (see section 2.D.2). However, it is important to recall that Thurstone's constructs can be extended to the analysis of data obtained by category methods, such as equal appearing and successive intervals. Although these models are subject to criticisms in addition to those applicable to his paired comparisons model, they have been widely used with considerable success. To date, no one has suggested a comparable extension of the present theory to deal with category scaling. The difficulty in doing so probably arises, at least in part, from the unresolved conceptual problem discussed in section 1.A.3.

In other areas of choice behavior, specifically motivation and learning, it has been generally assumed that scaling or measurement either is irrelevant or can be indefinitely postponed. Among the exceptional sorties are the papers of Hull et al. [1947] and Young [1947]. However, to one familiar with measurement ideas, the notions of incentive value and response strength are suggestive of scales.

In all of the fields in which scales have been important they have been constructed under the assumption that only data for pairs of stimuli are known. In economics this has not been a limitation because of the algebraic nature of their models and the assumed transitivity of preference. In the psychological models, in which discrimination is admittedly not perfect, the pairwise data have not been known to determine choices from larger sets, and the whole problem has remained unresolved. As we have seen (theorem 1), axiom 1, if accepted, justifies complacency on that score.

The purpose of this section is to show that for situations in which pairwise choice discrimination is imperfect axiom 1 implies the existence of a ratio scale, i.e., one that is unique except for its unit, independent of any assumptions about the structure of the set of alternatives. This formulation can be used to solve all of the classical problems in a very simple way.

2. Existence Theorem

Theorem 3. *Suppose that T is a finite subset of U, that $P(x, y) \neq 0, 1$ for all $x, y \in T$, and that axiom 1 holds for T and its subsets, then there exists a positive real-valued function v on T, which is unique up to multiplication by a positive constant, such that for every $S \subset T$*

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)}.$$

PROOF. Define $v(x) = kP_T(x)$, where $k > 0$; then by part i of axiom 1 and part iii of the probability axioms we have

$$\begin{aligned} P_S(x) &= \frac{P_T(x)}{P_T(S)} \\ &= \frac{kP_T(x)}{\sum_{y \in S} kP_T(y)} \\ &= \frac{v(x)}{\sum_{y \in S} v(y)}, \end{aligned}$$

so existence is ensured.

To show uniqueness, suppose that v' is another such function; then for any $x \in T$

$$v(x) = kP_T(x) = \frac{kv'(x)}{\sum_{y \in T} v'(y)}.$$

Let $k' = k \Big/ \sum_{y \in T} v'(y)$, and we have $v(x) = k'v'(x)$, which concludes the proof.

In essence, what we have shown is this. If we confine ourselves to a local region T in which all the pairwise discriminations are imperfect, and if the several probability measures are related to one another so that P_S, $S \subset T$, acts like a conditional probability relative to P_T (axiom 1), then the distribution $P_T(x)$ can be interpreted as a particular choice of unit of a ratio scale over T. By itself, this observation seems too trite to warrant comment; however, these local scales can be extended throughout U in a sensible manner which has implications for psychology that seem to have been overlooked.

The practical will note that the v-scale obtained in theorem 3 is not really very useful as it stands for two reasons: (i) the probabilities $P_T(x)$ will be extremely difficult to estimate when T is at all large, and (ii) the scale is defined only over a set having no pairwise perfect discriminations, which is probably only a small portion of any dimension we might wish to scale. The first difficulty is much mitigated when we notice that v can be expressed as

$$v(x) = kP(x, a)/P(a, x),$$

where a is an arbitrary but fixed element of T and k is a positive constant. This follows from the fact that

$$P_T(x) = P_T(a)P(x, a)/P(a, x),$$

according to lemma 3 Thus, if the pairwise probabilities can be estimated sufficiently accurately so that the ratio $P(x, a)/P(a, x)$ is reliable, then v can be determined.

Actually, in practice it would be most ill-advised to estimate the v-scale in this manner because too little of the available data is used. Fortunately, much more efficient—maximum likelihood—estimating schemes are described in the literature (see section 1.E.4 for the references).

3. Extension of v-Scale

For each subset of U in which pairwise discriminations are imperfect the v-scale is defined. Unless we are willing to suppose that in actuality all discriminations are imperfect, we have, at this point, a whole collection of very local scales. These are of interest only if circumstances can be found under which they can be welded together to form a single scale over the whole of U. The basic idea for doing this is simple. If R and S are two sets over which v-scales are defined, and if they overlap, then the arbitrary scale constants are chosen so that the scales coincide over the region of overlap. The problem is to give plausible sufficient conditions so that the extension is possible and unique. These are formulated as two definitions.

Definition 1. *The universal set U with pairwise probabilities $P(x, y)$ is said to be* finitely connected *if for every $a, b \in U$ for which $P(b, a) > \frac{1}{2}$, there exists a finite sequence $x_1, x_2, \cdots, x_n \in U$ such that*

$$\tfrac{1}{2} \leq P(x_1, a) < 1, \quad \tfrac{1}{2} \leq P(x_{i+1}, x_i) < 1, \quad \text{and} \quad \tfrac{1}{2} \leq P(b, x_n) < 1,$$

where $i = 1, 2, \cdots, n - 1$.

Intuitively, this definition means that any two stimuli are connected via a finite chain of imperfect discriminations. For all practical purposes this condition is met by every psychophysical continuum, and it is probably suitable for other domains provided that we are not too niggardly in defining U.

The next definition has been suggested by Marschak and others (Block and Marschak [1957] and Davidson and Marschak [1958]), and they have studied some of its relations to other concepts.

Definition 2. *The universal set U with pairwise probabilities $P(x, y)$ is said to satisfy the condition of* strong stochastic transitivity *if for every $x, y, z \in U$ such that $P(x, y) \geq \frac{1}{2}$ and $P(y, z) \geq \frac{1}{2}$ then $P(x, z) \geq max\,[P(x, y), P(y, z)]$.*

It is clear that if all pairwise discriminations are imperfect the first definition is satisfied. If all pairwise discriminations are imperfect, and if axiom 1 holds for all sets of three elements, then the second is also met, as was shown in 1.D.3.

Theorem 4. *Suppose that P_T is defined for every $T \subset U$ such that $|T| \leq 3$, that axiom 1 holds for such sets, that U is finitely connected, and that the condition of strong stochastic transitivity is met. Then there exists a positive ratio scale v on U such that for every $T \subset U$ for which part* i *of axiom* 1 *holds*

$$P_T(x) = \frac{v(x)}{\sum_{y \in T} v(y)}.$$

PROOF. Choose any $a \in U$ and set $v(a) = k$, where k is a fixed positive number. Consider any other $b \in U$. If $P(a, b) = \frac{1}{2}$, set $v(b) = k$. If $P(b, a) > \frac{1}{2}$, then by finite connectivity there exists a sequence $x_1, x_2, \cdots, x_n \in U$ forming a chain of imperfect discriminations from a to b.

Set

$$v(b) = k \frac{P(x_1, a)P(x_2, x_1) \cdots P(x_n, x_{n-1})P(b, x_n)}{P(a, x_1)P(x_1, x_2) \cdots P(x_{n-1}, x_n)P(x_n, b)}.$$

If $P(a, b) > \frac{1}{2}$, then a similar sequence exists from b to a, and the corresponding definition is made.

To complete the proof, it must be shown that the definition is independent of the particular sequence chosen and that for a set of imperfectly discriminated alternatives the definition given here coincides with the v-scale of theorem 3. Let us suppose that $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_m$ are two suitable sequences from a to b, where, with no loss of generality, $P(b, a) > \frac{1}{2}$. Now, either $P(x_1, y_1) \geqq \frac{1}{2}$ or $< \frac{1}{2}$. Suppose, again with no loss of generality, that the former holds. By strong stochastic transitivity,

$$P(x_1, y_1) \leqq P(x_1, a) < 1,$$

so by applying theorem 2 to $\{a, x_1, y_1\}$ we have

$$\frac{P(x_1, a)}{P(a, x_1)} = \frac{P(x_1, y_1)P(y_1, a)}{P(y_1, x_1)P(a, y_1)}.$$

Because

$$v(y_1) = k \frac{P(y_1, a)}{P(a, y_1)},$$

it follows that

$$k \frac{P(x_1, a)}{P(a, x_1)} = k \frac{P(x_1, y_1)P(y_1, a)}{P(y_1, x_1)P(a, y_1)}$$
$$= v(y_1) \frac{P(x_1, y_1)}{P(y_1, x_1)}.$$

Thus, we can begin the argument at y_1 rather than at a. Proceeding inductively, we can continue to move the starting point through the finite number of elements up to b, in which case uniqueness is trivial.

If $b, c \in U$ are such that $\frac{1}{2} \leqq P(c, b) < 1$ and if $v(b)$ is defined in terms of the sequence $x_1, x_2, \cdots, x_n$, then we may define $v(c)$ in terms of the sequence $x_1, x_2, \cdots, x_n, b$. Thus

$$\frac{v(c)}{v(b)} = \frac{P(c, b)}{P(b, c)}$$

because k and all the x_i terms are common to both definitions and so they cancel. This establishes the compatibility of the present scale with the one discussed in theorem 3, hence proving the final assertion of the present theorem.

The role of finite connectedness is to permit an extension of v throughout all of U, and the role of strong stochastic transitivity is to ensure one dimensionality and, thus, a unique extension of v. It should be pointed out that we do not need quite so strong a condition as definition 2; it would suffice to demand that if $P(x, y) \geqq \frac{1}{2}$, $P(y, z) \geqq \frac{1}{2}$, and $P(x, z) < 1$ then $P(x, y) < 1$ and $P(y, z) < 1$.

One important practical consequence of this theorem is that axiom 1 need hold only for subsets of three alternatives in order for the v-scale to exist. Thus any proposed counter example to axiom 1 will be of interest only if it is based on sets of three alternatives.

Another fact brought out by the theorem is that although axiom 1 implies unidimensionality when pairwise discrimination is perfect throughout (lemma 4) or when it is imperfect throughout (theorem 3) other restrictions must be added to axiom 1 to get unidimensionality when there are mixed perfect and imperfect pairwise discriminations. In the mixed case axiom 1 amounts to an assumption of local linearity. This strongly suggests that axiom 1 by itself admits a multidimensional scaling model when discriminations are mixed; however I have not yet been able to construct such a model.

4. Previous Work

The hypothesis that a numerical scale v might exist such that

$$P_T(x) = \frac{v(x)}{\sum_{y \in T} v(y)}$$

has appeared, at least for paired comparisons, from time to time in the literature as an *ad hoc* assumption. For example, Thurstone [1930] and, following him, Gulliksen [1953] postulated this in a learning theory in which v was interpreted as "response strength" (see Chapter 4). Undoubtedly it has appeared in other specific applications.

Of much greater importance, however, is the existence of a relatively extensive statistical literature based upon the two alternative version of this model. For a number of years R. A. Bradley has championed the assumption for paired comparisons data that $P(x, y) = v(x)/[v(x) + v(y)]$, and he and various colleagues have developed methods for estimating the scale values and for testing certain statistical hypotheses. The existence of their work, which complements the theoretical work described here and which is of utmost importance for empirical applications of the model, means that the statistical aspects of the present theory are much better understood than one would have any reason to hope a priori.

Since these developments are generally available, there is no need to summarize them except to indicate briefly what has been done. Bradley and Terry [1952] present maximum likelihood estimates for the v's in the paired comparisons case, and they develop several likelihood ratio tests for the null hypothesis that all of the v's are the same. Asymptotic results are obtained, and tables are included of the maximum likelihood estimates and the test statistics for small sample sizes. These tables are extended in Bradley [1954a]. In Bradley [1955] the power of the tests and the reliability of the estimators are studied. An extension of these methods to treatments that form a factorial set in paired comparisons is presented by Abelson and Bradley [1954]. Independently of this work, Ford [1957] has discussed the maximum likelihood estimates of the v's for the same model with, however, the generalization that the sample size may differ from pair to pair. Finally, in Bradley [1954b] the goodness of fit of the underlying model is discussed. He develops a test statistic having the χ^2 distribution for large sample sizes, which he shows is approximately the same as the ordinary test statistic based upon expected frequencies calculated from the maximum likelihood estimates. Twenty tests of the model, based upon data from two experiments, are given, and only one χ^2 is significant at the 0.05 level. In addition, Bradley refers to unpublished work of J. W. Hopkins in which extensive tests of the model have been conducted on taste sensations; he reports that these data give no reason to reject the model.

F. INDEPENDENCE-OF-UNIT CONDITION

1. Statement of Condition

In this section a condition about theory construction is developed that limits significantly the possible form that certain theories based on axiom 1 can assume. This condition must, I believe, be classed as extra-empirical, since it is intended to capture in part what we mean by an acceptable theory. Its application is not restricted to situations in which axiom 1 is assumed but holds whenever one or more of the variables involved form ratio scales. Since arguments of the type to be used have not often occurred in the behavioral sciences and may, therefore, seem suspect to some, it should be pointed out that they have adequate precedent from physics. For example, the condition that the laws of physics should be independent of translations and rotations of the coordinate system within which they are stated seems innocent enough, but it limits appreciably the possible physical laws. It is a condition about the nature of theory and the use of the word law, not an empirical hypothesis.

After discussing the condition and an empirical assumption, we will see how they may be used to analyze a problem of some inherent interest: the time- and space-order effects. Later they reappear in the analysis of the signal detectability problem (section 2.E) and of learning (Chapter 4).

By definition, a ratio scale is specified except for its unit, which is not

only unknown but unknowable. The unit is a matter of convention. Therefore, it seems illegitimate, if not actually inconsistent, for a theory to presuppose that the unit is known. And so, in its most general form, the condition to be imposed is that *any theory involving a ratio scale shall be independent of the unit chosen for that scale.* Indeed, if this were not so, empirical observations determining the form of the theory would permit us to evaluate the unit, and so the scale would be stronger than a ratio scale.

This condition must now be made specific to the present choice problem. Suppose, for example, that theorem 4 holds for choice probabilities over a set U of alternatives both before and after the occurrence of some event which is relevant to the organism making the choice. The event might be some physical stimulus in a psychophysical experiment, or it might be the occurrence of reward in a learning experiment, etc. In general, an event may be thought of as effecting a change of state in the organism. The scale values in one state come about as a modification, due to the event, of those that existed in the other state. It is appropriate to think in terms of states rather than events, even though only the latter interpretation is used in this book, because the general principle is concerned with the effect of different determiners of behavior upon the scale values, not just temporal events. Let us consider theories in which the scale value for alternative $x \in U$ is dependent upon only three things when the organism is in the state S_2: the states S_1 and S_2 and the scale value in S_1. Thus there is no loss of generality in writing the transformed value as $f[v(x)]$, where $v(x)$ is the scale value for state S_1 and f is a function which depends only upon x and the states S_1 and S_2. The condition, then, says that the mathematical form of f shall not depend upon our choice of unit, which is to say that if v transforms into $f(v)$ for a particular unit, and if we change the unit by multiplying throughout by a positive constant k, then kv must be transformed into $kf(v)$. In summary, we may state the condition as follows:

Independence-of-unit condition.

Suppose that the choice probabilities of an organism over subsets of U satisfy the conditions of theorem 4 both when the organism is in state S_1 and when it is in state S_2. Suppose, further, that the scale values for $x \in U$ can be written as $v(x)$ and $[v(x)]$ for S_1 and S_2, respectively, where f is a function that depends only upon x, S_1, and S_2. Then, for any $k > 0$

$$f[kv(x)] = kf[v(x)].$$

2. Behavioral Continuity

Although the point is rarely raised, implicit in most scaling theories is the assumption that any real number can appear as a scale value. Certainly, the evidence has not been so overwhelming that theorists have been forced to assume otherwise. Nonetheless, it is an assumption that may in fact be false, and some have felt, in particular, that an unbounded v-scale is counter intuitive. As we shall see later, when we study learning models, the assumption of a bounded v-scale leads us to certain peculiar results. So let us tentatively impose the following assumption:

Unboundedness assumption.

Any positive real number is a possible value on the v-scale.

The independence-of-unit condition, together with the unboundedness assumption, determines explicitly the form of the transformation f, since by the unboundedness assumption the number 1 is a possible scale value whatever the unit may be, and so, by the independence-of-unit condition

$$\begin{aligned} f(v) &= f(v1) \\ &= vf(1). \end{aligned}$$

This is to say, the only admissible transformations of the v's are multiplications by positive constants (positive because $f(1)$ must be a scale value). It is useful to replace $f(1)$ by a symbol which makes its dependencies explicit, e.g., α_{ix}, where i refers to the event that effects the transition from one state to another and x refers to the alternative.

3. Response Bias

The essential feature of any experiment in which the so-called time- or space-order errors appear is this. The subject is confronted by several stimuli that he is to rank according to some intrinsic but, to him, ambiguous property. Each stimulus is temporarily identified by some unambiguous label that is accessible both to the subject and to the experimenter and that is unrelated to the property the subject is judging. For example, if the stimuli are weights of nearly the same mass, the property to be judged can be relative heaviness and the temporary labels can be their serial positions in the order of lifting. Or, if the stimuli are patches of light to be judged according to relative brightness, then their location in space can be used as the temporary identifying label. Other identifications can be used so long as they are not correlated with the dimension being judged. It is a priori clear that however the objects may be labeled the subject may exhibit a bias among the labeling categories; it is doubtful that the bias really has much to do with space or time or order, and so following Irwin [1958] the more neutral term *response bias* is used.

Whatever it is called, the bias occurs and must be coped with in some manner—and the obvious idea of randomizing it out of existence only

muddies the data beyond use. The analysis that is to be proposed is most easily illustrated by a specific example; once that is understood, the generalizations will be obvious. Let the stimuli be three weights called H, M, and L (for heavy, medium, and light). These are presented sequentially, and the subject identifies the one that he thinks is heaviest by saying whether it was the first, second, or third presented. The data table consists of six rows, one for each of the orders HML, HLM, etc., and three columns, one for each of the response categories.

Prior to hefting the weights, a certain differential tendency to use the categories may be assumed to exist. Let the corresponding v-values be v_1, v_2, and v_3. After lifting the weights, these tendencies will be altered, and, if we are willing to suppose that the modification depends only upon the weights lifted, their order, and the value of the response category, then the independence-of-unit condition and the unboundedness assumption imply that the effect will be multiplicative. Thus the three v-values in row i may be written as

$$\alpha_{i1}v_1, \qquad \alpha_{i2}v_2, \quad \text{and} \quad \alpha_{i3}v_3.$$

Although it may be that the effect upon each response category depends upon all three of the weights, it would be much simpler if there were no interaction. Assuming this is so, then the effect actually depends only upon the weight that happens to correspond to that category. Thus there are only three parameters, say α corresponding to H, β to M, and γ to L. The model may then be summarized as

$$\begin{array}{r} \\ HML \\ HLM \\ MHL \\ LHM \\ MLH \\ LMH \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \alpha v_1 & \beta v_2 & \gamma v_3 \\ \alpha v_1 & \gamma v_2 & \beta v_3 \\ \beta v_1 & \alpha v_2 & \gamma v_3 \\ \gamma v_1 & \alpha v_2 & \beta v_3 \\ \beta v_1 & \gamma v_2 & \alpha v_3 \\ \gamma v_1 & \beta v_2 & \alpha v_3 \end{bmatrix} \end{array}.$$

Assuming imperfect discrimination, the probabilities in each row are determined according to theorem 4 by the values in each row. For example, the probabilities in the HLM row are

$$\frac{\alpha v_1}{\alpha v_1 + \gamma v_2 + \beta v_3}, \qquad \frac{\gamma v_2}{\alpha v_1 + \gamma v_2 + \beta v_3}, \qquad \frac{\beta v_3}{\alpha v_1 + \gamma v_2 + \beta v_3}.$$

Since the unit in each row may be changed without affecting the probabilities, the entire table may be divided by γv_1, yielding

$$\begin{bmatrix} \frac{\alpha}{\gamma} & \frac{\beta v_2}{\gamma v_1} & \frac{v_3}{v_1} \\ \frac{\alpha}{\gamma} & \frac{v_2}{v_1} & \frac{\beta v_3}{\gamma v_1} \\ \frac{\beta}{\gamma} & \frac{\alpha v_2}{\gamma v_1} & \frac{v_3}{v_1} \\ 1 & \frac{\alpha v_2}{\gamma v_1} & \frac{\beta v_3}{\gamma v_1} \\ \frac{\beta}{\gamma} & \frac{v_2}{v_1} & \frac{\alpha v_3}{\gamma v_1} \\ 1 & \frac{\beta v_2}{\gamma v_1} & \frac{\alpha v_3}{\gamma v_1} \end{bmatrix}.$$

This, in turn, can be decomposed by matrix multiplication into

$$\begin{bmatrix} \frac{\alpha}{\gamma} & \frac{\beta}{\gamma} & 1 \\ \frac{\alpha}{\gamma} & 1 & \frac{\beta}{\gamma} \\ \frac{\beta}{\gamma} & \frac{\alpha}{\gamma} & 1 \\ 1 & \frac{\alpha}{\gamma} & \frac{\beta}{\gamma} \\ \frac{\beta}{\gamma} & 1 & \frac{\alpha}{\gamma} \\ 1 & \frac{\beta}{\gamma} & \frac{\alpha}{\gamma} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{v_2}{v_1} & 0 \\ 0 & 0 & \frac{v_3}{v_1} \end{bmatrix}.$$

Observe that there are really only four parameters, not six, in this model. If four stimuli are used, the generalization is clear: there are 24 rows, four columns, and six parameters. And so on. In this fashion it is clear that we can separate out the so-called time- or space-order errors, making it possible to use psychophysical data to check axiom 1.

4. Estimation of Parameters

To apply this model, it is necessary to estimate the several parameters from data. No work has been done on optimal estimation methods, but

the following technique has been used with success. Its main virtue is algebraic simplicity.

Let P_{ij} denote the probability corresponding to the ith row and the jth column, and let $A_{ijk} = P_{ij}/P_{ik}$. Then the three-stimulus model implies the following matrix:

$$\begin{array}{c} \\ HML \\ HLM \\ MHL \\ LHM \\ MLH \\ LMH \end{array} \begin{array}{c} \begin{array}{ccc} A_{i12} & A_{i13} & A_{i23} \end{array} \\ \begin{bmatrix} \frac{\alpha v_1}{\beta v_2} & \frac{\alpha v_1}{\gamma v_3} & \frac{\beta v_2}{\gamma v_3} \\ \frac{\alpha v_1}{\gamma v_2} & \frac{\alpha v_1}{\beta v_3} & \frac{\gamma v_2}{\beta v_3} \\ \frac{\beta v_1}{\alpha v_2} & \frac{\beta v_1}{\gamma v_3} & \frac{\alpha v_2}{\gamma v_3} \\ \frac{\gamma v_1}{\alpha v_2} & \frac{\gamma v_1}{\beta v_3} & \frac{\alpha v_2}{\beta v_3} \\ \frac{\beta v_1}{\gamma v_2} & \frac{\beta v_1}{\alpha v_3} & \frac{\gamma v_2}{\alpha v_3} \\ \frac{\gamma v_1}{\beta v_2} & \frac{\gamma v_1}{\alpha v_3} & \frac{\beta v_2}{\alpha v_3} \end{bmatrix} \end{array}.$$

It is easy to see that

$$\frac{\alpha}{\beta} = \left(\frac{A_{112}A_{213}A_{423}}{A_{312}A_{513}A_{623}}\right)^{1/6}$$

$$\frac{\beta}{\gamma} = \left(\frac{A_{512}A_{313}A_{123}}{A_{612}A_{413}A_{223}}\right)^{1/6}$$

$$\frac{\alpha}{\gamma} = \left(\frac{A_{212}A_{113}A_{323}}{A_{412}A_{613}A_{523}}\right)^{1/6}$$

$$\frac{v_1}{v_2} = \left(\prod_{i=1}^{6} A_{i12}\right)^{1/6}$$

$$\frac{v_1}{v_3} = \left(\prod_{i=1}^{6} A_{i13}\right)^{1/6}$$

$$\frac{v_2}{v_3} = \left(\prod_{i=1}^{6} A_{i23}\right)^{1/6}$$

As a computational check, it is not difficult to show that the following equations must hold:

$$\left(\frac{\hat{\alpha}}{\beta}\right)\left(\frac{\hat{\beta}}{\gamma}\right)\Big/\left(\frac{\hat{\alpha}}{\gamma}\right) = 1$$

$$\left(\frac{\hat{v}_1}{v_2}\right)\left(\frac{\hat{v}_2}{v_3}\right)\Big/\left(\frac{\hat{v}_1}{v_3}\right) = 1.$$

G. ALGEBRAIC APPROXIMATIONS[5]

As pointed out in section 1.A.1, there is some disagreement as to whether an algebraic or probabilistic description of choice behavior is to be preferred. One argument favoring the algebraic approach is the apparent greater simplicity of the resulting mathematics, and some would hold that even if the probabilistic model were more accurate it should, nonetheless, be replaced by some algebraic approximation. In this section the properties of two such approximations to the pairwise discriminations are examined when axiom 1 is assumed to hold for sets of three alternatives.

1. Just Noticeable Differences

The most ancient and honorable, if frequently misunderstood, technique for passing from a probabilistic to an algebraic model is to introduce the concept of a just noticeable difference (jnd). This has been widely employed in psychophysics; however, there is no particular reason to restrict it to any special class of choice phenomena. The essential idea is to pick a probability cutoff π, $\frac{1}{2} < \pi < 1$, and to say that alternatives discriminated more than 100π per cent of the time are more than one jnd apart; those discriminated less often are one jnd or less apart. This can be cast in the language of binary relations as follows:

Definition 3. *Suppose that for every $x, y \in U$, $P(x, y)$ is defined, and let π be a fixed number, $\frac{1}{2} < \pi < 1$. The relation $L(\pi)$ on U is defined by $xL(\pi)y$ if and only if $P(x, y) > \pi$. The relation $I(\pi)$ on U is defined by $xI(\pi)y$ if and only if $1 - \pi \leq P(x, y) \leq \pi$.*

The intuitive meaning of $L(\pi)$ is "at least one π-jnd larger" and of $I(\pi)$, "not more than one π-jnd apart." It is, of course, necessary to specify the value of π, since these relations change with changes in π. That is to say, it is meaningless to speak of jnds without specifying the probability cutoff that was used to define them—a point unfortunately all too often ignored in the experimental literature.

[5] This section is included for completeness, but it is not necessary in order to understand any of the following work.

The question is what properties—axioms—these relations can be expected to meet. We might well expect $L(\pi)$ to be transitive—that when x is at least one jnd larger than y and y at least one jnd larger than z then x should be at least one jnd larger than z. On the other hand, we definitely do not expect $I(\pi)$ to be transitive. In Luce [1956] this question was treated abstractly in terms of conditions that two such relations might be expected to satisfy, and the following axiom system was offered:

Semiorder axioms. *Let L and I be binary relations on a set U. (L, I) is said to be a semiordering of U if for every $x, y, z, w \in U$*

(i) exactly one of xLy, yLx, or xIy obtains,
(ii) xIx,
(iii) xLy, yIz, zLw imply xLw,
(iv) xLy and yLz imply not both xIw and wIz.

Theorem 5. *Let T be any subset of U in which all pairwise discriminations are imperfect; suppose that P_S is defined for every $S \subset T$ such that $|S| \leq 3$; and suppose that for these subsets axiom 1 holds. Then for each π, $\frac{1}{2} < \pi < 1$, the relations $L(\pi)$ and $I(\pi)$ form a semiordering of T.*

PROOF. By theorem 4, there exists a scale v on T such that

$$P(x, y) = v(x)/[v(x) + v(y)]$$
$$= 1/[1 + v(y)/v(x)].$$

Thus $xL(\pi)y$ if and only if $1/[1 + v(y)/v(x)] > \pi$, which is equivalent to $v(x)/v(y) > \pi/(1 - \pi)$. Similarly, $xI(\pi)y$ if and only if $(1 - \pi)/\pi \leq v(x)/v(y) \leq \pi/(1 - \pi)$. Now the four semiorder axioms can be checked. The first two are trivial. The hypotheses of the third amount to

$$v(x)/v(y) > \pi/(1 - \pi), \qquad (1 - \pi)/\pi \leq v(y)/v(z) \leq \pi/(1 - \pi),$$
$$v(z)/v(w) > \pi/(1 - \pi).$$

Thus

$$\frac{v(x)}{v(w)} = \frac{v(x)}{v(y)}\frac{v(y)}{v(z)}\frac{v(z)}{v(w)}$$
$$> \frac{\pi}{(1 - \pi)}\frac{(1 - \pi)}{\pi}\frac{\pi}{(1 - \pi)}$$
$$= \frac{\pi}{1 - \pi},$$

as was to be shown. Suppose in the fourth axiom that $xI(\pi)w$, then we show $wL(\pi)z$. The hypotheses amount to

$$v(x)/v(y) > \pi/(1-\pi), \qquad v(y)/v(z) > \pi/(1-\pi),$$
$$(1-\pi)/\pi \leqq v(x)/v(w) \leqq \pi/(1-\pi).$$

Thus

$$\frac{v(w)}{v(z)} = \frac{v(y)}{v(z)}\frac{v(x)}{v(y)}\frac{v(w)}{v(x)} > \frac{\pi}{(1-\pi)}\frac{\pi}{(1-\pi)}\frac{(1-\pi)}{\pi} = \frac{\pi}{1-\pi},$$

which concludes the proof.

If we wish to treat two alternatives that are less than one jnd apart as being, in a sense, indifferent, then the foregoing algebraic system seems to be appropriate. It is, however, in some ways more difficult to work with than the one to be described next.

2. The Trace

In most of the algebraic models of choice it has been customary to work with weak orderings which have the important property that the indifference relation is transitive.

Weak order axioms. *Let R be a binary relation on a set U. R is said to be a weak ordering of U if for every $x, y, z \in U$*

(i) *either xRy or yRx or both,*
(ii) *(transitivity) xRy and yRz imply xRz.*

By defining xLy to mean xRy but not yRx and xIy to mean both xRy and yRx, it is not difficult to show that (L, I) satisfies the semiorder axioms; however, I is also transitive.

The importance of weak orders stems mainly from the fact that if v is a numerical mapping of U that preserves the order of a binary relation R on U, i.e.,

$$v(x) \geqq v(y) \text{ if and only if } xRy,$$

then R must be a weak ordering of U.

The problem now is whether there is a weak order that approximates a probabilistic model. As shown in Luce [1956], every semiorder induces a natural weak order, so from theorem 5 we know that if axiom 1 holds we can induce an infinity of weak orders, one for each value of π. These, however, are less interesting and less refined in a sense than the following relation defined in Luce [1958].

Definition 4. *Suppose that for every $x, y \in U$, $P(x, y)$ is defined. The relation $\gtrsim$ defined by $x \gtrsim y$ if and only if $P(x, z) \geqq P(y, z)$ for all $z \in U$ is called the* trace *of P.*

It is easy to see that the trace is a transitive relation, but without some restrictions on P it need not be a weak order. That is, there may exist incomparable pairs (x, y) in the sense that z and $z' \in U$ can be found such that $P(x, z) > P(y, z)$ and $P(x, z') < P(y, z')$. Unfortunately, we cannot show that the trace is a weak order under exactly the same conditions employed in theorem 5 because, unlike the relations $L(\pi)$ and $I(\pi)$ which are defined just in terms of the probabilities of two alternatives, the trace depends upon the relation of x and y to all other alternatives in U.

Theorem 6. *Suppose that P_S is defined for every subset $S \subset U$ such that $|S| \leqq 3$, that axiom 1 holds for such sets, and that all pairwise discriminations are imperfect. Then the trace is a weak order.*

PROOF. By definition of the trace, $x \gtrsim y$ if and only if $P(x, z) \geqq P(y, z)$, $z \in U$. Since imperfect discrimination and axiom 1 imply strong stochastic transitivity, theorem 4 holds, and so the preceding condition is equivalent to

$$\frac{v(x)}{v(x) + v(z)} \geqq \frac{v(y)}{v(y) + v(z)},$$

which in turn is equivalent to $v(x) \geqq v(y)$. Thus, for every $x, y \in U$, either $x > y$ or $y \gtrsim x$, as was to be shown.

Corollary. *Under the conditions of the theorem, $x \gtrsim y$ if and only if $P(x, y) \geqq \frac{1}{2}$.*

PROOF. Obvious.

This theorem can also be shown as follows: as noted in section 1.E.3, the hypotheses imply the condition of strong stochastic transitivity, and Block and Marschak [1957] have shown this condition to be equivalent to the trace being a weak order.