STUDIES IN THE METHODOLOGY AND FOUNDATIONS OF SCIENCE

SYNTHESE LIBRARY

MONOGRAPHS ON EPISTEMOLOGY,

LOGIC, METHODOLOGY, PHILOSOPHY OF SCIENCE,

SOCIOLOGY OF SCIENCE AND OF KNOWLEDGE,

AND ON THE MATHEMATICAL METHODS OF

SOCIAL AND BEHAVIORAL SCIENCES

PATRICK SUPPES

STUDIES IN THE METHODOLOGY AND FOUNDATIONS OF SCIENCE

Selected Papers from 1951 to 1969

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

ISBN 978-90-481-8320-3 ISBN 978-94-017-3173-7 (eBook)
DOI 10.1007/978-94-017-3173-7

TO ERNEST NAGEL AND ALFRED TARSKI

PREFACE

The twenty-three papers collected in this volume represent an important part of my published work up to the date of this volume. I have not arranged the paper chronologically, but under four main headings.

Part I contains five papers on methodology concerned with models and measurement in the sciences. This part also contains the first paper I published, 'A Set of Independent Axioms for Extensive Quantities', in *Portugaliae Mathematica* in 1951.

Part II also is concerned with methodology and includes six papers on probability and utility. It is not always easy to separate papers on probability and utility from papers on measurement, because of the close connection between the two subjects, but Articles 6 and 8, even though they have close relations to measurement, seem more properly to belong in Part II, because they are concerned with substantive questions about probability and utility.

The last two parts are concerned with the foundations of physics and the foundations of psychology. I have used the term *foundations* rather than *philosophy*, because the papers are mainly concerned with specific axiomatic formulations for particular parts of physics or of psychology, and it seems to me that the term *foundations* more appropriately describes such constructive axiomatic ventures. Part III contains four papers on the foundations of physics. The first paper deals with foundations of special relativity and the last three with the role of probability in quantum mechanics. I regret not including some of the earlier work with J. C. C. McKinsey on the foundations of classical mechanics, but I already have given an account in general terms of that work in the last chapter of my *Introduction to Logic.*

The largest number of papers is in the final part on the foundations of psychology. The greater concentration of papers here correctly reflects my interests over the past decade. In fact, the bulk of my papers that I think are of some conceptual importance and that are not included in this volume are papers that lie strictly within mathematical psychology as a

scientific discipline, and are consequently not really appropriate for inclusion in the present volume. Because so many of the papers in Part IV are concerned with the psychological foundations of mathematics in one form or another, it would almost have been appropriate to have so labeled Part IV. But I have included several papers that have nothing directly to do with his subject, and so I have kept the more general title.

Two of the papers in this collection were written with co-authors. The fourth paper, 'Foundational Aspects of Theories of Measurement', was written jointly with Dana Scott. The eighth paper, 'An Axiomatization of Utility Based on the Notion of Utility Differences', was written jointly with Muriel Winet. The appearance of these two papers here has been generously agreed to by them.

Two of the papers have not previously been published. These are the papers, 'Behaviorism' and 'On the Theory of Cognitive Processes'. The first of these two papers was written between 1963 and 1965, and was given in various revised forms at Swarthmore College and at the University of Illinois. The second of the two papers was written in 1967 and was given as an Arnold Isenberg Memorial Lecture at Michigan State University.

Acknowledgments for permission to publish the various papers are given at the bottom of the first page of each article, but thanks are extended here to the many editors and publishers who generously granted this permission. No substantive or real stylistic changes have been made in any of the articles; only the manner of referring to published articles and books has been standardized, as has the formatting of section headings. Footnotes in the original articles are numbered, beginning anew with each article; the bibliographic footnotes that originate with this publication are indicated by a dagger following the number. Some introductory remarks about each article and subsequent pertinent literature are to be found at the beginning of each part.

The publication of a series of papers spanning more than 18 years of work seems an appropriate occasion to acknowledge some of the intellectual debts I have incurred during those years. For my initial introduction to the philosophy of science and for continual intellectual counsel and advice, I owe a great deal to Ernest Nagel. Shortly after my arrival at Stanford University in 1950, J. C. C. McKinsey joined the faculty of the Department of Philosophy, and I learned from him the set-theoretical

tools that have been one of my main stocks in trade over the years. Various collaborative work that we planned beyond the several articles we originally published was abruptly halted by his untimely death in 1953. McKinsey acknowledged that the greatest influence on his scientific career had been Alfred Tarski. Both through McKinsey and also through direct acquaintance with Tarski, including attendance at his seminars at Berkeley during the years when I was first at Stanford and through his published works as well, I learned much of what I know about intellectual clarity and precision. And so, I take this occasion to acknowledge how much I owe to Tarski. I am indebted to William K. Estes for my first introduction to the conceptual and foundational problems of psychology, especially mathematical learning theory. We worked together intensively during 1955–56 when we were both Fellows at the Center for Advanced Study in the Behavioral Sciences at Stanford. In more recent years I have also learned much from Duncan Luce about both mathematical psychology and the theory of measurement.

It is important to record that I owe a very considerable debt to my younger colleagues as well. I mention especially Dana Scott and Richard C. Atkinson.

The idea of putting this volume together originated with Donald Davidson and Jaakko Hintikka. To both of them I owe a further debt for enlightening and penetrating conversations about philosophical matters. This is especially true of Davidson, who was my colleague at Stanford for many years. Finally, I want to acknowledge the dedicated and able editorial assistance of Miss Diana Axelson and Mrs. Lillian O'Toole, as well as the excellent help of Miss Anne Fagot in preparing the indexes and that of Mrs. Maria Jedd in preparing the illustrations.

PATRICK SUPPES

Stanford, California, January 1969

CONTENTS

PART I

METHODOLOGY: MODELS AND MEASUREMENT

The five papers in this part examine some of the many issues surrounding the general use of models in empirical science, and also the conceptual foundations of the theory of measurement. In discussing both the general use of models and the particular case of measurement, I have tried to show in these papers how set-theoretical tools standard in modern mathematics can also be used to good advantage in discussing methodological matters in the empirical sciences.

Because of restrictions on re-publication, I have not been able to include later papers that carry further some of the themes begun in these five papers. In connection with the theory of measurement, I would mention particularly the joint article with Joseph L. Zinnes, 'Basic Measurement Theory', which is Chapter 1 of the *Handbook of Mathematical Psychology*, published in 1963.[1] In that article, Zinnes and I give a more leisurely and general approach to the theory than is to be found in any one of the articles in Part I of the present volume, although most of the ideas worked out by Zinnes and me are anticipated in the fourth article in this volume, which I wrote jointly with Dana Scott. I mention the joint article with Zinnes especially for those who are interested in general questions about the theory of measurement, but find the fourth article somewhat heavy going. At the end of the article with Scott, there is a conjecture about finite axiomatizability. Tait (1959) has given a counterexample to this conjecture, and the subject remains as intractable as ever. On the other hand, some important, positive results giving necessary and sufficient conditions for various qualitative measurement structures to have a numerical representation are given in Scott (1964).

The ideas on empirical meaningfulness begun in the fifth article are also extended in the article with Zinnes. The development of a three-valued logic is extended further in Suppes (1965b).

The second article, the one on models of data, moves in a promising direction that I have not yet had the opportunity to explore in greater depth. More explicit and more extended analysis of the relation between

theories and data, once the data are expressed in canonical form, is much needed. From a general philosophical standpoint, the analysis of this relation brings back many aspects of atomism and earlier, simplistic versions of logical positivism. The difference in the present case, however, is that the selection of the canonical form of the data, that is, how the data of experiments are to be recorded, is not something fixed in nature or in the perceiving apparatus of men, but is something subject to modification in light of experience and according to demands of current theory. I continue to stand by what I say in this article, but I do recognize it as only a bare beginning.

The third article on extensive measurement is part of a continuing stream of articles on this subject since its publication in 1951; perhaps the best and latest treatment is to be found in an article by Luce and Marley (1969). At the end of the 1951 article, I mentioned two problems that remain unsolved by the analysis given there. The first is that the set of objects must contain an infinite number of elements because of the closure condition on the basic operation of combination introduced. The second problem concerns the absence of any theory of error as part of the basic conceptual framework introduced. The problem of error is still, I think, in an unsatisfactory state, although some significant progress is made in Krantz (1967). On the other hand, the problem of finiteness can be handled rather directly. What I consider to be the simplest and, in many ways, most attractive axioms for the finite case can be sketched in a few pages, and I think it may be of interest to do that here, especially because none of the developments in this set of axioms is at all technical in character. The axioms themselves are, in a strict logical sense, elementary in that the theory can be stated as a theory with standard formalization, that is, as a formalized theory within first-order predicate logic with identity.

We may develop the axioms of extensive measurement with at least three specific interpretations in mind. One is for the measurement of mass by means of an equal-arm balance, one is for the measurement of length of rigid rods, and one is for the measurement of subjective probability. Other interpretations are certainly possible, but I shall restrict detailed remarks to these three.

From a formal standpoint the basic structures are triples $\langle X, \mathscr{F}, \geqslant \rangle$, where X is a nonempty set, $\mathscr{F}$ is a family of subsets of X, and the

relation $\succcurlyeq$ is a binary relation on $\mathscr{F}$. By using subsets of X as objects, the need for a separate primitive concept of concatenation is avoided, contrary to the requirement in Article 3. As a general structural condition, it shall be required that $\mathscr{F}$ be an *algebra of sets* on X, which is just to require that $\mathscr{F}$ be nonempty and be closed under union and complementation of sets, i.e., if A and B are in $\mathscr{F}$ then $A \cup B$ and $\sim A$ are also in $\mathscr{F}$.

In addition to their finiteness, the distinguishing characteristic of the structures considered is that the objects are equally spaced in an appropriate sense along the continuum, so to speak, of the property being measured. The restrictions of finiteness and equal spacing enormously simplify the mathematics of measurement, but it is fortunately not the case that the simplification is accompanied by a total separation from realistic empirical applications. Finiteness and equal spacing are characteristic properties of many standard scales, for example, the ordinary ruler, the set of standard weights used with an equal-arm balance in the laboratory or shop, or almost any of the familiar gauges for measuring pressure, temperature, or volume.

The intended interpretations of the primitive concepts for the three cases mentioned is fairly obvious. In the case of mass, X is a set of physical objects, and for two subsets A and B, $A \succcurlyeq B$ if and only if the set A of objects is judged at least as heavy as the set B. It is probably worth emphasizing that several different uses of the equal-arm balance are appropriate for reaching a judgment of comparison. For example, if $A = \{x, y\}$ and $B = \{x, z\}$, it will not be possible literally to put A on one pan of the balance and simultaneously B on the other, because the object x is a member of both sets, but the comparison can be made in at least two different ways. One is just to compare the nonoverlapping parts of the two subsets, which in the present case just comes down to the comparison of $\{y\}$ and $\{z\}$. A rather different empirical procedure that even eliminates the need for the balance to be equal arm is to first just balance A with sand on the other pan (or possibly water, but in either case, sand or water in small containers), and then to compare B with this fixed amount of sand. No additional interpretations of these operations are required, even of union of sets, which serves as the operation of concatenation, when the standard meaning of the set-theoretical operations of intersection, union, and complementation is given.

In the case of the rigid rods, the set X is just the collection of rods, and $A \succcurlyeq B$ if and only if the set A of rods, when laid end to end in a straight line is judged longer than the set B of rods also so laid out. Variations on exactly how this qualitative comparison of length is to be made can easily be supplied by the reader.

In the case of subjective probability, the set X is the set of possible outcomes of the experiment or empirical situation being considered. The subsets of X in $\mathscr{F}$ are just events in the ordinary sense of probability concepts, and $A \succcurlyeq B$ if and only if A is judged at least as probable as B.

Axioms for extensive measurement, subject to the two restrictions of finitude and equal spacing, are given in the following definition. In Axiom 5, $\approx$ is the equivalence relation defined in the standard fashion in terms of $\succcurlyeq$; namely, $A \approx B$ if and only if $A \succcurlyeq B$ and $B \succcurlyeq A$.

DEFINITION: *A structure* $\chi = \langle X, \mathscr{F}, \succcurlyeq \rangle$ *is a finite, equally spaced extensive structure if and only if* X *is a finite set,* $\mathscr{F}$ *is an algebra of sets on* X*, and the following axioms are satisfied for every* A*,* B*, and* C *in* $\mathscr{F}$*:*

1. *The relation* $\succcurlyeq$ *is a weak ordering of* $\mathscr{F}$;
2. *If* $A \cap C = \emptyset$ *and* $B \cap C = \emptyset$*, then* $A \succcurlyeq B$ *if and only if* $A \cup C \succcurlyeq B \cup C$;
3. $A \succcurlyeq \emptyset$;
4. *Not* $\emptyset \succcurlyeq X$;
5. *If* $A \succcurlyeq B$ *then there is a* C *in* $\mathscr{F}$ *such that* $A \approx B \cup C$.

From the standpoint of the standard ideas about the measurement of mass or length, it would be natural to strengthen Axiom 3 to assert that if $A \neq \emptyset$, then $A > \emptyset$, but because this is not required for the representation theorem and is unduly restrictive in the case of subjective probability, the weaker axiom seems more appropriate.

In stating the representation and uniqueness theorem, we use the notion of an additive measure μ from $\mathscr{F}$ to the real numbers, i.e., a function μ such that for any A and B in $\mathscr{F}$

(i) $\quad \mu(\emptyset) = 0,$

(ii) $\quad \mu(A) \geqslant 0,$

(iii) $\quad$ if $A \cap B = \emptyset$ then $\mu(A \cup B) = \mu(A) + \mu(B),$

where $\emptyset$ is the empty set, and it is also required for the applications intended here that $\mu(X) > 0$. A surprisingly strong representation theorem can be proved to the effect that there are only two nonequivalent sorts of atoms.

THEOREM: *Let* $\chi=\langle X, \mathscr{F}, \geqslant\rangle$ *be a finite, equally spaced extensive structure. Then there exists an additive measure* μ *such that for every* A *and* B *in* $\mathscr{F}$

$$\mu(A)\geqslant\mu(B) \quad \textit{if and only if} \quad A\geqslant B,$$

and the measure μ *is unique up to a positive similarity transformation. Moreover, there are at most two equivalence classes of atomic events in* $\mathscr{F}$; *and if there are two rather than one, one of these contains the empty event.*[2]

Proof: It will suffice to restrict ourselves to the atomic events, for by Axiom 2, the results can then easily be extended to any event. From the finiteness of X, it follows at once that there are a finite number of atomic events, which by Axiom 1 may be arranged in an ordered set of equivalence classes, where $\approx$ is the equivalence relation. Let $\mathscr{A}_0, \mathscr{A}_1, \ldots, \mathscr{A}_n$ be these classes with $\mathscr{A}_m<\mathscr{A}_n$ if $m<n$, and if A_0 is in $\mathscr{A}_0$, set $\mu(A_0)=0$. If A_1 is in $\mathscr{A}_1$,

$$\mu(A_1)=1.$$

Now consider an atomic event A_2 in $\mathscr{A}_2$. By virtue of Axiom 5, there must exist an event C such that if A_1 is any atomic event in $\mathscr{A}_1$

$$A_1\cup C\approx A_2,$$

but clearly C can have as members only atoms belonging to $\mathscr{A}_1$. Thus if k is the cardinality of $A_1\cup C$, we assign $\mu(A_2)=k$, and again this same measure to every atomic event in $\mathscr{A}_2$.

To prove that every $\mathscr{A}_n$ is an integer multiple of $\mathscr{A}_1$ in the sense just indicated, suppose there is some equivalence class that is not. Let $\mathscr{A}_n$ be the first such in the ordering. Then there must exist a C such that for any A_n in $\mathscr{A}_n$

$$A_{n-1}\cup C\approx A_n.$$

Clearly, C must contain only atoms that precede $\mathscr{A}_n$ in the ordering, whence by hypothesis $A_{n-1}\cup C$ must be an integer multiple of A_1, and consequently so must A_n, contrary to hypothesis.

We now want to show that each $\mathscr{A}_i$ is empty for $i>1$. Let

$$\mathscr{A}_1=\{A_1,\ldots,A_r\}.$$

Let us suppose, by way of contradiction, that some $\mathscr{A}_i$ is nonempty for $i>1$, and, in fact, let i be the least such i. So there is a B in $\mathscr{A}_i$, and we

know at once that

$$A_r < B,$$

because A_r is in $\mathscr{A}_1$. Now let

$$C = A_1 \cup \cdots \cup A_{r-1}.$$

Then by virtue of Axiom 2

$$A_r \cup C < B \cup C,$$

and thus by Axiom 5 there exists a D in $\mathscr{F}$ such that

$$A_r \cup C \cup D \approx B \cup C.$$

Without loss of generality, we may assume that

$$(A_r \cup C) \cap D = \emptyset,$$

because we may always take

$$D^* = D - (A_r \cup C) \in \mathscr{F}$$

and so

$$A_r \cup D \approx B.$$

Now if A is any atomic event and $A \subseteq D$ then A must be in $\mathscr{A}_0 \cup \mathscr{A}_1$ by virtue of our supposition about $\mathscr{A}_i$. Moreover, D must contain at least one such atomic event, and so for some j, $1 \leqslant j \leqslant r$,

$$A = A_j,$$

but then

$$(A_r \cup C) \cap D \neq \emptyset,$$

since $A_r \cup C$ is equal to $\cup \mathscr{A}_1$, and this last inequality contradicts an earlier equation.

Finally, it is easy to check that μ is an additive measure when we set $\mu(\emptyset) = 0$, and that from the construction for any A and B in $\mathscr{F}$

$$\mu(A) \geqslant \mu(B) \quad \text{if and only if} \quad A \geqslant B.$$

For the interpretation of subjective probability, we obtain a standard probability measure P by the normalization:

$$P(A) = \frac{\mu(A)}{\mu(X)}.$$

NOTES

[1] A consolidated list of references referred to in these introductory remarks or in the articles themselves is to be found at the end of the book.
[2] This last observation I owe to Robert Titiev.

1. A COMPARISON OF THE MEANING AND USES OF MODELS IN MATHEMATICS AND THE EMPIRICAL SCIENCES*

I. MEANING

Consider the following quotations:

A possible realization in which all valid sentences of a theory T are satisfied is called a model of T [Tarski, 1953, p. 11].

In the fields of spectroscopy and atomic structure, similar departures from classical physics took place. There had been accumulated an overwhelming mass of evidence showing the atom to consist of a heavy, positively charged nucleus surrounded by negative, particle-like electrons. According to Coulomb's law of attraction between electric charges, such a system will collapse at once unless the electrons revolve about the nucleus. But a revolving charge will, by virtue of its acceleration, emit radiation. A mechanism for the emission of light is thereby at once provided.

However, this mechanism is completely at odds with experimental data. The two major difficulties are easily seen. First, the atom in which the electrons revolve continually should emit light all the time. Experimentally, however, the atom radiates only when it is in a special, 'excited' condition. Second, it is impossible by means of this model to account for the occurrence of spectral lines of a single frequency (more correctly, of a narrow range of frequencies). The radiating electron of our model would lose energy; as a result it would no longer be able to maintain itself at the initial distance from the nucleus, but fall in toward the attracting center, changing its frequency of revolution as it falls. Its orbit would be a spiral ending in the nucleus. By electrodynamic theory, the frequency of the radiation emitted by a revolving charge is the same as the frequency of revolution, and since the latter changes, the former should also change. Thus our model is incapable of explaining the sharpness of spectral lines [Lindsay and Margenau, 1936, pp. 390–91].

The author (Gibbs) considers his task not as one of establishing physical theories directly, but as one of constructing statistic-mechanical models which have some analogies in thermodynamics and some other parts of physics; hence he does not hesitate to introduce some very special hypotheses of a statistical character [Khinchin, 1949, p. 4].

Thus, the model of rational choice as built up from pair-wise comparisons does not seem to suit well the case of rational behavior in the described game situation [Arrow, 1951, p. 21].

In constructing the model we shall assume that each variable is some kind of average or aggregate for members of the group. For example, D might be measured by locating the

* Reprinted from *Synthese* **12** (1960), 287–301.

opinions of group members on a scale, attaching numbers to scale positions and calculating the standard deviation of the members' opinions in terms of these numbers. Even the intervening variables, although not directly measured, can be thought of as averages of the values for individual members [Simon, 1957, p. 116].

This work on mathematical models for learning has not attempted to formalize any particular theoretical system of behavior; yet the influences of Guthrie and Hull are most noticeable. Compared with the older attempts at mathematical theorizing, the recent work has been more concerned with detailed analyses of data relevant to the models and with the design of experiments for directly testing quantitative predictions of the models [Bush and Estes, 1959, p. 3].

I shall describe ... various criteria used in adopting a mathematical model of an observed stochastic process ... For example, consider the number of cars that have passed a given point by time t. The first hypothesis is a typical mathematical hypothesis, suggested by the facts and serving to simplify the mathematics. The hypothesis is that the stochastic process of the model has independent increments ... The next hypothesis, that of stationary increments, states that, if $s < t$, the distribution of $x(t) - x(s)$ depends only on the time interval length $t - s$. This hypothesis means that we cannot let time run through both slack and rush hours. Traffic intensity must be constant.

The next hypothesis is that events occur one at a time. This hypothesis is at least natural to a mathematician. Because of limited precision in measurements it means nothing to an observer ... The next hypothesis is of a more quantitative kind, which also is natural to anyone who has seen Taylor's theorem. It is that the probability that at least one car should pass in a time interval of length h should be $ch + o(h)$ [Doob, 1960, p. 27].

The first of these quotations is taken from a book on mathematical logic, the next two from books on physics, the following three from works on the social sciences, and the last one from an article on mathematical statistics. Additional uses of the word 'model' could easily be collected in another batch of quotations. One of the more prominent senses of the word missing in the above quotations is the very common use in physics and engineering of 'model' to mean an actual physical model as, for example, in the phrases 'model airplane' and 'model ship'.

It may well be thought that it is impossible to put under one concept the several uses of the word 'model' exhibited by these quotations. It would, I think, be too much to claim that the word 'model' is being used in exactly the same sense in all of them. The quotation from Doob exhibits one very common tendency, namely, to confuse or to amalgamate what logicians would call the model and the theory of the model. It is very widespread practice in mathematical statistics and in the behavioral sciences to use the word 'model' to mean the set of quantitative assumptions of the theory, that is, the set of sentences which in a precise treatment

would be taken as axioms, or, if they are themselves not adequately exact, would constitute the intuitive basis for formulating a set of axioms. In this usage a model is a linguistic entity and is to be contrasted with the usage characterized by the definition from Tarski, according to which a model is a non-linguistic entity in which a theory is satisfied.

There is also a certain technical usage in econometrics of the word 'model' that needs to be noted. In the sense of the econometricians a model is a class of models in the sense of logicians, and what logicians call a model is called by econometricians a *structure.*

It does not seem to me that these are serious difficulties. I claim that the concept of model in the sense of Tarski may be used without distortion and as a fundamental concept in all of the disciplines from which the above quotations are drawn. In this sense I would assert that the meaning of the concept of model is the same in mathematics and the empirical sciences. The difference to be found in these disciplines is to be found in their use of the concept. In drawing this comparison between constancy of meaning and difference of use, the sometimes difficult semantical question of how one is to explain the meaning of a concept without referring to its use does not actually arise. When I speak of the meaning of the concept of a model I shall always be speaking in well-defined technical contexts and what I shall be claiming is that, given this technical meaning of the concept of model, mathematicians ask a certain kind of question about models and empirical scientists tend to ask another kind of question.

Perhaps it will be useful to defend this thesis about the concept of model by analyzing uses of the word in the above quotations. As already indicated, the quotation from Tarski represents a standard definition of 'model' in mathematical logic. Our references to models in pure mathematics will, in fact, be taken to refer to mathematical logic, that branch of pure mathematics explicitly concerned with the theory of models. The technical notion of possible realization used in Tarski's definition need not be expounded here. Roughly speaking, a possible realization of a theory is a set-theoretical entity of the appropriate logical type. For example, a possible realization of the theory of groups is any ordered couple whose first member is a set and whose second member is a binary operation on this set. The intuitive idea of a model as a possible realization in which a theory is satisfied is too familiar in the literature of mathematical logic to

need recasting. The important distinction that we shall need is that a theory is a linguistic entity consisting of a set of sentences and models are non-linguistic entities in which the theory is satisfied (an exact definition of theories is also not necessary for our uses here).

I would take it that the use of the notion of models in the quotation from Lindsay and Margenau could be recast in these terms in the following manner. The orbital theory of the atom is formulated as a theory. The question then arises, does a possible realization of this theory in terms of entities defined in close connection with experiments actually constitute a model of the theory, or, put another way which is perhaps simpler, do models of an orbital theory correspond well to data obtained from physical experiments with atomic phenomena? It is true that many physicists want to think of a model of the orbital theory of the atom as being more than a certain kind of set-theoretical entity. They envisage it as a very concrete physical thing built on the analogy of the solar system. I think it is important to point out that there is no real incompatibility in these two viewpoints. To define formally a model as a set-theoretical entity which is a certain kind of ordered tuple consisting of a set of objects and relations and operations on these objects is not to rule out the physical model of the kind which is appealing to physicists, for the physical model may be simply taken to define the set of objects in the set-theoretical model. Because of the importance of this point it may be well to illustrate it in somewhat greater detail. We may axiomatize classical particle mechanics in terms of the five primitive notions of a set P of particles, an interval T of real numbers corresponding to elapsed times, a position function s defined on the Cartesian product of the set of particles and the time interval, a mass function m defined on the set of particles, and a force function f defined on the Cartesian product of the set of particles, the time interval and the set of positive integers (the set of positive integers enters into the definition of the force function simply in order to provide a method of naming the forces). A possible realization of the axioms of classical particle mechanics, that is, of the theory of classical particle mechanics, is then an ordered quintuple $\mathscr{P} = \langle P, T, s, m, f \rangle$.

A model of classical particle mechanics is such an ordered quintuple. It is simple enough to see how an actual physical model in the physicist's sense of classical particle mechanics is related to this set-theoretical sense of models. We simply can take the set of particles to be in the case of the

solar system the set of planetary bodies. Another slightly more abstract possibility is to take the set of particles to be the set of centers of mass of the planetary bodies. This generally exemplifies the situation. The abstract set-theoretical model of a theory will have among its parts a basic set which will consist of the objects ordinarily thought to constitute the physical model (for a discussion of the axiomatic foundations of classical particle mechanics in greater detail along the lines just suggested see Suppes, 1957, Chap. 12).

In the preceding paragraph we have used the phrases, 'set-theoretical model' and 'physical model'. There would seem to be no use in arguing about which use of the word 'model' is primary or more appropriate in the empirical sciences. My own contention in this paper is that the set-theoretical usage is the more fundamental. The highly physically minded or empirically minded scientists who may disagree with this thesis and believe that the notion of a physical model is the more important thing in a given branch of empirical science may still agree with the systematic remarks I am making.

An historical illustration of this point is Kelvin's and Maxwell's efforts to find a mechanical model of electromagnetic phenomena. Without doubt they both thought of possible models in a literal physical sense, but it is not difficult to recast their published memoirs on this topic into a search for set-theoretical models of the theory of continuum mechanics which will account for observed electromagnetic phenomena. Moreover, it is really the formal part of their memoirs which has had permanent value. Ultimately it is the mathematical theory of Maxwell which has proved important, not the physical image of an ether behaving like an elastic solid.

The third quotation is from Khinchin's book on statistical mechanics, and the phrase, 'the author', refers to Gibbs whom Khinchin is discussing at this point. The use of the word 'model' in the quotation of Khinchin is particularly sympathetic to the set-theoretical viewpoint, for Khinchin is claiming that in his work on the foundations of statistical mechanics Gibbs was not concerned to appeal directly to physical reality or to establish true physical theories, but rather, to construct models or theories having partial analogies to the complicated empirical facts of thermodynamics and other branches of physics. Again in this quotation we have as in the case of Doob, perhaps even more directly, the tendency toward a

confusion of the logical type of theories and models, but again this does not create a difficulty. Anyone who has examined Gibb's work or Khinchin's will readily admit the ease and directness of formulating their work in such a fashion as to admit explicitly and exactly the distinction between theories and models made in mathematical logic. The abstractness of Gibb's work in statistical mechanics furnishes a particularly good example for applying the exact notion of model used by logicians, for there is not a direct and immediate tendency to think of Gibbs' statistical mechanical theories as being the theories of the one physical universe.

I think the following observation is empirically sound concerning the use of the word 'model' in physics. In old and established branches of physics which correspond well with the empirical phenomena they attempt to explain, there is only a slight tendency ever to use the word 'model'. The language of theory, experiment and common sense is blended into one realistic whole. Sentences of the theory are asserted as if they are the one way of describing the universe. Experimental results are described as if there were but one obvious language for describing them. Notions of common sense refined perhaps here and there are taken to be appropriately homogeneous with the physical theory. On the other hand, in those branches of physics which give as yet an inadequate account of the detailed physical phenomena with which they are concerned there is a much more frequent use of the word 'model'. Connotation of the use of the word is that the model is like a model of an airplane or ship. It simplifies drastically the true physical phenomena and only gives account of certain major or important aspects of it. Again, in such uses of the word 'model', it is to be emphasized that there is a constant interplay between the model as a physical or non-linguistic object and the model as a theory.

The quotation from Arrow which follows the one from Khinchin exemplifies in the social sciences this latter tendency in physics. Arrow, I would say, refers to the *model* of rational choice because the theory he has in mind does not give a very adequate description of the phenomena with which it is concerned but only provides a highly simplified schema. The same remarks apply fairly well to the quotation from Simon. In Simon we have an additional phenomenon exemplified which is very common in the social and behavioral sciences. A certain theory is stated in broad and general terms. Some qualitative experiments to test this theory are performed. Because of the success of these experiments scien-

tists interested in more quantitative and exact theories then turn to what is called 'the construction of a model' for the original theory. In the language of logicians, it would be more appropriate to say that rather than constructing a model they are interested in constructing a quantitative theory to match the intuitive ideas of the original theory.

In the quotation from Bush and Estes and the one from Doob, there is introduced an important line of thought which is, in fact, very closely connected with the concept of model as used by logicians. I am thinking here of the notion of model in mathematical statistics, the extensive literature on estimating parameters in models and testing hypotheses about them. In a statistical discussion of the estimation of the parameters of a model it is usually a trivial task to convert the discussion into one where the usage of terms is in complete agreement with that of logicians. The detailed consideration of statistical questions almost requires the consideration of models as mathematical or set-theoretical rather than simple physical entities. The question, "How well does the model fit the data?" is a natural one for statisticians and behavioral scientists. Only recently has it begun to be so for physicists, and it is still true that much of the experimental literature in physics is reported in terms of a rather medieval brand of statistics.

It may be felt by some readers that the main difficulty with the thesis being advanced in this paper is the lack of substantive examples in the empirical sciences. Such a reader would willingly admit that there are numerous examples of exactly formulated theories in pure mathematics, and thereby an exact basis is laid for precisely defining the models in which these theories are satisfied. But it might be held the situation is far different in any branch of empirical sciences. The formulation of theory goes hand in hand with the development of new experiments and new experimental techniques. It is the practice of empirical scientists, so it might be claimed, not to formulate theories in exact fashion but only to give them sufficient conceptual definiteness to make their connections with current experiments sufficiently clear to other specialists in the field.

He who seeks an exact characterization of the theory and thus of models in such branches of science as non-vertebrate anatomy, organic chemistry or nuclear physics is indeed barking up the wrong tree. In various papers and books I have attempted to provide some evidence against this view. In the final chapter of my *Introduction to Logic* I have formulated axio-

matically a theory of measurement and a version of classical particle mechanics which satisfy, I believe, the standards of exactness and clarity customary in the axiomatic formulation of theories in pure mathematics.

In Estes and Suppes (1959a) such a formulation is attempted for a branch of mathematical learning theory. In Rubin and Suppes (1954) an exact formulation of relativistic mechanics is considered and in Suppes (1959a)[1†] such a formulation of relativistic kinematics is given. These references are admittedly egocentric; it is also pertinent to refer to the work of Woodger (1957), Hermes (1938), Adams (1959), Debreu (1959), Noll (1959) and many others. Although it is not possible to pinpoint a reference to every branch of empirical science which will provide an exact formulation of the fundamental theory of the discipline, sufficient examples do now exist to make the point that there is no systematic difference between the axiomatic formulation of theories in well-developed branches of empirical science and in branches of pure mathematics.

By remarks made from a number of different directions, I have tried to argue that the concept of model used by mathematical logicians is the basic and fundamental concept of model needed for an exact statement of any branch of empirical science. To agree with this thesis it is not necessary to rule out or to deplore variant uses or variant concepts of model now abroad in the empirical sciences. As has been indicated, I am myself prepared to admit the significance and practical importance of the notion of physical model current in much discussion in physics and engineering. What I have tried to claim is that in the exact statement of the theory or in the exact analysis of data the notion of model in the sense of logicians provides the appropriate intellectual tool for making the analysis both precise and clear.

II. USES

The uses of models in pure mathematics are too well known to call for review here. The search in every branch of mathematics for representation theorems is most happily characterized in terms of models. To establish a representation theorem for a theory is to prove that there is a class of models of the theory such that every model of the theory is isomorphic to some member of this class. Examples now classical of such representation theorems are Cayley's theorem that every group is isomorphic to a group

of transformations and Stone's theorem that every Boolean algebra is isomorphic to a field of sets. Many important problems in mathematical logic are formulated in terms of classes of models. For a statement of many interesting results and problems readers are referred to Tarski (1954).

When a branch of empirical science is stated in exact form, that is, when the theory is axiomatized within a standard set-theoretical framework, the familiar questions raised about models of the theory in pure mathematics may also be raised for models of the precisely formulated empirical theory. On occasion such applications have philosophical significance. Many of the discussions of reductionism in the philosophy of science may best be formulated as a series of problems using the notion of a representation theorem. For example, the thesis that biology may be reduced to physics would be in many people's minds appropriately established if one could show that for any model of a biological theory it was possible to construct an isomorphic model within physical theory. The diffuse character of much biological theory makes any present attempt to realize such a program rather hopeless. An exact result of this character can be established for one branch of physics in relation to another. An instance of this is Adams' (1959) result that for a suitable characterization of rigid body mechanics every model of rigid body mechanics is isomorphic to a model defined within simple particle mechanics. But I do not want to give the impression that the application of models in the empirical sciences is mainly restricted to problems which interest philosophers of science. The attempt to characterize exactly models of an empirical theory almost inevitably yields a more precise and clearer understanding of the exact character of the theory. The emptiness and shallowness of many classical theories in the social sciences is well brought out by the attempt to formulate in any exact fashion what constitutes a model of the theory. The kind of theory which mainly consists of insightful remarks and heuristic slogans will not be amenable to this treatment. The effort to make it exact will at the same time reveal the weakness of the theory.

An important use of models in the empirical sciences is in the construction of Gedanken experiments. A Gedanken experiment is given precision and clarity by characterizing a model of the theory which realizes it. A standard and important method for arguing against the general plausibility of a theory consists of extending it to a new domain by constructing a model of the theory in that domain. This aspect of the use of models need not

however be restricted to Gedanken experiments. A large number of experiments in psychology are designed with precisely this purpose in mind, that is, with the extension of some theory to a new domain, and the experimenter's expectation is that the results in this domain will not be those predicted by the theory.

It is my own opinion that a more exact use of the theory of models in the discussion of Gedanken experiments would often be of value in various branches of empirical science. A typical example would be the many discussions centering around Mach's proposed definition of the mass of bodies in terms of their mutually induced accelerations. Because of its presumed simplicity and beauty this definition is frequently cited. Yet from a mathematical standpoint and any exact theory of models of the theory of mechanics, it is not a proper definition at all. For a very wide class of axiomatizations of classical particle mechanics, it may be proved by Padoa's principle that a proper definition of mass is not possible. Moreover, if the number of interacting bodies is greater than seven, a knowledge of the mutually induced acceleration of the particles is not sufficient for unique determination of the ratios of the masses of the particles. The fundamental weakness of Mach's proposal is that he did not seem to realize a definition in the theory cannot be given for a single model, but must be appropriate for every model of the theory in order to be acceptable in the standard sense.

Another significant use of models, perhaps peculiar to the empirical sciences, is in the analysis of the relation between theory and experimental data. The importance of models in mathematical statistics has already been mentioned. The homogeneity of the concept of model used in that discipline with that adopted by logicians has been remarked upon. The striking thing about the statistical analysis of data is that it is shot through and through with the kind of comparison of models that does not ordinarily arise in pure mathematics. Generally speaking, in some particular branch of pure mathematics, the comparison of models involves comparison of two models of the same logical type. The representation theorems mentioned earlier provide good examples. Even in the case of embedding theorems, which establish that models of one sort may be extended in a definite manner to models of another sort, the logical type of the two models is very similar. The situation is often radically different in the comparison of theory and experiment. On the one hand, we may have a

rather elaborate set-theoretical model of the theory which contains continuous functions or infinite sequences, and, on the other hand, we have highly finitistic set-theoretical models of the data. It is perhaps necessary to explain what I mean by 'models of the data'. The maddeningly diverse and complex experience which constitutes an experiment is not the entity which is directly compared with a model of a theory. Drastic assumptions of all sorts are made in reducing the experimental experience, as I shall term it, to a simple entity ready for comparison with a model of the theory.

Perhaps it would be well to conclude with an example illustrating these general remarks about models of the data. I shall consider the theory of linear response models set forth in Estes and Suppes (1959a). For simplicity, let us assume that on every trial the organism can make exactly one of two responses, A_1 or A_2, and after each response it receives a reinforcement, E_1 or E_2, of one of the two possible responses. A learning parameter Θ, which is a real number such that $0<\Theta\leqslant 1$, describes the rate of learning in a manner to be made definite in a moment. A possible realization of the theory is an ordered triple $\mathscr{X}=\langle X, P, \Theta\rangle$ of the following sort. X is the set of all sequences of ordered pairs such that the first member of each pair is an element of some set A and the second member an element of some set B, where A and B each have two elements. Intuitively, the set A represents the two possible responses and the set B the two possible reinforcements. P is a probability measure on the Borel field of cylinder sets of X, and Θ is a real number as already described. (Actually there is a certain arbitrariness in the characterization of possible realizations of theories whose models have a rather complicated set-theoretical structure, but this is a technical matter into which we shall not enter here.) To define the models of the theory, we need a certain amount of notation. Let $A_{i,n}$ be the event of response A_i on trial n, $E_{j,n}$ the event of reinforcement E_j on trial n, where $i, j=1, 2$, and for x in X let x_n be the equivalence class of all sequences in X which are identical with x through trial n. A possible realization of the linear response theory is then a model of the theory if the following two axioms are satisfied in the realization:

Axiom 1: *If* $P(E_{i,n}A_{i',n}x_{n-1})>0$ *then*

$$P(A_{i,n+1} \mid E_{i,n}A_{i',n}x_{n-1}) = (1-\Theta)\,P(A_{i,n} \mid x_{n+1}) + \Theta.$$

Axiom 2: If $P(E_{j,n}A_{i',n}x_{n-1})>0$ *and* $i\neq j$ *then*

$$P(A_{i,n+1} \mid E_{j,n}A_{i',n}x_{n-1}) = (1-\Theta)\,P(A_{i,n} \mid x_{n-1}).$$

As is clear from the two axioms, this linear response theory is intuitively very simple. The first axiom just says that when a response is reinforced the probability of making that response on the next trial is increased by a simple linear transformation. And the second axiom says that if some other response is reinforced, the probability of making the response is decreased by a second linear transformation. In spite of the simplicity of this theory, it gives a reasonably good account of a number of experiments, and from a mathematical standpoint, it is by no means trivial to characterize asymptotic properties of its models.

The point of concern here, however, is to relate models of this theory to models of the data. Again for simplicity, let us consider the case of simple noncontingent reinforcement. On every trial, the probability of an E_1 reinforcement, independent of any preceding events, is π. The experimenter decides on an experiment of, say, 400 trials for each subject, and he uses a table of random numbers to construct for each subject a finite reinforcement sequence of 400 trials. The experimental apparatus might be described as follows.

The subject sat at a table of standard height. Mounted vertically on the table top was a 125 cm wide by 75 cm high black panel placed 50 cm from the end of the table. The experimenter sat behind the panel, out of view of the subject. The apparatus, as viewed by the subject, consisted of two silent operating keys mounted 20 cm apart on the table top and 30 cm from the end of the table; upon the panel, three milk-glass panel lights were mounted. One of these lights, which served as the signal for the subject to respond, was centered between the keys at a height of 42 cm from the table top. Each of the two remaining lights, the reinforcing signals, was at a height of 28 cm directly above one of the keys. On all trials the signal light was lighted for 3.5 sec; the time between successive signal exposures was 10 sec. The reinforcing light followed the cessation of the signal light by 1.5 sec and remained on for 2 sec.

The model of the data incorporates very little of this description. On each trial the experimenter records the response made and the reinforcement given. Expressions on the subject's face, the movement of his limbs, and in the present experiment even how long he takes to make the choice of

which key to punch, are ignored and not recorded. Even though it is clear exactly what the experimenter records, the notion of a possible realization of the data is not unambiguously clear. As part of the realization it is clear we must have a finite set D consisting of all possible finite sequences of length 400 where, as previously, the terms of the sequences are ordered couples, the first member of each couple being drawn from some pair set A and the second member from some pair set B. If a possible realization consisted of just such a set D, then any realization would also be a model of the data. But it seems natural to include in the realization a probability measure P on the set of all subsets of D, for by this means we may impose upon models of the data the experimental schedule of reinforcement.

In these terms, a possible realization of the data is an ordered couple $\mathscr{D}=\langle D, P\rangle$ and, for the case of noncontingent reinforcement, a realization is a model if and only if the probability measure P has the property of being a Bernouilli distribution with parameter π on the second members of the terms of the finite sequences in D, i.e., if and only if for every n from 1 to 400, $P(E_{1,n} \mid x_{n-1})=\pi$ when $P(x_n)>0$.

Unfortunately, there are several respects in which this characterization of models of the data may be regarded as unsatisfactory. The main point is that the models are still too far removed even from a highly schematized version of the experiment. No account has been taken of the standard practice of randomization of response A_1 as the left key for one subject and the right key for another. Secondly, a model of the data, as defined above, contains 2^{400} possible response sequences. An experiment that uses 30 or 40 subjects yields but a small sample of these possibilities. A formal description of this sample is easily given, and it is easily argued that the 'true' model of the data is this actual sample, not the much larger model just defined. Involved here is the formal relation between the three entities labeled by statisticians the 'sample', the 'population', and the 'sample space'. A third difficulty is connected with the probability measure that I have included as part of the model of the data. It is certainly correct to point out that a model of the data is hardly appropriately experimental if there is no indication given of how the probability distribution on reinforcements is produced.

It is not possible in this paper to enter into a discussion of these criticisms or the possible formal modifications in models of the data which might be made to meet them. My own conviction is that the set-theoretical

concept of model is a useful tool for bringing formal order into the theory of experimental design and analysis of data. The central point for me is the much greater possibility than is ordinarily realized of developing an adequately detailed formal theory of these matters.

NOTE

[1†] Article 12 in this volume.

2. MODELS OF DATA*

I. INTRODUCTION

To nearly all the members of this Congress, the logical notion of a model of a theory is too familiar to need detailed review here. Roughly speaking, a model of a theory may be defined as a possible realization in which all valid sentences of the theory are satisfied, and a possible realization of the theory is an entity of the appropriate set-theoretical structure. For instance, we may characterize a possible realization of the mathematical theory of groups as an ordered couple whose first member is a nonempty set and whose second member is a binary operation on this set. A possible realization of the theory of groups is a model of the theory if the axioms of the theory are satisfied in the realization, for in this case (as well as in many others), the valid sentences of the theory are defined as those sentences which are logical consequences of the axioms. To provide complete mathematical flexibility I shall speak of theories axiomatized within general set theory by defining an appropriate set-theoretical predicate (e.g., 'is a group') rather than of theories axiomatized directly within first-order logic as a formal language. For the purposes of this paper, this difference is not critical. In the set-theoretical case, it is convenient sometimes to speak of the appropriate predicate's being satisfied by a possible realization. But whichever sense of formalization is used, essentially the same logical notion of model applies.[1]

It is my opinion that this notion of model is the fundamental one for the empirical sciences as well as mathematics. To assert this is not to deny a place for variant uses of the word 'model' by empirical scientists, as, for example, when a physicist talks about a physical model, or a psychologist refers to a quantitative theory of behavior as a mathematical model. On this occasion I do not want to argue for this fundamental character of the logical notion of model, for I tried to make out a detailed case at a collo-

* Reprinted from *Logic, Methodology and Philosophy of Science: Proceedings of the* 1960 *International Congress* (ed. by E. Nagel, P. Suppes, and A. Tarski), Stanford University Press, Stanford, Calif., 1962, pp. 252–261.

quium in Utrecht last January, also sponsored by the International Union of History and Philosophy of Science (Suppes, 1960b).[2†] Perhaps the most persuasive argument which might be singled out for mention here is that the notion of model used in any serious statistical treatment of a theory and its relation to experiment does not differ in any essential way from the logical notion of model.

The focus of the present paper is closely connected to the statistical analysis of the empirical adequacies of theories. What I want to try to show is that exact analysis of the relation between empirical theories and relevant data calls for a hierarchy of models of different logical type. Generally speaking, in pure mathematics the comparison of models involves comparison of two models of the same logical type, as in the assertion of representation theorems. A radically different situation often obtains in the comparison of theory and experiment. Theoretical notions are used in the theory which have no direct observable analogue in the experimental data. In addition, it is common for models of a theory to contain continuous functions or infinite sequences although the confirming data are highly discrete and finitistic in character.

Perhaps I may adequately describe the kind of ideas in which I am interested in the following way. Corresponding to possible realizations of the theory, I introduce possible realizations of the data. Models of the data of an experiment are then defined in the customary manner in terms of possible realizations of the data. As should be apparent, from a logical standpoint possible realizations of data are defined in just the same way as possible realizations of the theory being tested by the experiment from which the data come. The precise definition of models of the data for any given experiment requires that there be a theory of the data in the sense of the experimental procedure, as well as in the ordinary sense of the empirical theory of the phenomena being studied.

Before analyzing some of the consequences and problems of this viewpoint, it may be useful to give the ideas more definiteness by considering an example.

II. EXAMPLE FROM LEARNING THEORY

I have deliberately chosen an example from learning theory because it is conceptually simple, mathematically non-trivial and thoroughly probabilistic. More particularly, I consider linear response theory as developed

by Estes and myself (1959a). To simplify the presentation of the theory in an inessential way, let us assume that on every trial the organism in the experimental situation can make exactly one of two responses, A_1 or A_2, and after each response it receives a reinforcement, E_1 or E_2, of one of the two possible responses. A possible experimental outcome in the sense of the theory is an infinite sequence of ordered pairs, where the nth term of the sequence represents the observed response – the first member of the pair – and the actual reinforcement – the second member of the pair – on trial n of the experiment.

A possible realization of the theory is an ordered triple $\mathscr{X}=\langle X, P, \theta\rangle$ of the following sort. The set X is the set of all sequences of ordered pairs such that the first member of each pair is an element of some set A, and the second member an element of some set B, where A and B each have two elements. The set A represents the two possible responses and the set B the two possible reinforcements. The function P is a probability measure on the smallest Borel field containing the field of cylinder sets of X; and θ, a real number in the interval $0<\theta\leqq 1$, is the learning parameter. (Admittedly, for theories whose models have a rather complicated set-theoretical structure, the definition of possible realization is at points arbitrary, but this is not an issue which affects in any way the development of ideas central to this paper.)

There are two obvious respects in which a possible realization of the theory cannot be a possible realization of experimental data. The first is that no actual experiment can include an infinite number of discrete trials. The second is that the parameter θ is not directly observable and is not part of the recorded data.

To pursue further relations between theory and experiment, it is necessary to state the axioms of the theory, i.e., to define models of the theory. For this purpose a certain amount of notation is needed. Let $A_{i,n}$ be the event of response A_i on trial n, $E_{j,n}$ the event of reinforcement E_j on trial n, where $i, j=1, 2$, and for x in X let x_n be the equivalence class of all sequences in X which are identical with x through trial n. A possible realization of the linear response theory is then a model of the theory if the following two axioms are satisfied in the realization:

Axiom 1: *If* $P(E_{i,n}A_{i',n}x_{n-1})>0$, *then*

$$P(A_{i,n+1} \mid E_{i,n}A_{i',n}x_{n-1}) = (1-\theta)\,P(A_{i,n} \mid x_{n-1}) + \theta.$$

Axiom 2: *If* $P(E_{j,n}A_{i',n}x_{n-1})>0$ *and* $i\neq j$, *then*

$$P(A_{i,n+1} \mid E_{j,n}A_{i',n}x_{n-1}) = (1-\theta)\,P(A_{i,n} \mid x_{n-1}).$$

The first axiom asserts that when a response is reinforced, the probability of making that response on the next trial is increased by a simple linear transformation. The second axiom asserts that when a different response is reinforced, the probability of making the response is decreased by a second linear transformation. To those who are concerned about the psychological basis of this theory, it may be remarked that it is derivable from a much more complicated theory that assumes processes of stimulus sampling and conditioning. The linear response theory is the limiting case of the stimulus sampling theory as the number of stimuli approaches infinity.

For still greater definiteness, it will be expedient to consider a particular class of experiments to which the linear response theory has been applied, namely, those experiments with simple contingent reinforcement schedules. On every trial, if an A_1 response is made, the probability of an E_1 reinforcement is π_1, independent of the trial number and other preceding events. If an A_2 response is made, the probability of an E_2 reinforcement is π_2. Thus, in summary for every n,

$$P(E_{1,n} \mid A_{1,n}) = \pi_1 = 1 - P(E_{2,n} \mid A_{1,n}),$$
$$P(E_{2,n} \mid A_{2,n}) = \pi_2 = 1 - P(E_{1,n} \mid A_{2,n}).$$

This characterization of simple contingent reinforcement schedules has been made in the language of the theory, as is necessary in order to compute theoretical predictions. This is not possible for the finer details of the experiment. Let us suppose the experimenter decides on 600 trials for each subject. A brief description (cf. Suppes and Atkinson, 1960, pp. 81–83) of the experimental apparatus might run as follows.

The subject sits at a table of standard height. Mounted vertically in front of the subject is a large opaque panel. Two silent operating keys (A_1 and A_2 responses) are mounted at the base of the panel 20 cm apart. Three milk-glass panel lights are mounted on the panel. One of these lights, which serves as the signal for the subject to respond, is centered between the keys at the subject's eye level. Each of the other two lights, the reinforcing events E_1 and E_2, is mounted directly above one of the keys. On all trials the signal light is on for 3.5 sec; the time between successive signal exposures is 10 sec. A reinforcing light comes on 1.5 sec after the cessation of the signal light and remains on for 2 sec.

It is not surprising that this description of the apparatus is not incorporated in any direct way at all into the theory. The important point is to take the linear response theory and this description as two extremes between which a hierarchy of theories and their models is to be fitted in a detailed analysis.

In the class of experiments we are considering, the experimenter records only the response made and reinforcement given on each trial. This suggests the definition of the possible realizations of the theory that is the first step down from the abstract level of the linear response theory itself. This theory I shall call the *theory of the experiment*, which term must not be taken to refer to what statisticians call the theory of experimental design – a topic to be mentioned later. A possible realization of the theory of the experiment is an ordered couple $\mathscr{Y} = \langle Y, P \rangle$, where (i) Y is a finite set consisting of all possible finite sequences of length 600 with, as previously, the terms of the sequences being ordered pairs, the first member of each pair being drawn from some pair set A and correspondingly for the second members, and (ii) the function P is a probability measure on the set of all subsets of Y.

A possible realization $\mathscr{Y} = \langle Y, P \rangle$ of the theory of the experiment is a model of the theory if the probability measure P satisfies the defining condition for a simple contingent reinforcement schedule. Models of the experiment thus defined are entities still far removed from the actual data. The finite sequences that are elements of Y may indeed be used to represent any possible experimental outcome, but in an experiment with, say, 40 subjects, the observed 40 sequences are an insignificant part of the 4^{600} sequences in Y. Consequently, a model closer to the actual situation is needed to represent the actual conditional relative frequencies of reinforcement used.

The appropriate realization for this purpose seems to be an N-tuple Z of elements from Y, where N is the number of subjects in the experiment. An N-tuple rather than a subset of Y is selected for two reasons. The first is that if a subset is selected there is no direct way of indicating that two distinct subjects had exactly the same sequence of responses and reinforcements – admittedly a highly improbable event. The second and more important reason is that the N-tuple may be used to represent the time sequence in which subjects were run, a point of some concern in considering certain detailed questions of experimental design. It may be noted that in

using an N-tuple as a realization of the data rather than a more complicated entity that could be used to express the actual times at which subjects were run in the experiment, we have taken yet another step of abstraction and simplification away from the bewilderingly complex complete experimental phenomena.[3]

The next question is, When is a possible realization of the data a model of the data? The complete answer, as I see it, requires a detailed statistical theory of goodness of fit. Roughly speaking, an N-tuple realization is a model of the data if the conditional relative frequencies of E_1 and E_2 reinforcements fit closely enough the probability measure P of the model of the experiment. To examine in detail statistical tests for this goodness of fit would be inappropriate here, but it will be instructive of the complexities of the issues involved to outline some of the main considerations. The first thing to note is that no single simple goodness-of-fit test will guarantee that a possible realization Z of the data is an adequate model of the data. The kinds of problems that arise are these: (i) (*Homogeneity*) Are the conditional relative frequencies (C.R.F.) of reinforcements approximately π_i or $1-\pi_i$, as the case may be, for each subject? To answer this we must compare members of the N-tuple Z. (ii) (*Stationarity*) Are the C.R.F. of reinforcements constant over trials? To answer this practically we sum over subjects, i.e., over members of Z, to obtain sufficient data for a test. (iii) (*Order*) Are the C.R.F. of reinforcements independent of preceding reinforcements and responses? To answer this we need to show that the C.R.F. define a zero order process – that serial correlations of all order are zero. Note, of course, that the zero order is with respect to the conditional events E_i given A_j, for $i, j=1, 2$. These three questions are by no means exhaustive; they do reflect central considerations. To indicate their essentially formal character, it may be helpful to sketch their formulation in a relatively classical statistical framework. Roughly speaking, the approach is as follows. For each possible realization Z of the data, we define a statistic $T(Z)$ for each question. This statistic is a random variable with a probability distribution – preferably a distribution that is (asymptotically) independent of the actual C.R.F. under the null hypothesis that Z is a model of the data. In statistical terminology, we "accept" the null hypothesis if the obtained value of the statistic $T(Z)$ has a probability equal to or greater than some significance level α on the assumption that indeed the null hypothesis is true.

For the questions of homogeneity, stationarity, and order stated above, maximum likelihood or chi-square statistics would be appropriate. There is not adequate space to discuss details, but these statistics are standard in the literature. For the purposes of this paper, it is not important that some subjectivists like L. J. Savage might be critical of the unfettered use of such classical tests. A more pertinent caveat is that joint satisfaction of three statistical tests (by 'satisfaction' I mean acceptance of the null hypothesis with a level of significance $\geqq 0.05$) corresponding to the three questions does not intuitively seem completely sufficient for a possible realization Z to be a model of the data.[4] No claim for completeness was made in listing these three, but it might also be queried as to what realistic possibility there is of drawing up a finite list of statistical tests which may be regarded as jointly sufficient for Z to be a model of the data. A skeptical non-formalistic experimenter might claim that given any usable set of tests he could produce a conditional reinforcement schedule that would satisfy the tests and yet be intuitively unsatisfactory. For example, suppose the statistical tests for order were constructed to look at no more than fourth-order effects, the skeptical experimenter could then construct a possible realization Z with a non-random fifth-order pattern. Actually the procedure used in well-constructed experiments makes such a dodge rather difficult. The practice is to obtain the C.R.F. from some published table of random numbers whose properties have been thoroughly investigated by a wide battery of statistical tests. From the systematic methodological standpoint, it is not important that the experimenter himself perform the tests on Z.

On the other hand, in the experimental literature relevant to this example, it is actually the case that greater care needs to be taken to guarantee that a possible realization Z of the data is indeed a model of the data for the experiment at hand. A typical instance is the practice of restricted randomization. To illustrate, if $P(E_{1,n} \mid A_{1,n})=0.6$, then some experimenters would arrange that in every block of 10 A_1 responses, exactly 6 are followed by E_1 reinforcements, a result that should have a probability of approximately zero for a large number of trials.[5]

The most important objection of the skeptical experimenter to the importance of models of the data has not yet been examined. The objection is that the precise analysis of these models includes only a small portion of the many problems of experimental design. For example, by most

canons of experimental design, the assignment of A_1 to the left (or to the right) for every subject would be a mistake. More generally, the use of an experimental room in which there was considerably more light on the left side of subjects than on the right would be considered mistaken. There is a difference, however, in these two examples. The assignment of A_1 to the left or right for each subject is information that can easily be incorporated into models of the data – and requirements of randomization can be stated. Detailed information about the distribution of physical parameters characterizing the experimental environment is not a simple matter to incorporate in models of data and is usually not reported in the literature; roughly speaking, some general *ceteris paribus* conditions are assumed to hold.

The characterization of models of data is not really determined, however, by relevant information about experimental design which can easily be formalized. In one sense there is scarcely any limit to information of this kind; it can range from phases of the moon to *I.Q.* data on subjects.

The central idea, corresponding well, I think, to a rough but generally clear distinction made by experimenters and statisticians, is to restrict models of the data to those aspects of the experiment which have a parametric analogue in the theory. A model of the data is designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory. The point I want to make is not as simple or as easily made precise as I could wish. Table I is meant to

TABLE I

Hierarchy of theories, models, and problems

Theory of	Typical problems
Linear response models	Estimation of θ, goodness of fit to models of data
Models of experiment	Number of trials, choice of experimental parameters
Models of data	Homogeneity, stationarity, fit of experimental parameters
Experimental design	Left-right randomization, assignment of subjects
Ceteris paribus conditions	Noises, lighting, odors, phases of the moon

indicate a possible hierarchy of theories, models, and problems that arise at each level to harass the scientist. At the lowest level I have placed *ceteris paribus* conditions. Here is placed every intuitive consideration of experimental design that involves no formal statistics. Control of loud noises, bad odors, wrong times of day or season go here. At the next level formal problems of experimental design enter, but of the sort that far exceed the limits of the particular theory being tested. Randomization of A_1 as the left or right response is a problem for this level, as is random assignment of subjects to different experimental groups. All the considerations that enter at this level can be formalized, and their relation to models of the data, which are at the next level, can be made explicit – in contrast to the seemingly endless number of unstated *ceteris paribus* conditions.

At the next level, models of the experiment enter. They bear the relation to models of the data already outlined. Finally, at the top of the hierarchy are the linear response models, relatively far removed from the concrete experimental experience. It is to be noted that linear response models are related directly to models of the data, without explicit consideration of models of the experiment. Also worth emphasizing once again is that the criteria for deciding if a possible realization of the data is a model of the data in no way depend upon its relation to a linear response model. These criteria are to determine if the experiment was well run, not to decide if the linear response theory has merit.

The dependence is actually the other way round. Given a model of the data, we ask if there is a linear response model to which it bears a satisfactory goodness-of-fit relation. The rationale of a maximum likelihood estimate of θ is easily stated in this context: given the experimental parameters π_1 and π_2 we seek that linear response model, i.e., the linear response model with learning parameter $\hat{\theta}$, which will maximize the probability of the observed data, as given in the model of the data.

It is necessary at this point to break off rather sharply discussion of this example from learning theory, but there is one central point that has not been sufficiently mentioned. The analysis of the relation between theory and experiment must proceed at every level of the hierarchy shown in Table I. Difficulties encountered at all but the top level reflect weaknesses in the experiment, not in the fundamental learning theory. It is unfortunate that it is not possible to give here citations from the experimental literature

of badly conceived or poorly executed experiments that are taken to invalidate the theory they presume to test, but in fact do not.

III. THE THEORY OF MODELS IN THE EMPIRICAL SCIENCES

I began by saying that I wanted to try to show that exact analysis of the relation between empirical theories and relevant data calls for a hierarchy of models of different logical type. The examination of the example from learning theory was meant to exhibit some aspects of this hierarchy. I would like to conclude with some more general remarks that are partially suggested by this example.

One point of concern on my part has been to show that in moving from the level of theory to the level of experiment we do not need to abandon formal methods of analysis. From a conceptual standpoint, the distinction between pure and applied mathematics is spurious – both deal with set-theoretical entities, and the same is true of theory and experiment.

It is a fundamental contribution of modern mathematical statistics to have recognized the explicit need of a model in analyzing the significance of experimental data. It is a paradox of scientific method that the branches of empirical science that have the least substantial theoretical developments often have the most sophisticated methods of evaluating evidence. In such highly empirical branches of science, a large hierarchy of models is not necessary, for the theory being tested is not a theory with a genuine logical structure, but a collection of heuristic ideas. The only models needed are something like the models of the experiment and models of the data discussed in connection with the example from learning theory.

Present statistical methodology is less adequate when a genuine theory is at stake. The hierarchy of models outlined in our example corresponds in a very rough way to statisticians' concepts of a sample space, a population, and a sample. It is my own opinion that the explicit and exact use of the logical concept of model will turn out to be a highly useful device in clarifying the theory of experimental design, which many statisticians still think of as an "art" rather than a "science". Limitations of space have prevented working out the formal relations between the theory of experimental design and the theory of models of the data, as I conceive it.

However, my ambitions for the theory of models in the empirical

sciences are not entirely such practical ones. One of the besetting sins of philosophers of science is to overly simplify the structure of science. Philosophers who write about the representation of scientific theories as logical calculi then go on to say that a theory is given empirical meaning by providing interpretations or coordinating definitions for some of the primitive or defined terms of the calculus. What I have attempted to argue is that a whole hierarchy of models stands between the model of the basic theory and the complete experimental experience. Moreover, for each level of the hierarchy, there is a theory in its own right. Theory at one level is given empirical meaning by making formal connections with theory at a lower level. Statistical or logical investigation of the relations between theories at these different levels can proceed in a purely formal, set-theoretical manner. The more explicit the analysis, the less place there is for non-formal considerations. Once the empirical data are put in canonical form (at the level of models of data in Table I), every question of systematic evaluation that arises is a formal one. It is important to notice that the questions to be answered are formal, but not mathematical – not mathematical in the sense that their answers do not in general follow from the axioms of set theory (or some other standard framework for mathematics). It is precisely the fundamental problem of scientific method to state the principles of scientific methodology that are to be used to answer these questions – questions of measurement, of goodness of fit, of parameter estimation, of identifiability, and the like. The principles needed are entirely formal in character in the sense that they have as their subject matter set-theoretical models and their comparison. Indeed, the line of argument I have tried to follow in this paper leads to the conclusion that the only systematic results possible in the theory of scientific methodology are purely formal, but a general defense of this conclusion cannot be made here.

NOTES

[1] For a detailed discussion of axiomatization of theories within set theory, see Suppes (1957, Chap. 12).

[2†] Article 1 in this volume.

[3] The exact character of a model $\mathscr{Y}$ of the experiment and a model Z of the data is not determined uniquely by the experiment. It would be possible, for instance, to define $\mathscr{Y}$ in terms of N-tuples.

[4] For use at this point, a more explicit definition of models of the data would run as

follows. Z is an N-fold model of the data for experiment $\mathscr{Y}$ if and only if there is a set Y and a probability measure P on subsets of Y such that $\mathscr{Y} = \langle Y, P \rangle$ is a model of the theory of the experiment, Z is an N-tuple of elements of Y, and Z satisfies the statistical tests of homogeneity, stationarity, and order. A fully formal definition would spell out the statistical tests in exact mathematical detail. For example, a chi-square test of homogeneity for E_1 reinforcements following A_1 responses would be formulated as follows. Let N_j be the number of A_1 responses (excluding the last trial) for subject j, i.e., as recorded in Z_j – the jth member of the N-tuple Z, and let v_j be the number of E_1 reinforcements following A_1 responses for subject j. Then

$$\chi_H^2(Z) = \sum_{j=1}^{N} \frac{(v_j - N_j\pi_1)^2}{N_j\pi_1} + \frac{(N_j - v_j - N_j(1 - \pi_1))^2}{N_j(1 - \pi_j)}$$

$$= \sum_{j=1}^{N} \frac{(v_j - N_j\pi_1)^2}{N_j\pi_1(1 - \pi_1)},$$

and this χ^2 has N degrees of freedom. If the value $\chi_H^2(Z)$ has probability greater than 0.05 the null hypothesis is accepted, i.e., with respect to homogeneity Z is satisfactory. [5] To emphasize that conceptually there is nothing special about this particular example chosen from learning theory, it is pertinent to remark that much more elaborate analyses of sources of experimental error are customary in complicated physical experiments. In the literature of learning theory it is as yet uncommon to report the kind of statistical tests described above which play a role analogous to the physicists' summary of experimental errors.

3. A SET OF INDEPENDENT AXIOMS FOR EXTENSIVE QUANTITIES*[1]

I. INTRODUCTION

The modern viewpoint on quantities goes back at least to Newton's *Universal Arithmetick*. Newton asserts that the relation between any two quantities of the same kind can be expressed by a real, positive number.[2] In 1901, O. Hoelder gave a set of 'Axiome der Quantitaet', which are sufficient to establish an isomorphism between any realization of his axioms and the additive semigroup of all positive real numbers. Related work of Hilbert, Veronese and others is indicative of a general interest in the subject of quantities in the abstract on the part of mathematicians of this period. During the last thirty years, from another direction, philosophers of science have become interested in the logical analysis of empirical procedures of measurement.[3] The interests of these two groups overlap insofar as the philosophers have been concerned to state the formal conditions which must be satisfied by empirical operations measuring some characteristic of physical objects (or other entities). Philosophers have divided quantities (that is, entities or objects considered relatively to a given characteristic, such as mass, length or hardness) into two kinds. Intensive quantities are those which can merely be arranged in a serial order; extensive quantities are those for which a "natural" operation of addition or combination can also be specified. Another, more exact, way of making a distinction of this order is to say that intensive quantities are quantities to which numbers can be assigned uniquely up to a monotone transformation, and extensive quantities are quantities to which numbers can be assigned uniquely up to a similarity transformation (that is, multiplication by a positive constant).[4] This last condition may be said to be the criterion of formal adequacy for a system of extensive quantities.

Hoelder's system satisfies this criterion of adequacy for extensive quantities, and his system has in fact been used by some philosophers

* Reprinted from *Portugaliae Mathematica* **10** (1951), 163–172.

(see, for instance, Nagel, 1931) in methodological studies of measurement. But from the methodological standpoint, there are at least two serious defects in Hoelder's system. The first is that he does not axiomatize the relation designated by '=' but instead, treats it as the logical relation of identity. However, it is ordinarily admitted that two distinct line segments may have the same numerical length or two distinct physical objects the same mass; and consequently, '=' should designate an equivalence relation which is not the logical one of identity.[5] The second defect of Hoelder's system is that it is too strong for a general characterization of extensive quantities. His system is categorical in the sense that any two realizations of it are isomorphic, and, in addition, isomorphic to the additive semigroup of all positive real numbers. But these requirements are certainly too demanding, for it is intuitively obvious that a set of extensive quantities need not even have the density property of the rational numbers. The masses of objects in a given set could, for instance, surely be determined, even if relatively to some unit, the mass of every object in the set were a positive integer.

The purpose of the present paper is to present a formally adequate system of axioms for extensive quantities, from which these two defects are eliminated. In addition, proofs of the independence of the axioms and the primitives of the system are given.

II. AXIOMS

We consider a system consisting of a nonempty set K of arbitrary elements $x, y, z \ldots$, a binary relation Q defined over K, and a binary function $*$ defined over K. Such a system may be regarded as the ordered triple $\langle K, Q, * \rangle$. Variables 'm', 'n', etc., take as values the natural numbers; the notation 'nx' is defined in the usual recursive way: $1x = x$, and $nx = (n-1)x * x$.

DEFINITION: *A system $\langle K, Q, * \rangle$ will be said to be a system of extensive quantities if it satisfies the following seven axioms:*

*A*I. *If x, y and z are in K, and if $x\ Q\ y$ and $y\ Q\ z$, then $x\ Q\ z$.*

*A*II. *If x and y are in K, then $x * y$ is in K.*

*A*III. *If x, y and z are in K, then $(x * y) * z\ Q\ x * (y * z)$.*

*A*IV. *If x, y and z are in K and $x\ Q\ y$, then $x * z\ Q\ z * y$.*

*A*V. *If x and y are in K and not $x\ Q\ y$, then there is a z in K such that $x\ Q\ y * z$ and $y * z\ Q\ x$.*

*A*VI. *If x and y are in K, then not $x*y\ Q\ x$.*

*A*VII. *If x and y are in K and $x\ Q\ y$, then there is a number n such that $y\ Q\ nx$.*

If $*$ is interpreted as $+$ and Q as $\leqslant$, it may easily be seen that these axioms are satisfied by any additive semigroup of positive numbers closed under subtraction of smaller numbers from larger ones. The formal adequacy of these axioms, in the sense defined in Section I, is established in Section V; that they are mutually independent is established in Section VI.

III. ELEMENTARY THEOREMS

In the statement and proof of theorems which follow from the seven axioms just given, the statement of the condition that elements be in K is omitted for brevity. The proofs are all elementary in character and are therefore considerably abbreviated.

THEOREM 1: $x\ Q\ x$.

Proof: Assume: not $x\ Q\ x$. Then, by A.V, there is a z such that $x*z\ Q\ x$, but this contradicts A.VI.

THEOREM 2: $x*y\ Q\ y*x$.

Proof: Use Th. 1 and A.IV.

THEOREM 3: *If $x\ Q\ y$ and $u\ Q\ v$, then $x*u\ Q\ y*v$.*

Proof: $x*u\ Q\ u*y$ and $u*y\ Q\ y*v$, by A.IV and hypothesis. Then use A.I.

THEOREM 4: $x*(y*z)\ Q\ (x*y)*z$.

Proof: Using Th. 2, we get: $x*(y*z)\ Q\ (y*z)*x$. Then using Th. 3, Th. 2, A.I and A.III on this, we get: $x*(y*z)\ Q\ z*(y*x)$. Using Th. 2, Th. 3, and A.I, we get theorem.

THEOREM 5: $x\ Q\ y$ or $y\ Q\ x$.

Proof: Assume: not $x\ Q\ y$ and not $y\ Q\ x$. Then $y*z_1\ Q\ x$ and $x*z_2\ Q\ y$, by A.V. From this by Th. 1 and Th. 3, we get: $(y*z_1)*z_2\ Q\ x*z_2$, and then, by A.I, $(y*z_1)*z_2\ Q\ y$. From this, using Th. 4 and A.I, we get: $y*(z_1*z_2)\ Q\ y$, which contradicts A.VI.

THEOREM 6: *If $x*u\ Q\ y*u$, then $x\ Q\ y$.*

Proof: Assume: not $x\ Q\ y$. Then by A.V, there is a z such that $y*z\ Q\ x$. Using Th. 3, hypothesis, Th. 4 and A.I, we get: $y*(z*u)\ Q\ y*u$. From this, by Th. 3, A.III and A.I, we obtain: $(y*u)*z\ Q\ y*u$, which contradicts A.VI.

THEOREM 7: *If* $y*z\ Q\ u$ *and* $x\ Q\ y$, *then* $x*z\ Q\ u$.

Proof: $x*z\ Q\ y*z$, by Th. 1, hypothesis, and Th. 3. Then, $x*z\ Q\ u$, by hypothesis and A.I.

THEOREM 8: *If* $u\ Q\ x*z$ *and* $x\ Q\ y$, *then* $u\ Q\ y*z$.

Proof: Similar to Th. 7.

THEOREM 9: $mx*nx\ Q\ (m+n)x$.

Proof: We use mathematical induction on n. For $n=1$, the proof is immediate. For $n+1$, we begin with: $mx*(n+1)\,x\ Q\ mx*(nx*x)$. Using principally Th. 1, Th. 3, and Th. 6 and the hypothesis that theorem holds for n, we obtain: $mx*(n+1)x\ Q\ (m+n+1)x$.

THEOREM 10: $(m+n)x\ Q\ mx*nx$.

Proof: Similar to Th. 9.

THEOREM 11: $n(mx)\ Q\ (nm)x$.

Proof: We use mathematical induction on n. For $n=1$, the proof is immediate. For $n+1$, we begin with: $(n+1)\,(mx)\ Q\ n(mx)*mx$. Using principally Th. 10, the hypothesis that theorem holds for n, and Th. 9, we get: $(n+1)\,(mx)\ Q\ ((n+1)\,m)\,x$.

THEOREM 12: $(nm)\,x\ Q\ n(mx)$.

Proof: Similar to Th. 11.

THEOREM 13: $n(x*y)\ Q\ nx*ny$.

Proof: Again we use mathematical induction on n. The proof is obvious for $n=1$. For $n+1$, we begin with: $(n+1)(x*y)\ Q\ n(x*y)*(x*y)$, which follows from Th. 10. Using hypothesis that theorem holds for n, Th. 1, Th. 3, A.III, and A.I, we get then: $(n+1)(x*y)\ Q\ nx*(ny*(x*y))$. Starting now from Th. 6, then using Th. 2, Th. 1, Th. 3, A.III, A.I, Th. 1, and Th. 3 again, we get: $nx*(ny*(x*y))\ Q\ nx*(x*(ny*y))$. Combining this with previous result, using A.I and Th. 4, and definition of 'nx', we get the theorem.

THEOREM 14: $nx*ny\ Q\ n(x*y)$.

Proof: Similar to Th. 13.

THEOREM 15: *If* $x\ Q\ y$, *then* $nx\ Q\ ny$.

Proof: We use mathematical induction on n. For $n=1$ the proof is immediate. From hypothesis that theorem holds for n we get immediately: $nx\ Q\ ny$. And, then, by use of Th. 3 and $x\ Q\ y$, we get: $nx*x\ Q\ ny*y$, from which we get the theorem immediately.

THEOREM 16: *If* $nx\ Q\ ny$, *then* $x\ Q\ y$.

Proof: Assume: not $x\ Q\ y$. Then by A.V, there is a z such $y*z\ Q\ x$;

and from this by Th. 15, Th. 14, and A.I, we obtain: $ny*nz\ Q\ nx$. By use of hypothesis and A.I, this yields: $ny*nz\ Q\ ny$, which contradicts A.VI.

THEOREM 17: *If* $m \leqslant n$, *then* $mx\ Q\ nx$.

Proof: If $m=n$, then the theorem immediately by Th. 1. This leaves the case of $m<n$. Thus, $n=m+k$. Now assume: not $mx\ Q\ nx$. Then by Th. 5, $nx\ Q\ mx$, that is, $(m+k)\,x\ Q\ mx$. From this, by Th. 9 and A.I, we get: $mx*kx\ Q\ mx$, which contradicts A.VI.

THEOREM 18: *There is a number n such that* $x\ Q\ ny$.

Proof: By Th. 5, $x\,Q\,y$ or $y\,Q\,x$. Case 1. $x\,Q\,y$. Let $n=1$. Case 2. $y\ Q\ x$. Theorem follows immediately from A.VII.

IV. SYSTEM OF MAGNITUDES

If magnitudes are defined as certain equivalence classes of quantities, a system of extensive magnitudes may be developed, which is useful for proving the formal adequacy of our axioms for extensive quantities. Conceived this way, there would seem to be a proper place for magnitudes as well as quantities, and there need be no interminable debate about the relative merit of each.[6]

The relation defined by the logical product of Q and its converse is obviously reflexive, symmetrical and transitive, that is, it is an equivalence relation, which we may designate by 'C':

$$xCy =_{df} (xQy \text{ and } yQx).$$

Thus, C defines a partition of K, that is, a set of pair-wise disjoint, nonempty subsets of K whose union equals K. We designate the C-equivalence class of which x is a member (that is, the coset x/C) by '$[x]$', and the partition of K by 'K/C'. The relation C has the substitution property relatively to Q and $*$, that is, (i) if $x\,C\,y$ and $y\,Q\,z$, then $x\,Q\,z$, and if $x\,C\,y$ and $z\,Q\,y$, then $z\,Q\,x$, and (ii) if $x\,C\,y$ and $u\,C\,v$, then $x*u\ C\ y*v$. (i) is trivial and (ii) follows immediately from Th. 3 and the definition of 'C'. Thus we may define a relation $\leqslant$ and an operation $+$ in K/C:

(i) $[x] \leqslant [y]$ if and only if $x\ Q\ y$;

(ii) $[x]+[y]$ is the C-equivalence class in K/C which consists of the elements in K standing in relation C to the element $x*y$. Also, '$n[x]$' is defined recursively, just as 'nx' was previously: $1[x]=[x]$ and $n[x]=$

$(n-1)[x]+[x]$. In fine, where $\mathfrak{M}=\langle K, Q, *\rangle$ is a system of extensive quantities, $\mathfrak{M}/C=\langle K/C, \leqslant, +\rangle$ is the equivalence-class (or coset) system of $\mathfrak{M}$ under relation C, and we shall call $\mathfrak{M}/C$ a system of extensive magnitudes.

On the basis of the axioms and theorems already given, it is easy to prove the following theorems for extensive magnitudes, which we shall begin numbering with 21. The theorems are arranged in an order to bring out clearly the algebraic structure of a system of extensive magnitudes. For brevity we write '$[x]<[y]$' for 'not $([y]\leqslant[x])$'.

Th. 21: *If* $[x]$ *and* $[y]$ *are in* K/C, *then* $[x]+[y]$ *is in* K/C.

Th. 22: *If* $[x]$, $[y]$ *and* $[z]$ *are in* K/C, *then* $([x]+[y])+[z]=[x]+([y]+[z])$.

Th. 23: *If* $[x]$ *and* $[y]$ *are in* K/C, *then* $[x]+[y]=[y]+[x]$.

Th. 24: *If* $[x]$, $[y]$ *and* $[z]$ *are in* K/C *and* $[x]+[z]=[y]+[z]$, *then* $[x]=[y]$.

Th. 25: *If* $[x]$ *and* $[y]$ *are in* K/C, $[x]\leqslant[y]$ *and* $[y]\leqslant[x]$, *then* $[x]=[y]$.

Th. 26: *If* $[x]$, $[y]$ *and* $[z]$ *are in* K/C, $[x]\leqslant[y]$ *and* $[y]\leqslant[z]$, *then* $[x]\leqslant[z]$.

Th. 27: *If* $[x]$ *and* $[y]$ *are in* K/C, *then* $[x]\leqslant[y]$ *or* $[y]\leqslant[x]$.

Th. 28: *If* $[x]$ *and* $[y]$ *are in* K/C, *and* $[y]<[x]$, *then there is a* $[z]$ *in* K/C *such that* $[x]=[y]+[z]$.

Th. 29: *If* $[x]$ *and* $[y]$ *are in* K/C, *then* $[x]<[x]+[y]$.

Th. 30: *If* $[x]$ *and* $[y]$ *are in* K/C *and* $[x]\leqslant[y]$, *then there is a number* n *such that* $[y]\leqslant n[x]$.

It is apparent that obvious analogues of all the theorems in Section III may be easily proved. From the theorems stated here, we see that the algebraic structure of a system of extensive magnitudes is that of a simply ordered, 'Archimedean', Abelian semigroup which does not have a zero element and which is closed under subtraction of 'smaller' elements from 'larger' ones.

V. ADEQUACY OF AXIOMS

The formal adequacy of our axioms is proved by making essential use of the theorems on extensive magnitudes. The reason for this is that a system of extensive quantities is in general merely homomorphic to an additive semigroup of positive real numbers, which is to be expected,

since in measurement of objects relative to a certain characteristic the same number is often assigned to distinct objects. In particular, a given number is assigned to a C-equivalence class of objects, which leads to the following metatheorem.

METATHEOREM A: *If* $\mathfrak{M}=\langle K, Q, *\rangle$ *is a system of extensive quantities, then the system of extensive magnitudes* $\mathfrak{M}/C$ *is isomorphic to an additive semigroup of positive real numbers, closed under subtraction of smaller numbers from larger ones.*[7†]

Proof: The proof of this metatheorem follows along standard lines, as given, for instance, in Hoelder (1901) or Birkhoff (1948, p. 226). (Birkhoff's proof for simply-ordered, Archimedean groups need be only slightly modified; Birkhoff also gives detailed references to the literature.) It will therefore suffice briefly to describe the construction of a mapping f with the desired properties. We define the set $S_{[x][e]}$ where $[x]$ and $[e]$ are in K/C, as the set of all rational fractions m/n such that $n[x]\leqslant m[e]$. It is easy to show that $S_{[x][e]}$ has a greatest lower bound, which we define as the number assigned to $[x]$, that is, the mapping f is defined as follows:

$$f_{[e]}([x]) \quad \text{is the greatest lower bound of} \quad S_{[x][e]}.$$

Since it may be shown that $f_{[e]}([e])=1$, the choice of $[e]$ corresponds to the choice of a unit. And, using the theorems of Section IV, it may be shown in a straightforward manner that $f_{[e]}$ has the desired properties: If $[x]\leqslant[y]$, then $f_{[e]}([x])\leqslant f_{[e]}([y])$; $f_{[e]}([x]+[y])=f_{[e]}([x])+f_{[e]}([y])$; and if $[x]\neq[y]$, then $f_{[e]}([x])\neq f_{[e]}([y])$.[8]

The following metatheorem establishes the desired uniqueness property of our axioms. It is equivalent to saying that in the measurement of extensive quantities, only the choice of a unit is arbitrary.

METATHEOREM B: *If* $\mathfrak{M}=<K, Q, *>$ *is a system of extensive quantities, then any two additive semigroups of positive real numbers, which are isomorphic to* $\mathfrak{M}/C$, *are related by a similarity transformation.*

Proof: Consider any additive semigroup of positive real numbers isomorphic to $\mathfrak{M}/C$ under the mapping g. Then it will be sufficient to show that there exists a positive constant c such that for every $[x]$ in K/C, $g([x])=cf_{[e]}([x])$, where $f_{[e]}$ is the mapping defined above. Let $g([e])=c$. Then, assume that there exists an $[x]$ in K/C such that $g([x])<cf_{[e]}([x])$. On this assumption, we may find an m/n such that

$$(1) \qquad g([x])/c < m/n < f_{[e]}([x]).$$

It is clear from definition of $f_{[e]}$ that then $m[e]<n[x]$, and therefore, on hypothesis for g, $mg([e])<ng([x])$, that is, $m/n<g([x])/c$, but this contradicts (1). Similarly, on the assumption that there exists an $[x]$ in K/C such that $cf_{[e]}([x])<g([x])$, we also get a contradiction. Q.E.D.

It may be remarked that a system of extensive magnitudes $\mathfrak{M}/C$ is also isomorphic to (nonadditive) semigroups in the number domain which are not related by a similarity transformation. The last realization of our axioms given in Section VII below is an example of this kind.

VI. INDEPENDENCE OF AXIOMS

The following seven examples establish the mutual independence of our axioms for extensive quantities. The first example provides an interpretation of K, Q and $*$ that is satisfied by all but the first axiom, etc. Since the examples are all of an elementary character, all proofs are omitted.

I. Let K be the set of all positive integers; let $x\,Q\,y$ if and only if $x\leqslant y+1$; and define $x*y$ as $x+y+2$.

II. Let K be simply the set consisting of the number one; let $x\,Q\,y$ if and only if $x\leqslant y$; and define $*$ as ordinary addition.

III. Let K be the set of all positive rational numbers; let $x\,Q\,y$ if and only if $x\leqslant y$; and define $x*y$ as $\mathrm{Max}\,(x, y)+\frac{1}{2}\,\mathrm{Min}\,(x, y)$.

IV. Let K be the set of all positive rational numbers; let $x\,Q\,y$ if and only if $x\leqslant y$; and define $x*y$ as $x+10y$.

V. Let K be the set of all positive integers with the exception of the number one; let $x\,Q\,y$ if and only if $x\leqslant y$; and define $*$ as ordinary addition.

VI. Let K be the set consisting of the number one; let $x\,Q\,y$ if and only if $x\leqslant y$; and define $*$ as ordinary multiplication.

VII. Let K be the set consisting of (i) all ordered pairs whose first members are positive integers and whose second members are integers, together with (ii) all ordered pairs whose first members are zero and whose second members are positive integers; where $x=\langle a, b\rangle$ and $y=\langle c, d\rangle$, let $x\,Q\,y$ if and only if $a<c$, or $a=c$ and $b\leqslant d$; define $\langle a, b\rangle * \langle c, d\rangle$ as $\langle a+c, b+d\rangle$.

VII. INDEPENDENCE OF PRIMITIVES

Using Padoa's principle[9], we may establish the mutual independence of the three primitives K, Q and $*$ of our axioms for extensive quantities. The application of Padoa's principle requires that we find for each primitive two different realizations of our axioms such that the other two primitives are given the same interpretation for both realizations.

I. Independence of K. For the first realization, let K' be the set of positive integers; let $x\,Q'\,y$ if and only if $x \leqslant y$; and define $*'$ as ordinary addition. And, for the second realization, let K'' be the set of even positive integers; $Q''=Q'$; $*''=*'$.

II. Independence of Q. For the first realization, let K' be the set of all ordered pairs of positive integers; where $x=\langle a, b\rangle$ and $y=\langle c, d\rangle$, let $x\,Q'\,y$ if and only if $a \leqslant c$; and $\langle a, b\rangle *' \langle c, d\rangle = \langle a+c, b+d\rangle$. For the second realization, $K''=K'$; $*''=*'$; where $x=\langle a, b\rangle$ and $y=\langle c, d\rangle$, let $x\,Q''\,y$ if and only if $b \leqslant d$. Thus, we have, for instance, $\langle 1, 2\rangle Q' \langle 2, 1\rangle$, and not $\langle 1, 2\rangle Q'' \langle 2, 1\rangle$.

III. Independence of $*$. For the first realization, let K' be the set of positive real numbers; let $x\,Q'\,y$ if and only if $x \leqslant y$; and define $*'$ as ordinary addition. For the second realization, $K''=K'$; $Q''=Q'$; $x *'' y = \sqrt{x^2+y^2}$. Thus, we have, for instance, $1 *' 2=3$, and $1 *'' 2 \neq 3$.

VIII. EMPIRICAL REALIZATIONS

Our system of axioms for extensive quantities was designed to eliminate the two defects of Hoelder's system, which were mentioned in Section I. In this concluding section, I would like to point out, from the standpoint of the methodological analysis of measurement, two more fundamental defects common to both systems.

Given any realization of our axioms, it is apparent, in the first place, that the set K must contain an infinite number of elements. This flagrantly violates obvious finitistic requirements of empirical measurement. And it is apparent, in the second place, that the realization of Q must be a perfectly transitive relation, which entails that the measuring instrument used to determine whether or not two objects stand in the relation Q must possess perfect sensitivity. However, a lack of such perfect sensitivity seems characteristic of nearly all measuring instruments. An equal-arm

balance, for instance, can only differentiate between objects having a mass-difference greater than some finite amount.

The standard axiomatic theory of quantities must be altered rather profoundly in order to take account of these two problems. At least from a methodological standpoint, such an altered formal system, mirroring more accurately the facts of actual, imperfect measurement, would be of interest.

NOTES

[1] I am grateful to J. C. C. McKinsey for a number of helpful suggestions in connection with the present paper.

[2] Newton (1769, p. 2).

[3] The work of Norman R. Campbell (1920) and (1928) has been outstanding in this direction.

[4] It may be remarked that this traditional classification is not very satisfactory, since there are also quantities which are assigned numbers uniquely up to a variety of other groups of transformations. However, this issue is irrelevant here, since we are solely concerned with extensive quantities in the sense just defined, and the problem of precisely how many formally different kinds of quantities it is useful to distinguish need not concern us.

[5] This criticism would also seem to apply to the axioms for the measurement of utility given by J. von Neumann and O. Morgenstern (1947): '=' should designate the relation of indifference rather than that of identity.

[6] For some aspects of this debate, see Russell (1903, Chaps. 19, 20) and Nagel (1931).

[7†] I would now call Metatheorem A the 'Representation theorem' for extensive quantities, and Metatheorem B the 'Uniqueness theorem'.

[8] Another method of proof of this metatheorem is to show that $\mathfrak{M}/C$ can be uniquely embedded in an Archimedean, simply ordered group. And it is well known (see Birkhoff, 1948) that any such group is isomorphic to a subgroup of the additive group of all real numbers.

[9] Padoa (1901); a clear statement of this principle is also to be found in McKinsey (1935).

4. FOUNDATIONAL ASPECTS OF THEORIES OF MEASUREMENT*[1]

I. DEFINITION OF MEASUREMENT

It is a scientific platitude that there can be neither precise control nor prediction of phenomena without measurement. Disciplines as diverse as cosmology and social psychology provide evidence that it is nearly useless to have an exactly formulated quantitative theory, if empirically feasible methods of measurement cannot be developed for a substantial portion of the quantitative concepts of the theory. Given a physical concept like that of mass or a psychological concept like that of habit strength, the point of a theory of measurement is to lay bare the structure of a collection of empirical relations which may be used to measure the characteristic of empirical phenomena corresponding to the concept. Why a collection of relations? From an abstract standpoint, a set of empirical data consists of a collection of relations between specified objects. For example, data on the relative weights of a set of physical objects are easily represented by an ordering relation on the set; additional data, and a fortiori an additional relation, are needed to yield a satisfactory quantitative measurement of the masses of the objects.

The major source of difficulty in providing an adequate theory of measurement is to construct relations which have an exact and reasonable numerical interpretation, and, yet also, have a technically practical empirical interpretation. The classical analyses of the measurement of mass, for instance, have the embarrassing consequence that the basic set of objects measured must be infinite. Here the relations postulated have acceptable numerical interpretations, but are utterly unsuitable empirically. Conversely, as we shall see in the last section of this paper, the structure of relations which have a sound empirical meaning often cannot be succinctly characterized so as to guarantee a desired numerical interpretation.

* Reprinted from *The Journal of Symbolic Logic* **23** (1958), 113–128. Written jointly with Dana Scott.

Nevertheless, this major source of difficulty will not here be carefully scrutinized in a variety of empirical contexts. The main point of the present paper is to show how foundational analyses of measurement may be grounded in the general theory of models, and to indicate the kind of problems relevant to measurement which may then be stated (and perhaps answered) in a precise manner.

Before turning to problems connected with construction of theories of measurement, we want to give a precise set-theoretical meaning to the notions involved. To begin with, we treat sets of empirical data as being (finitary) relational systems, that is to say, finite sequences of the form $\mathfrak{A}=\langle A, R_1, \ldots, R_n\rangle$, where A is a nonempty set of elements called the *domain* of the relational system $\mathfrak{A}$, and $R_1, \ldots, R_n$ are finitary relations on A. The relational system $\mathfrak{A}$ is called *finite* if the set A is finite; otherwise, *infinite*. It should be obvious from this definition that we are mainly considering *qualitative* empirical data. Intuitively we may think of each particular relation R_i (an m_i-ary relation, say) as representing a complete set of 'yes' or 'no' answers to a question asked of every m_i-termed sequence of objects in A. The point of this paper is not to consider that aspect of measurement connected with the actual collection of data, but rather the analysis of relational systems and their numerical interpretations.

If $s=\langle m_1, \ldots, m_n\rangle$ is an n-termed sequence of positive integers, then a relational system $\mathfrak{A}=\langle A, R_1, \ldots, R_n\rangle$ is of *type* s if for each $i=1, \ldots, n$ the relation R_i is an m_i-ary relation. Two relational systems are *similar* if there is a sequence s of positive integers such that they are both of type s. Notice that the type of a relational system is uniquely determined only if all the relations are nonempty; the avoiding of this ambiguity is not worthwhile. Suppose that two relational systems $\mathfrak{A}=\langle A, R_1, \ldots, R_n\rangle$ and $\mathfrak{B}=\langle B, S_1, \ldots, S_n\rangle$ are of type $s=\langle m_1, \ldots, m_n\rangle$. Then $\mathfrak{B}$ is a *homomorphic image of* $\mathfrak{A}$ if there is a function f from A onto B such that, for each $i=1, \ldots, n$ and for each sequence $\langle a_1, \ldots, a_{m_i}\rangle$ of elements of A, $R_i(a_1, \ldots, a_{m_i})$ *if and only if* $S_i(f(a_1), \ldots, f(a_{m_i}))$. If the function f is one-one, then $\mathfrak{B}$ is an *isomorphic image of* $\mathfrak{A}$, or simply $\mathfrak{A}$ and $\mathfrak{B}$ are *isomorphic*. $\mathfrak{A}$ is a *subsystem* of $\mathfrak{B}$ if $A \subseteq B$ and, for each $i=1, \ldots, n$, the relation R_i is the restriction of the relation S_i to A. $\mathfrak{A}$ is *imbeddable* in $\mathfrak{B}$ if some subsystem of $\mathfrak{B}$ is a homomorphic image of $\mathfrak{A}$.[2] A *numerical relational system* is simply a relational system whose domain of elements is the set Re of all

real numbers. A *numerical assignment* for a relational system $\mathfrak{A}$ with respect to a numerical relational system $\mathfrak{N}$ is a function which imbeds $\mathfrak{A}$ in $\mathfrak{N}$. A numerical assignment is not required to be one-one.

Within the framework of the preceding formal definitions, it is now possible to give an exact characterization of a theory of measurement. First of all, the general outlines of a theory are determined by fixing a finite sequence s of positive integers and only considering relational systems of type s. Next, a numerical relational system $\mathfrak{N}$ of type s is selected which corresponds to the intended numerical interpretation of the theory, and only relational systems imbeddable in $\mathfrak{N}$ are permitted. Moreover, the theory need not concern all relational systems of type s imbeddable in $\mathfrak{N}$, but only a distinguished subclass. Since it is reasonable that no special set of objects be preferred, we require that the distinguished subclass be closed under isomorphism. We thus arrive at the following characterization of theories of measurement as definite entities: a theory of measurement *is* a class $\boldsymbol{K}$ of relational systems closed under isomorphism for which there exists a finite sequence s of positive integers and a numerical relational system $\mathfrak{N}$ of type s such that all relational systems in $\boldsymbol{K}$ are of type s and imbeddable in $\mathfrak{N}$.[3]

Some readers may object that the definition of theories of measurement should be linguistic rather than set-theoretical in character, since a theory is ordinarily thought of as a linguistic entity. To be sure, many theories of measurement have a natural formalization in first-order predicate logic with identity. Notice, however, that first-order axioms by themselves are not adequate, for if they admit one infinite relational system as a model then they have models of every infinite cardinality, and it is difficult to see how any natural connection can be established between numerical models and models of arbitrary cardinality. Even neglecting this criticism, first-order axioms are not adequate to express properties involving arbitrary natural numbers, for example, that a relational system is finite or that as an ordering it has Archimedean properties. Any linguistic definition of theories which will permit expression of these more general properties would require extensive machinery and be immediately involved in some of the deepest problems of modern metamathematics. On the other hand, we do not wish to give the impression that we reject any linguistic questions. In fact, we use our set-theoretical definition as a point of departure for asking just such questions.

On the basis of the definition of theories of measurement adopted, two questions naturally arise, to each of which we devote a section. In the first place, is a given class of relational systems a theory of measurement? And in the second place, given a theory of measurement, in what sense can it be axiomatized?

II. EXISTENCE OF MEASUREMENT

A simple counterexample shows that not every class of relational systems of a given type closed under isomorphism is a theory of measurement. Let $\boldsymbol{O}$ be the class of all relational systems of type $\langle 2 \rangle$ that are simple orderings. Let $\langle A, R \rangle$ be a system in $\boldsymbol{O}$ where R well-orders A and A has a power not equal to or less than that of the continuum. Such a relational system can be proved to exist even without the help of the axiom of choice, but of course with aid of this axiom, the existence is obvious. By way of contradiction, suppose that $\boldsymbol{O}$ is a theory of measurement relative to a numerical relational system $\langle \mathrm{Re}, S \rangle$. From the definition, it follows that $\langle A, R \rangle$ is imbeddable in $\langle \mathrm{Re}, S \rangle$ and that there is a numerical assignment f mapping A onto a subset of Re such that *xRy if and only if* $f(x)\ S\ f(y)$ for all elements $x, y \in A$. Let a, b be elements of A such that $f(a)=f(b)$. From the hypothesis that R is a simple ordering, we can assume without loss of generality that aRb. Hence, we have $f(a)\ S\ f(b)$, and then $f(b)\ S\ f(a)$, and finally, bRa. R is antisymmetric, and so $a=b$. This argument shows that the function f is one-one. Hence A has the same power as a subset of Re, which is impossible. This proof shows that every theory of measurement included in the class $\boldsymbol{O}$ contains only relational systems of power at most that of the continuum. It is an unsolved problem of set-theory closely connected with the continuum hypothesis whether the class $\boldsymbol{O}$ restricted to systems of power at most that of the continuum is actually a theory of measurement.[4] At least it can be very easily shown that $\boldsymbol{O}$ so restricted is not a theory of measurement relative to the system $\langle \mathrm{Re}, \leqq \rangle$, where the relation $\leqq$ is the usual ordering of the real numbers.[5] Indeed, the exact condition that a relational system in $\boldsymbol{O}$ must satisfy to be imbeddable in $\langle \mathrm{Re}, \leqq \rangle$ is not really elementary, and the proof of the necessity involves the axiom of choice.[6]

Let $\boldsymbol{O}'$ be $\boldsymbol{O}$ restricted to countable relational systems.[7] It was proved by Cantor that $\boldsymbol{O}'$ is a theory of measurement relative to $\langle \mathrm{Re}, \leqq \rangle$, to

formulate somewhat irreverently his classical result in the terminology of this paper. This restriction to countable relational systems is always sufficient, for it can be shown that the class of *all* countable relational systems of a given type is a theory of measurement; however, the numerical relational system required is so bizarre as to be of no practical value.

A primary aim of measurement is to provide a means of convenient computation. Practical control or prediction of empirical phenomena requires that unified, widely applicable methods of analyzing the important relationships between the phenomena be developed. Imbedding the discovered relations in various numerical relational systems is the most important such unifying method that has yet been found. But among the morass of all possible numerical relational systems only a very few are of any computational value, indeed only those definable in terms of the ordinary arithmetical notions. From an empirical standpoint, most sets of qualitative data can find numerical interpretation by relations defined in terms of addition and ordering alone. By way of example, we may cite the measurement of masses, distances, sensation intensities, and subjective probabilities. Frequently the consideration of weighted averages requires also the use of the multiplication of numbers. However, in the examples given in this paper, we shall restrict ourselves to the notions of addition and ordering.

No natural scientific situation would seem strictly to require the consideration of sets of infinite data. This state of affairs suggests that theories of measurement containing only finite relational systems would suffice for empirical purposes. The problem is delicate, however, for the measurement of a meteorological quantity such as temperature by an automatic recording device is usually treated as continuous, both in its own scale and in time. Yet the important problem of measurement does not really lie in the correct use of such recording devices, but rather in their initial calibration, a process proceeding from a finite number of qualitative decisions. Because of the awkwardness of the uniform application of finite relational systems, we shall not generally make this restriction.

Further remarks about establishing the existence of measurement are best motivated by reference to a concrete example. In a recent paper (1956), Luce has introduced a generalization of simple orderings which he calls *semiorders*. A *semiorder* is a relational system $\langle A, P \rangle$ of type $\langle 2 \rangle$

which satisfies the following axioms for all $x, y, z, w \in A$:

S1. *Not xPx.*

S2. *If xPy and zPw, then either xPw or zPy.*

S3. *If xPy and zPx, then either wPy or zPw.*[8]

Such relations are most likely to occur in situations where objects are to be arranged in order, and where it is difficult to say exactly when two objects are indifferent. For example, to say that xPy might be interpreted as meaning that the pitch of the sound x is *definitely higher* than the pitch of y, or that the hue of color x is *definitely brighter* than the hue of color y, or that the weight of the object x is *noticeably greater* than that of y, etc. *Indifference* between two objects x and y (in symbols: xIy) is defined as not xPy, and not yPx. The point of Luce's axioms is that the relation I of indifference is not always transitive, a fact easily appreciated for each of the intuitive interpretations given above.

In his paper, Luce gives a certain numerical interpretation for certain kinds of semiorders, but he does not show that any particular class of semiorders is a theory of measurement in the sense used here, because his interpretations are not relative to a fixed numerical relation. However, in the finite case, the situation becomes relatively simple. Let $\gg$ be that relation between real numbers defined by the condition: $x \gg y$ *if and only if* $x > y+1$. Clearly, if x and y are real numbers such that $x \gg y$, then it is fair to say that x is *definitely greater than y*, or better, x is *noticeably greater than y*. It is in fact a simple exercise to prove that the relational system $\langle \mathrm{Re}, \gg \rangle$ is a semiorder. Further, we shall give the proof of the following result:

The class of finite semiorders is a theory of measurement relative to the numerical relational system $\langle \mathrm{Re}, \gg \rangle$.

Before presenting the proof of the above, it would be well to outline a general method in proofs of the existence of measurement which we shall call the *method of cosets*. Let $\mathfrak{A} = \langle A, R_1, \ldots, R_n \rangle$ be a relational system of type $\langle m_1, \ldots, m_n \rangle$. A uniquely determined equivalence relation E is introduced into $\mathfrak{A}$ by the condition: xEy *if and only if for each* $i = 1, \ldots, n$ *and each pair* $\langle z_1, \ldots, z_{m_i} \rangle, \langle w_1, \ldots, w_{m_i} \rangle$ *of* m_i*-termed sequences of elements of A, if* $z_j \neq w_j$ *implies* $\{z_j, w_j\} = \{x, y\}$ for $j = 1, \ldots, m_i$, *then* $R_i(z_1, \ldots, z_{m_i})$ *if and only if* $R_i(w_1, \ldots, w_{m_i})$.

Even though the above definition is complicated to state in general, the meaning of the relation xEy is simple: elements x and y stand in the

relation E just when they are perfect substitutes for each other with respect to all the relations R_i.[9]

The notion of a weak ordering can serve as an example. Let $\mathfrak{A}=\langle A, R\rangle$ where the binary relation R is connected and transitive. Then xEy is equivalent to the condition: *For all* $z\in A$, xRz *if and only if* yRz, *and* zRx *if and only if* zRy. However, this simplifies finally to: xRy *and* yRx.

Returning now to the general case, define, for each $x\in A$, $[x]$ to be the class of all y such that xEy. $[x]$ is called the *coset* of x. Let A^* be the class of all $[x]$ for $x\in A$. Directly from the definition of E we can deduce that it is permissible to define m_i-ary relations R_i^* over A^* such that, for all $x_1, \ldots, x_{m_i}\in A$, $R_i^*([x_1], \ldots, [x_{m_i}])$ *if and only if* $R_i(x_1, \ldots, x_{m_i})$. The relational system $\mathfrak{A}^*=\langle A_1^*, R_1^*, \ldots, R_n^*\rangle$ is called the *reduction of* $\mathfrak{A}$ *by cosets.*

It is at once obvious that $\mathfrak{A}^*$ is a homomorphic image of $\mathfrak{A}$ and that $\mathfrak{A}^{**}$ is isomorphic with $\mathfrak{A}^*$. What is not quite obvious is the following: *If* $\mathfrak{B}$ *is a homomorphic image of* $\mathfrak{A}$, *then* $\mathfrak{A}^*$ *is a homomorphic image of* $\mathfrak{B}$.

By way of proof, let f be a homomorphism of $\mathfrak{A}$ onto $\mathfrak{B}$. We wish to show that if $f(x)=f(y)$, then $[x]=[y]$. Instead of the general case, assume for simplicity that $\mathfrak{A}$ and $\mathfrak{B}$ are of type $\langle 2\rangle$ and $\mathfrak{A}=\langle A, R\rangle$, $\mathfrak{B}=\langle B, S\rangle$. We must show that if $f(x)=f(y)$, then xEy, or in other words, for all $z\in A$, xRz if and only if yRz, and zRx if and only if zRy. Assume xRz. It follows that $f(x)\ S\ f(z)$, and hence $f(y)\ S\ f(z)$, which implies that yRz. The argument is clearly symmetric. We have therefore shown that there is a function g from B onto A^* such that $g(f(x))=[x]$ for $x\in A$. It is trivial to verify that g is a homomorphism of $\mathfrak{B}$ onto $\mathfrak{A}^*$.

Notice the following relation between the concepts of homomorphic image and subsystem: if $\mathfrak{B}$ is a homomorphic image of $\mathfrak{A}$, then $\mathfrak{B}$ is isomorphic to a subsystem of $\mathfrak{A}$. For let f be a homomorphism of $\mathfrak{A}$ onto $\mathfrak{B}$. Let g be any function from B into A such that $f(g(y))=y$ for all $y\in B$. The restriction of $\mathfrak{A}$ to the range of g yields the subsystem of $\mathfrak{A}$ isomorphic with $\mathfrak{B}$.

Using the above remarks, we can establish at once the equivalence: $\mathfrak{A}$ *is imbeddable in* $\mathfrak{B}$ *if and only if* $\mathfrak{A}^*$ *is imbeddable in* $\mathfrak{B}$.

Further, it follows that any function imbedding $\mathfrak{A}^*$ in $\mathfrak{B}$ is always an isomorphism of $\mathfrak{A}^*$ onto a subsystem of $\mathfrak{B}$, and of all homomorphic images of $\mathfrak{A}$ this property is characteristic of $\mathfrak{A}^*$.

Let $\boldsymbol{K}$ now be any class of relational systems closed under isomorphism. Let $\boldsymbol{K}^*$ be the class of all systems isomorphic to some $\mathfrak{A}^*$ for $\mathfrak{A} \in \boldsymbol{K}$. In effect we have shown above:

(i) $\boldsymbol{K}$ is a theory of measurement relative to a numerical relational system $\mathfrak{N}$ if and only if $\boldsymbol{K}^*$ is also.

(ii) If $\boldsymbol{K}$ in addition is closed under the formation of subsystems, then $\boldsymbol{K}^*$ is the class of all systems in $\boldsymbol{K}$ possessing only one-one numerical assignments.

To use our example again, if $\boldsymbol{K}$ is the class of weak orders, then $\boldsymbol{K}^*$ is the class of simple orders. Notice that the proof in the first paragraph of this section is a special case of (ii).

It should be remarked that for a relational system $\mathfrak{A}$, $\mathfrak{A}$ and $\mathfrak{A}^*$ always satisfy exactly the same formulas of first-order logic not involving the notion of identity. Hence, if $\boldsymbol{K}$ is the class of all relational systems satisfying first-order axioms without identity, then $\boldsymbol{K}^*$ is the class of all systems satisfying the axioms for $\boldsymbol{K}$ and in addition satisfying the axiom:

(*) If xEy, then $x=y$.

The application of this remark to weak orderings and simple orderings is left to the reader.

Consider again the case of semiorders. Let $\boldsymbol{S}$ be the class of all finite semiorders. For any $\langle A, P\rangle \in \boldsymbol{S}$, consider the relation I of indifference defined above. In terms of I one can establish a simplified characterization of E: *xEy if and only if for all $z \in A$, xIz if and only if yIz.*

Introduce (*) as a new axiom S4. The class of all $\mathfrak{A} \in \boldsymbol{S}$ satisfying S4 is just the class $\boldsymbol{S}^*$. Notice that unlike the pleasant situation with weak orderings and simple orderings, the class $\boldsymbol{S}^*$ is not closed under the formation of subsystems even though $\boldsymbol{S}$ is.

For any semiorder $\langle A, P\rangle$ introduce a further relation R as follows: *xRy if and only if for all z, if zPx then zPy, and if yPz then xPz.*

We leave to the reader the elementary verification of the fact that R is a weak ordering of A, and that xEy if and only if xRy and yRx. Thus, if $\langle A, P\rangle \in \boldsymbol{S}^*$, then R is a simple ordering of A. The connection between P and R is clearer if one notices that xPy implies xRy, and that, if xRx_1, x_1Py_1, and y_1Ry, then xPy.

Now let $\mathfrak{A}=\langle A, P\rangle$ be a fixed member of $\boldsymbol{S}^*$. We wish to show that $\mathfrak{A}$ has an assignment in $\langle \mathrm{Re}, \gg \rangle$. Under the relation R, A is simply ordered. Let $A=\{x_0, \ldots, x_n\}$ where x_iRx_{i-1} and $x_i \neq x_{i-1}$. Define by a course of

values recursion a sequence $a_0, \ldots, a_n$ of rational numbers determined uniquely by the following two conditions:

(1) If $x_i I x_0$, then $a_i = i/(i+1)$

(2) If $x_i I x_j$ and $x_i P x_{j-1}$ where $j>0$, then $a_i = i/(i+1)a_j + 1/(i+1)a_{j-1} + 1$.

Notice that in (2) the hypothesis implies that $j \leqq i$, while in the case $j=i$ the formula for a_i simplifies to $a_i = a_{i-1} + i + 1$. Notice further that every element x_i comes either under (1) or (2); for letting x_j be the first element such that $x_j I x_i$, there are two cases: $j=0, j>0$. Clearly we always have $a_i \geqq 0$.

We show first that $a_i > a_{i-1}$ by induction on i. For case (1), this is obvious. Passing to (2), assume that $x_i I x_j$ and $x_i P x_{j-1}$. If $x_{i-1} I x_0$, then $a_{i-1} < 1$ while $a_i > 1$. Hence we can assume not $x_{i-1} I x_0$, or in other words $x_{i-1} P x_0$. Let x_k be the first element such that $x_{i-1} I x_k$ and $x_{i-1} P x_{k-1}$. By definition $a_{i-1} = (i-1)/i\, a_k + 1/i\, a_{k-1} + 1$. If $j=i$, there is no problem. Assume then that $j<i$. Now $x_{i-1} R x_j$, $x_i R x_{i-1}$, and $x_j I x_i$, hence $x_j I x_{i-1}$, and so by our choice of k we have $k \leqq j$. By the induction hypothesis on i, it follows that $a_j > a_{j-1}$ and $a_k > a_{k-1}$. If $k=j$, the required inequality is obvious. If $k \leqq j-1$, then $a_i > a_{j-1} + 1$. Similarly $a_{i-1} < a_k + 1$, but again, by the induction hypothesis, $a_k \leqq a_{j-1}$, and hence $a_i > a_{i-1}$.

The next step is to prove that, if $x_i P x_k$, then $a_i > a_k + 1$. Let x_j be the first element such that $x_i I x_j$ and $x_i P x_{j-1}$. We have $j-1 \geqq k$, and, in view of the preceding argument, $a_{j-1} \geqq a_k$. But $a_{j-1} + 1 < a_i$, whence $a_i > a_k + 1$.

Conversely we must show that, if $a_i > a_k + 1$, then $x_i P x_k$. The hypothesis of course implies $i>k$. Assume by way of contradiction that not $x_i P x_k$. It follows that $x_i I x_k$. Let x_j be the first element such that $x_i I x_j$; then $k \geqq j$ and $a_k \geqq a_j$. If $j=0$, then $x_i I x_0$ and $x_k I x_0$, because $x_i R x_k$. But then $0 \leqq a_i < 1$ and $0 \leqq a_k < 1$, which contradicts the inequality $a_i > a_k + 1$. We can conclude that $j>0$. Now $a_i < a_j + 1$, but $a_k \geqq a_j$, and thus $a_i < a_k + 1$, which again is a contradiction. All cases have been covered, and the argument is complete.

Finally, define a function f on A such that $f(x_i) = a_i$. We have actually shown that f imbeds $\mathfrak{A}$ in $\langle \mathrm{Re}, \gg \rangle$. Thus it has been proved that $\boldsymbol{S}^*$ is a theory of measurement relative to $\langle \mathrm{Re}, \gg \rangle$, and, by the general remarks on the method of cosets, we conclude that $\boldsymbol{S}$ is also a theory of measurement relative to $\langle \mathrm{Re}, \gg \rangle$.

Notice that the above proof would also work in the infinite case as long as the ordering R is a well-ordering of type ω.

Let us now summarize the steps in establishing the existence of measurement using as examples simple orderings and semiorders. First, after one is given a class, $\boldsymbol{K}$ say, of relational systems, the numerical relational system should be decided upon. The numerical relational system should be suggested naturally by the structure of the systems in $\boldsymbol{K}$, and as was remarked, it is most practical to consider numerical systems where all the relations can be simply defined in terms of addition and ordering of real numbers. Second, if the proof that $\boldsymbol{K}$ is a theory of measurement is not at once obvious, the cardinality of systems in $\boldsymbol{K}$ should be taken into consideration. The restriction to countable systems would always seem empirically justified, and adequate results are possible with a restriction to finite systems. Third, the proof of the existence of measurement can often be simplified by the reduction of each relational system in $\boldsymbol{K}$ by the method of cosets. Then, instead of trying to find numerical assignments for each member of $\boldsymbol{K}$, one concentrates only on the reduced systems. This plan was helpful in the case of semiorders. Instead of cosets, it is sometimes feasible to consider imbedding by subsystems. That is to say, one considers some convenient subclass $\boldsymbol{K}' \subseteq \boldsymbol{K}$ such that every element of $\boldsymbol{K}$ is a subsystem of some system in $\boldsymbol{K}'$. If $\boldsymbol{K}'$ is a theory of measurement, then so is $\boldsymbol{K}$. In the case of semiorders we could have used either plan: cosets or subsystems.

After the existence of measurement has been established, there is one question which is often of interest: For a given relational system, what is the class of all its numerical assignments? We present an example.

Consider relational systems $\mathfrak{A} = \langle A, D \rangle$ of type $\langle 4 \rangle$. For such systems we introduce the following definitions: *xRy if and only if $xyDyy$. xyM^1zw if and only if $xyDzw$, $zwDxy$, yRz and zRy. $xyM^{n+1}zw$ if and only if there exist $u, v \in A$ such that xyM^nuv and uvM^1zw.*

Let $\boldsymbol{H}$ be the class of all such relational systems which satisfy the following axioms for every $x, y, z, u, v, w \in A$:

*A*1. *If $xyDzw$ and $zwDuv$, then $xyDuv$.*

*A*2. *$xyDzw$ or $zwDxy$.*

*A*3. *If $xyDzw$, then $xzDyw$.*

*A*4. *If $xyDzw$, then $wzDyx$.*

*A*5. *If xRy and $yzDuv$, then $xzDuv$.*

A6. There is a $z \in A$ such that $xzDzy$ and $zyDxy$.

A7. If not $xyDzw$ and not xRy, then there is a $u \in A$ such that $zwDxu$, not xRu, and not uRy.

A8. If $xyDzw$ and not xRy, then there are $u, v \in A$ and an n such that $zuM^n vw$ and $zuDxy$.

These axioms imply that for a system $\mathfrak{A}$ in $\boldsymbol{H}$, the relation R is a weak ordering of A, and the intuitive interpretation of $xyDzw$ in case yRx and wRz is that the interval between x and y is not greater than the interval between z and w. Making heavy use of the last three existence axioms, it can be shown that $\boldsymbol{H}$ is a theory of measurement relative to the numerical relational system $\langle \mathrm{Re}, \Delta \rangle$ where Δ is the quaternary relation defined by the condition *$xy\Delta zw$ if and only if* $x-y \leqq z-w$ for all $x, y, z, w \in \mathrm{Re}$. It must be stressed that the Archimedean property of the ordering embodied in A8 cannot be formulated in first-order logic, because it implies that all systems in $\boldsymbol{H}^*$ have cardinality not more than the power of the continuum. In addition, it can be shown that, if $\mathfrak{A}$ is in $\boldsymbol{H}$, and f and g are two numerical assignments of $\mathfrak{A}$ relative to $\langle \mathrm{Re}, \Delta \rangle$, then f and g are related by a positive linear transformation;[10†] that is, there exist $\alpha, \beta \in \mathrm{Re}$ with $\alpha > 0$ such that, for all $x \in \mathrm{Re}$, $f(x) = \alpha g(x) + \beta$. This gives in a certain sense the answer to the question above: if we know one numerical assignment for $\mathfrak{A}$, we know them all. Except for very special systems in $\boldsymbol{H}$, nothing more specific can really be expected.

Notice that all relational systems in $\boldsymbol{H}$ are necessarily infinite. In the next section we shall consider in detail the theory of measurement $\boldsymbol{F}$ consisting of all finite relational systems imbeddable in $\langle \mathrm{Re}, \Delta \rangle$. Here the situation is quite hopeless. There simply is no apparent general statement that can be made about the relation between assignments. Inasmuch as any function φ which imbeds $\langle \mathrm{Re}, \Delta \rangle$ in itself is necessarily a linear transformation and conversely, it follows that, if $\mathfrak{A}$ is a system in $\boldsymbol{F}$ and f is an assignment for $\mathfrak{A}$, then f composed with a linear transformation is also an assignment. The main difficulty with $\boldsymbol{F}$ is that two assignments for the same system in $\boldsymbol{F}$ need not be related by a linear transformation.

III. AXIOMATIZABILITY

Given a theory of measurement, it is natural to ask various questions about its axiomatizability, for the axiomatic analysis of any mathematical

theory usually throws considerable light on the structure of the theory. In particular, given an extrinsic characterization of a theory of measurement via a particular numerical relational system, it is quite desirable to have an intrinsic axiomatic characterization of the theory to be able better to recognize when a relational system actually belongs to the theory. In view of the paucity of metamathematical results concerning the axiomatics of higher-order theories, we shall restrict ourselves to the problem of axiomatizing theories of measurement in first-order logic.

It is a well-known result that if a set of first-order axioms has one infinite model, then it has models of unbounded cardinalities. Since for the most part we are interested in one-one assignments with values in the set of real numbers, unbounded cardinalities are hardly an asset. That is to say, the class of all relational systems that are models of a given set of first-order axioms is usually not a theory of measurement. To remove such difficulties without having to understand them, we simply restrict the cardinalities under consideration. Even a restriction to finite cardinalities is not too strong and leads to some rather difficult questions. Thus for the remainder of this section we shall consider only *finitary theories of measurement*, i.e., theories containing only finite relational systems. Such a theory is called *axiomatizable*, if there exists a set of sentences of first-order logic (the axioms of the theory) such that a finite relational system is in the theory if and only if the system satisfies all the sentences in the set. A theory is *finitely axiomatizable* if it has a finite set of axioms. A theory is *universally axiomatizable* if it has a set of axioms each of which is a universal sentence (i.e., a sentence in prenex normal form with only universal quantifiers).

It should be observed, first, that *any* finitary theory of measurement is axiomatizable. This is no deeper than saying that in first-order logic we can write down a sentence completely describing the isomorphism type of each finite relational system not in the given theory, and clearly the negations of these sentences can serve as the required set of axioms. It is of course quite obvious that we cannot in each instance give an effective method for writing down the axioms, since there are clearly a continuum number of distinct finitary theories of measurement. Notice also that if the theory is closed under subsystems then the axioms may be taken as universal sentences, and conversely. In case one considers theories consisting of all finite relational systems imbeddable in a given numerical relational

system, then the problem of a recursive or effective axiomatization is simply the problem of whether the class of universal sentences true in the given numerical relational system is recursively enumerable or not. It is not difficult to establish that this last problem is equivalent to the problem of giving a recursive enumeration of all the relation types of finite relational systems *not* imbeddable in the given numerical relational system. For numerical relational systems whose relations are definable in first-order logic in terms of $+$ and $\leqq$, these problems do not arise since the first-order theory of $+$ and $\leqq$ is decidable, and it is to these relational systems that we shall primarily restrict our further attention.

In the second place, in all domains of mathematics a finite axiomatization of a theory is usually felt to be the most satisfactory result. No doubt the psychological basis for such a feeling rests on the fact that only a finite characterization can in one step explicitly lay bare the full structure of a theory. Of course an extremely complicated axiomatization may be of little practical value, and as regards theories of measurement, there is a further complication. Namely, if an axiomatization in first-order logic, no matter how elegant it may be, involves a combination of several universal and existential quantifiers, then the confirmation of this axiom may be highly contingent on the relatively arbitrary selection of the particular domain of objects. From the empirical standpoint, aside from the possible requirement of a fixed minimal number of objects, results ought to be independent of an exact specification of the extent of the domain.

We are thus brought to our third observation: A finite universal axiomatization of a theory of measurement always yields a characterization independent of accidental object selection. To be precise, consider a fixed universal sentence. This formula will obviously contain just a finite number of variables. Hence, to verify the truth of the sentence in a particular relational system, we need consider only subsets of the domain of a uniformly bounded cardinality. Furthermore, verification for each subset is completely independent of any relationships with the complementary set.

Simple orderings and semiorders are examples of this last point. To determine whether a finite relational system of type $\langle 2 \rangle$ is a simple ordering, one has only to consider triples of objects; for semiorders, quadruples. In constructing an experiment, say, on the simple ranking of

objects with respect to a certain property, the design is ordinarily such that connectivity and antisymmetry of the relation are satisfied, because for each pair of objects the subject is required to decide the ranking one way or the other, but not in both directions. Analysis of the data then reduces to searching for intransitive triads.

Vaught (1954) has provided a useful criterion for certain classes of relational systems to be axiomatizable by means of a universal sentence. A straightforward analysis of his proof yields immediately the following criterion for finitary theories of measurement.

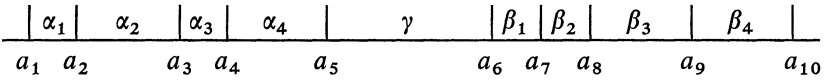
A finitary theory of measurement $\boldsymbol{K}$ *is axiomatizable by a universal sentence, if and only if* $\boldsymbol{K}$ *is closed under subsystems and there is an integer* n *such that, if any finite relational system* $\mathfrak{A}$ *has the property that every subsystem of* $\mathfrak{A}$ *with no more than* n *elements is in* $\boldsymbol{K}$*, then* $\mathfrak{A}$ *is in* $\boldsymbol{K}$.

Though classes of finite simple orderings and finite semiorders are two examples of finitary theories of measurement axiomatizable by a universal sentence, there are interesting examples of finitary theories of measurement closed under subsystems which are *not* axiomatizable by a universal sentence. We now turn to the proof for one such case.

Let $\boldsymbol{F}$ be the class of all finitary relational systems of type $\langle 4 \rangle$ imbeddable in the numerical relational system $\langle \mathrm{Re}, \Delta \rangle$. A wide variety of sets of empirical data are in $\boldsymbol{F}$. In fact, all sets of psychological data based upon judgments of differences of sensation intensities or of differences in utility qualify as candidates for membership in $\boldsymbol{F}$. For example, in an experiment concerned with the subjective measurement of loudness of n sounds, the appropriate empirical data would be obtained by asking subjects to compare each of the n sounds with every other and then to compare the difference of loudness in every pair of sounds with every other. More elaborate interpretations are required to obtain appropriate data on utility differences for individuals or social groups (cf. Davidson *et al.*, 1957; Suppes and Winet, 1955).[11†] It may be of some interest to mention one probabilistic interpretation closely related to the classical scaling method of paired comparisons. Subjects are asked to choose only between objects, but they are asked to make this choice a number of times. There are many situations in which they vacillate in their choice, and the probability p_{xy} that x will be chosen over y may be estimated from the relative frequency with which x is so chosen. From inequalities of the form $p_{xy} \leqq p_{zw}$ we may obtain a set of empirical data, that is, a

finite relational system of type $\langle 4 \rangle$, which is a candidate for membership in $\boldsymbol{F}$. The intended interpretation is that, if $p_{xy} \geqq \frac{1}{2}$ and $p_{zw} \geqq \frac{1}{2}$, then $p_{xy} \leqq p_{zw}$ if and only if the difference in sensation intensity or difference in utility between x and y is equal to or less than that between z and w, the idea being, of course, that if x and y are closer together than z and w in the subjective scale, then the relative frequency of choice of x over y is closer to one-half than that of z over w.

Before formally proving that the theory of measurement $\boldsymbol{F}$ is not axiomatizable by a universal sentence, we intuitively indicate for a relational system of ten elements the kind of difficulty which arises in any attempt to axiomatize $\boldsymbol{F}$. Let the ten elements be $a_1, \ldots, a_{10}$ ordered as shown on the following diagram with atomic intervals given the designations indicated.



Let α be the interval (a_1, a_5), let β be the interval (a_6, a_{10}), and let y be larger than α or β. We suppose further that $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ is equal in size to $\beta_2, \beta_4, \beta_1, \beta_3$, respectively, but α is less than β.[12]

The size relationships among the remaining intervals may be so chosen that any subsystem of nine elements is imbeddable in $\langle \mathrm{Re}, \Delta \rangle$, whereas the full system of ten elements is clearly not.

Generalizing this example and using the criterion derived from Vaught's theorem we now prove:

THEOREM: *The theory of measurement $\boldsymbol{F}$ is not axiomatizable by a universal sentence.*

Proof: In order to apply the criterion of axiomatizability by a universal sentence, we need to show that for every n there is a finite relational system $\mathfrak{A}$ of type $\langle 4 \rangle$ such that every subsystem of $\mathfrak{A}$ with n elements in its domain is in $\boldsymbol{F}$ but $\mathfrak{A}$ is not.

To this end, for every even integer $n = 2m \geqq 10$ we construct a finite relational system $\mathfrak{A}$ of type $\langle 4 \rangle$ such that every subsystem of $2m-1$ elements is in $\boldsymbol{F}$. (A fortiori every subsystem of $2m-k$ elements for $k < 2m$ is in $\boldsymbol{F}$.) To make the construction both definite and compact, we take numbers as elements of the domain and disrupt exactly one numerical relationship. Let now m be an even integer equal to or greater than 10. The selection of numbers $a_1, \ldots, a_{2m}$ may be most easily described by

specifying the numerical size of the atomic intervals. We define $\alpha_i = a_{i+1} - a_i$ for $i = 1, \ldots, m-1$ and $\beta_i = a_{m+i+1} - a_{m+i}$ for $i = 1, \ldots, m-1$. We then set $a_1 = 1$, $\alpha_i = 2^i$ for $i = 1, \ldots, m-1$, and $a_{m+1} = 2^{2m}$. In fixing the size of β_i, we have two cases to consider depending on the parity of m.

Case 1: *m is even.* Then $m-1$ is odd, and we set $\beta_i = \alpha_{i/2}$ for $i = 2, 4, \ldots, m-2$ and $\beta_i = \alpha_{(m+i-1)/2}$ for $i = 1, 3, \ldots, m-1$.

Case 2: *m is odd.* Then $m-1$ is even, and we set $\beta_i = \alpha_{i/2}$ for $i = 2, 4, \ldots, m-1$ and $\beta_i = \alpha_{(m+i)/2}$ for $i = 1, 3, \ldots, m-2$. Thus if $n = 2m = 12$, we have $\alpha_1 = \beta_2$, $\alpha_2 = \beta_4$, $\alpha_3 = \beta_1$, $\alpha_4 = \beta_3$, $\alpha_5 = \beta_5$. With the set $A = \{a_1, \ldots, a_{2m}\}$ defined, we now define the relation D as the expected numerical relation except for permutations of a_1, a_m, a_{m+1} and a_{2m}. If $x, y, z, w \in A$ and $\langle x, y, z, w \rangle$ is not some permutation of $\langle a_1, a_m, a_{m+1}, a_{2m} \rangle$, then $\langle x, y, z, w \rangle \in D$ if and only if

$$(1) \qquad x - y \leqq z - w.$$

Moreover, let $a = a_1$, $b = a_m$, $c = a_{m+1}$, $d = a_{2m}$. Then we put the following nine permutations of $\langle a, b, c, d \rangle$ in D:

$$\begin{array}{lll} \langle b, a, d, c \rangle & \langle a, b, d, c \rangle & \langle c, b, d, a \rangle \\ \langle b, d, a, c \rangle & \langle a, c, d, b \rangle & \langle c, d, a, b \rangle \\ \langle b, d, c, a \rangle & \langle a, d, c, b \rangle & \langle c, d, b, a \rangle. \end{array}$$

(These nine permutations correspond exactly to the strict inequalities following from $b - a < d - c$. All nine are needed to make the subsystems of $\langle A, D \rangle$ have the appropriate properties.)

From the choice of the numbers in A and the definition of D, it is obvious that $\langle A, D \rangle$ is not imbeddable in $\langle \mathrm{Re}, \Delta \rangle$, that is, that $\langle A, D \rangle$ is not in $\boldsymbol{F}$; for the atomic intervals between a_1 and a_m must add up to a length equal to the sum of the atomic intervals between a_{m+1} and a_{2m}, but by hypothesis the interval (a_1, a_m) is less than the interval (a_{m+1}, a_{2m}). It remains to show that every subsystem of $2m-1$ elements is in $\boldsymbol{F}$. Two cases naturally arise.

Case 1: The element omitted in the subsystem is a_1, a_m, a_{m+1} or a_{2m}. Then the nine permutations of (2) are not in D restricted to the subsystem, and the subsystem is not merely imbeddable in $\langle \mathrm{Re}, \Delta \rangle$, but by virtue of (1) is a subsystem of it.

Case 2: The element omitted is neither a_1, a_m, a_{m+1} nor a_{2m}. Let a_i be the element not in the subsystem. There are two cases to consider.

Case 2a: $a_i < a_m$. For this situation we may use for our numerical

assignment the function f defined by $f(a_{i-j})=a_{i-j}+1$ for $j=1,\ldots,i-1$, $f(a_{i+j})=a_{i+j}$ for $j=1,\ldots,n-i$. It is straightforward but tedious to verify that f is a numerical assignment, that is, that it preserves the relation D as defined by (1) and (2). Only two observations are crucial to this verification. First, regarding atomic intervals (in the full system), if $a_{i-j+1}-a_{i-j}=a_{k+1}-a_k$ for $k>i$, then $f(a_{i-j+1})-f(a_{i-j})=(a_{i-j+1}-1)-(a_{i-j}-1)=a_{k+1}-a_k=f(a_{k+1})-f(a_k)$. Second, the numbers in A were so chosen that, if $x, y, z, w\in A$, and (z, w) is not an atomic interval, and $(x, y)\neq(z, w)$ and $x-y\leqq z-w$, then $x-y+2\leqq z-w$. Then it is clear from the definition of f that $f(x)-f(y)\leqq f(z)-f(w)$. (Note that the above implies the weaker result that no two distinct nonatomic intervals have the same size.)

Case 2b: $a_i>a_m+1$. Here we may use a numerical assignment f defined, as would be expected from the previous case, by $f(a_{i-j})=a_{i-j}$ for $j=1,\ldots,i-1$, $f(a_{i+j})=a_{i+j}+1$ for $j=1,\ldots,n-i$. This completes the proof of the theorem.

It would be pleasant to report that we could prove a stronger result about the theory of measurement $\boldsymbol{F}$, namely, that it is not finitely axiomatizable. Unfortunately, there seems to be a paucity of tools available for studying such questions for classes of relational systems. However, we would like to state a conjecture which if true would provide one useful tool for studying the finite axiomatizability of finitary theories of measurement like $\boldsymbol{F}$ which are closed under submodels. We say that two sentences are *finitely equivalent* if and only if they are satisfied by the same finite relational systems, and we conjecture: *If* S *is a sentence such that if it is satisfied by a finite model it is satisfied by every submodel of the finite model, then there is a universal sentence finitely equivalent to* S. If this conjecture is true, it follows that any finitary theory of measurement closed under submodels is finitely axiomatizable if and only if it is axiomatizable by a universal sentence.

The proof (or disproof) of this conjecture appears difficult. It easily follows from Tarski's results (1954) on universal (arithmetical) classes in the wider sense that, if the finitistic restrictions are removed throughout in the conjecture, the thus modified conjecture is true; for the class of relational systems satisfying S, being closed under submodels, is a universal class in the wider sense and is axiomatizable by a denumerable set of universal sentences. Since S is logically equivalent to this set of

universal sentences, it is a logical consequence of some finite subset of them; but because it implies the full set, it also implies the finite subset and is thus equivalent to it.

Our conjecture is one concerning the general theory of models and its pertinence is not restricted to theories of measurement. In conclusion we should like to mention an unsolved problem typical of those which arise in the special area of measurement. *Let R be any binary numerical relation definable in an elementary manner in terms of plus and less than. Is the finitary theory of measurement of all systems imbeddable in R finitely axiomatizable?* (If our conjecture about finite models is true, then the theory of measurement F is not finitely axiomatizable and shows that the answer to this problem is negative for quaternary relations definable in terms of plus and less than.)

NOTES

[1] We would like to record here our indebtedness to Alfred Tarski, whose clear and precise formulation of the mathematical theory of models has greatly influenced our presentation (Tarski, 1954, 1955). Although our theories of measurement do not constitute special cases of the arithmetical classes of Tarski, the notions are closely related, and we have made use of results and methods from the theory of models.

[2] Although in most mathematical contexts imbeddability is defined in terms of isomorphism rather than homomorphism, for theories of measurement this is too restrictive. However, the notion of homomorphism used here is actually closely connected with isomorphic imbeddability and the facts are explained in detail in Section II.

[3] In some contexts we shall say that the class $\boldsymbol{K}$ is *a theory of measurement of type s relative to* $\mathfrak{R}$. Notice that a consequence of this definition is that if $\boldsymbol{K}$ is a theory of measurement, then so is every subclass of $\boldsymbol{K}$ closed under isomorphism. Moreover, the class of all systems imbeddable in members of $\boldsymbol{K}$ is also a theory of measurement.

[4] In this connection see Sierpinski (1934, Section 7, pp. 141ff.) in particular *Proposition* C_{75}, where of course different terminology is used.

[5] It is sufficient here to consider a relational system isomorphic to the ordering of the ordinals of the second number class or to the lexicographical ordering of all pairs of real numbers.

[6] A simple ordering is imbeddable in $\langle \mathrm{Re}, \leqslant \rangle$ if and only if it contains a countable dense subset. For the exact formulation and a sketch of a proof, see Birkhoff (1948, pp. 31–32, Theorem 2).

[7] The word 'countable' means at most denumerable, and it refers to the cardinality of the domains of the relational systems.

[8] See Luce (1956, Section 2, p. 181). The axioms given here are actually a simplification of those given by Luce.

[9] The authors are indebted to the referee for pointing out the work by Hailperin (1954), which suggested this general definition.

[10†] The proofs of both these facts about H are very similar to the corresponding proofs in Suppes and Winet (1955; Article 8 in this volume).

[11†] Article 8 in this volume.

[12] Essentially this example was first given in another context by Herman Rubin to show that a particular set of axioms is defective.

5. MEASUREMENT, EMPIRICAL MEANINGFULNESS, AND THREE-VALUED LOGIC*[1]

I. INTRODUCTION

The predominant current opinion appears to be that it is scarcely possible to set up criteria of empirical meaningfulness for individual statements. What is required, it is said, is an analysis of theories taken as a whole. There is even some skepticism regarding this, and it has been romantically suggested that the entire fabric of experience and language must be considered and taken into account in any construction of general categories of meaning or analyticity. What I have to say makes no contribution to the attempt to find a general criterion of meaning applicable to arbitrary statements. Rather I am concerned to exemplify a general method that will yield specific positive criteria for specific branches of science.

A brief analysis of two simple examples will indicate the sort of thing I have in mind. Consider the statement:

(i) The mass of the sun is greater than 10^6.

If a physicist were asked if (i) is true or false, he would most likely reply that it depends on what unit of mass is implicitly understood in uttering (i). On the other hand, if we were to ask him about the truth of the sentence:

(ii) The mass of the sun is at least ten times greater than that of the earth,

he would, without any reservation about units of measurement, state that (ii) is true, and perhaps even add that its truth is known to every schoolboy. Now my main point is that we may insist that our systematic language of physics (or of any other empirical science) has no hidden references to units of measurement. The numerals occurring in the

* Reprinted from *Measurement: Definitions and Theories* (ed. by C. West Churchman and P. Ratoosh), Wiley, New York, 1959, pp. 129–143.

language are understood to be designating "pure" numbers. An excellent example of a physical treatise written without reference to units is provided by the first two books of Newton's *Principia.* (Units are introduced in the consideration of data in Book III, and occasionally in examples in the earlier books.) Newton avoids any commitment to units of measurement by speaking of one quantity being proportional to another or standing in a certain ratio to it. Thus he formulates his famous second law of motion:

> The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed [Cajori edition, p. 13].

Systematic reasons for adopting Newton's viewpoint as the fundamental one are given in later sections. My only concern at the moment is to establish that adoption of this viewpoint does not represent a gross violation of the use of physical concepts and language by physicists. It seems obvious that, in *using* a unitless language, we would not find occasion to use (i), for there would be no conceivable way of establishing its truth or falsity, either by empirical observation or logical argument. In contrast, (ii) would be acceptable. Yet it is difficult to see how to develop a simple and natural syntactical or semantical criterion within, say, a formal language for expressing the results of measurements of mass, which would rule out sentences like (i) and admit sentences like (ii). The central purpose of this paper is to explore some of the possibilities for classifying as meaningless well-formed sentences like (i), or, more exactly, the analogues of (i) in a formalized language. Formalization of a certain portion of the unitless language of physicists is not absolutely necessary for expressing the ideas I want to put forth, but it is essential to a clear working out of details. Moreover, the exact formal construction seems to pose some interesting problems which could scarcely be stated for a natural language. In the final section, the possibility is explored of interpreting this formalized language in terms of a three-valued logic of truth, falsity, and meaninglessness.

II. INVARIANCE AND MEANINGFULNESS

In connection with any measured property of an object, or set of objects, it may be asked how unique is the number assigned to measure the

property. For example, the mass of a pebble may be measured in grams or pounds. The number assigned to measure mass is unique once a unit has been chosen. A more technical way of putting this is that the measurement of mass is unique up to a similarity transformation.[2] The measurement of temperature in °C or °F has different characteristics. Here an origin as well as a unit is arbitrarily chosen: technically speaking, the measurement of temperature is unique up to a linear transformation.[3] Other formally different kinds of measurement are exemplified by (1) the measurement of probability, which is absolutely unique (unique up to the identity transformation), and (2) the ordinal measurement of such physical properties as hardness of minerals, or such psychological properties as intelligence and racial prejudice. Ordinal measurements are commonly said to be unique up to a monotone-increasing transformation.[4]

Use of these different kinds of transformations is basic to the main idea of this paper. An empirical hypothesis, or any statement in fact, which uses numerical quantities is empirically meaningful only if its truth value is invariant under the appropriate transformations of the numerical quantities involved. As an example, suppose a psychologist has an ordinal measure of *I.Q.*, and he thinks that scores $S(a)$ on a certain new test T have ordinal significance in ranking the intellectual ability of people. Suppose further that he is able to obtain the ages $A(a)$ of his subjects. The question then is: Should he regard the following hypothesis as empirically meaningful?

HYPOTHESIS 1: *For any subjects a and b if* $S(a)/A(a)<S(b)/A(b)$, *then I.Q.* $(a)<I.Q.\ (b)$.

From the standpoint of the invariance characterization of empirical meaning, the answer is negative. To see this, let $I.Q.\ (a)\geqslant I.Q.\ (b)$, let $A(a)=7$, $A(b)=12$, $S(a)=3$, $S(b)=7$. Make no transformations on the *I.Q.* data, and make no transformations on the age data. But let ϕ be a monotone-increasing transformation which carries 3 into 6 and 7 into itself. Then we have

$$\tfrac{3}{7}<\tfrac{7}{12},$$

but

$$\tfrac{6}{7}\geqslant\tfrac{7}{12},$$

and the truth value of Hypothesis 1 is not invariant under ϕ.

The empirically significant thing about the transformation characteristic of a quantity is that it expresses in precise form how unique is the structural isomorphism between the empirical operations used to obtain a given measurement and the corresponding arithmetical operations or relations. If, for example, the empirical operation is simply that of ordering a set of objects according to some characteristic, then the corresponding arithmetical relation is that of less than (or greater than), and any two functions which map the objects into numbers in a manner preserving the empirical ordering are adequate. More exactly, a function f is adequate if, and only if, for any two objects a and b in the set, a stands in the given empirical relation to b if and only if

$$f(a) < f(b).^{5}$$

It is then easy to show that if f_1 and f_2 are adequate in this sense, then they are related by a monotone-increasing transformation. Only those arithmetical operations and relations which are invariant under monotone-increasing transformations have any empirical significance in this situation.

The key notion referred to in the last sentence is that of invariance. In order to make the notion of invariance or the related notion of meaningfulness completely precise, we can do one of two things: set up an exact set-theoretical framework for our discussion (e.g., for classical mechanics, see McKinsey and Suppes, 1955), or formalize a language adequate to express empirical hypotheses and facts involving numerical quantities. Here we shall formalize a simple language for expressing the results of mass measurements. It should be clear that the method of approach is applicable to any other kind of measurement, or combinations thereof.

III. EMPIRICAL MEANINGFULNESS IN THE LANGUAGE L_M

To avoid many familiar details, we shall use as a basis the formal language of Tarski's monograph (1951) enriched by individual variables 'a', 'b', 'c',..., 'a_1', 'b_1', 'c_1',..., the individual constants: $o_1, \ldots, o_{10}$, which designate ten, not necessarily distinct, physical objects, and the mass term 'm', where '$m(a)$' designates a real number, the mass of a. The values of the individual variables are physical objects. The numerical variables are 'x', 'y', 'z',..., 'x_1', 'y_1', 'z_1',.... Tarski's numerical constants are: 1, 0, -1.

We shall include, for purposes of examples, numerical constants for the positive and negative integers less than 100 in absolute value. The operation signs are those for addition and multiplication. We also include the standard sign for exponentiation with the fixed base 2. A *term* is any arithmetically meaningful expression built up from this notation in the usual manner. (We omit an exact definition.) Thus the following are terms: $m(a)$, $5 \cdot m(a)+3$, $2+1$, $x+3$, 2^x. Our two relation symbols are the usual sign of equality and the greater than sign. An *atomic formula* is then an expression of the form

$$(\alpha = \beta), \quad (\alpha > \beta),$$

where α and β are terms with the restriction in the case of $(\alpha > \beta)$ that α and β are both numerical terms, that is, neither α nor β is an individual variable or constant. When no confusion will result, parentheses are omitted. *Formulas* are constructed from atomic formulas by means of sentential connectives and quantifiers. The symbol '$-$' is used for negation; the ampersand '&' for conjunction; the symbol '$\vee$' for disjunction (to be read 'or'); the arrow '$\rightarrow$' for implication (to be read 'if... then...'); the double arrow '$\leftrightarrow$' for equivalence (to be read 'if and only if'); the reverse '$\exists$' is the existential quantifier; and the upside down '$\forall$' the universal quantifier. Thus the following are formulas: $(\exists x)(m(a) = x)$, $(\exists x)(\exists y)(x > y)$, $0 > x \rightarrow m(b) > x$. We also use the standard symbol '$\neq$' for negating an equality. A formula is a *sentence* if it contains no free variables, that is, every occurrence of a variable is bound by some quantifier.

Sentences are true or false, but unlike the situation in the language of Tarski's monograph (1951), the truth or falsity of many sentences in the language L_M constructed here depends on empirical observation and contingent fact. For example, the truth of the sentence:

$$(1) \qquad (\exists a)(\forall b)(b \neq a \rightarrow m(a) > 5 \cdot m(b))$$

is a matter of physics, not arithmetic.

Pursuing now in more detail the remarks in the first section, the intuitive basis for our classification of certain formulas of L_M as empirically meaningless may be brought out by considering the simple sentence:

$$(2) \qquad m(o_1) = 4.$$

It must first be emphasized that in the language L_M, the numeral '4' occurring in Sentence 2 designates a "pure" number. There is no convention, explicit or implicit, that '4' stands for '4 g', '4 lb', or the like. It is to be clearly understood that no unit of mass is assumed in the primitive notation of L_M. With this understanding in mind, it is obvious that no experiment with apparatus for determining the masses of physical objects could determine the truth or falsity of Sentence 2. It is equally obvious that no mathematical argument can settle this question. On the other hand, it is clear that sentences like:

(3) $\quad m(o_1) > m(o_2)$

or

(4) $\quad m(o_3) = 5 \cdot m(o_4),$

which are concerned with numerical relations between the masses of certain objects can be determined as true or false on the basis of experiment without prior determination of a unit of mass.

It seems to me that the use of 'pure' numerals in L_M is more fundamental than the use of what we may term 'unitized numerals'. The justification of this view is that the determination of units and an appreciation of their empirical significance comes *after*, not before, the investigation of questions of invariance and meaningfulness. The distinction between Sentence 2 and the other three Sentences 1, 3, and 4 is that the latter sentences remain true (or false) under any specification of units. In other words, the truth value of these sentences is independent of the arbitrary choice of a unit. Paraphrasing Weyl, we may say[6]: *only the numerical masses of bodies relative to one another have an objective meaning.*

My claim regarding fundamentals may be supported by an axiomatic, operational analysis of any actual experimental procedure for measuring mass. Most such procedures may be analyzed in terms of three formal notions: the set A of physical objects, a binary operation Q of comparison, and a binary operation $*$ of combination. The formal task is to show that under the intended empirical interpretation the triple $\mathfrak{A} = \langle A, Q, * \rangle$ has such properties that it may be proved that there exists a real-valued function $\mathbf{m}$ defined on A such that for any a and b in A

(i) $\quad a \, Q \, b$ if and only if $\mathbf{m}(a) \leqslant \mathbf{m}(b)$,

(ii) $\quad \mathbf{m}(a * b) = \mathbf{m}(a) + \mathbf{m}(b)$.

The empirically arbitrary character of the choice of a unit is established by showing that the functional composition of any similarity transformation ϕ with the function $\mathbf{m}$ yields a function $\phi \circ \mathbf{m}$ which also satisfies (i) and (ii), where $\circ$ is the operation of functional composition.[7†]

We may think of such an operational analysis as supporting the choice of L_M, where the term 'm' of L_M designates a numerical representing function satisfying (i) and (ii). Roughly speaking, because this representing function is only unique up to a similarity transformation, we then expect any sentence to be empirically meaningful in L_M if and only if its truth value is the same when 'm' is replaced by any expression which designates multiplication of the representing function by a positive number. However, there are certain difficulties with deciding exactly how to make this intuitive definition of empirical meaningfulness precise. For example, if the definition applies to any sentences, then we have the somewhat paradoxical result that Sentence 2 and its negation are both empirically meaningless, but their disjunction:

$$(5) \qquad m(o_1) = 4 \vee m(o_1) \neq 4$$

is meaningful, since it is always true.

To facilitate our attempts to meet this problem, we first need to introduce the semantical notion of a *model* of L_M. For simplicity in defining the notion of model, and without any loss of generality, we shall from this point on consider L_M as not having any individual constants that designate physical objects.

On this basis, a model $\mathfrak{M}$ of L_M is an ordered triple $\langle \mathfrak{S}, A, \mathbf{m} \rangle$ where

(i) $\mathfrak{S}$ is the usual system of real numbers under the operations of addition, multiplication, and exponentiation with the base 2, and the relation less than with the appropriate numbers corresponding to their numerical designations in L_M[8];

(ii) A is a finite, nonempty set;

(iii) $\mathbf{m}$ is a function on A which takes positive real numbers as values.

The intended interpretation of A is as a set of physical objects whose masses are being determined; the function $\mathbf{m}$ is meant to be a possible numerical function used to represent experimental results. We assume the semantical notion of *satisfaction* and suppose it to be understood under what conditions a sentence of L_M is said to be *satisfied* in a model $\mathfrak{M}$. Roughly speaking, a sentence S of L_M is satisfied in $\mathfrak{M} = \langle \mathfrak{S}, A, \mathbf{m} \rangle$

if S is true when the purely arithmetical symbols of S are given the usual interpretation in terms of $\mathfrak{S}$, when the individual variables occurring in S range over the set A, and when the symbol 'm', if it occurs in S, is taken to designate the function **m**.

We say that a sentence of L_M is *arithmetically true* if it is satisfied in every model of L_M. And we deal with the arithmetical truth of formulas with free variables by considering the truth of their *closures*. By the closure of a formula we mean the sentence resulting from the formula by adding sufficient universal quantifiers to bind all free variables in the formula. Thus '$(\forall a)(m(a)>0)$' is the closure of '$m(a)>0$', and is also the closure of itself.

Using these notions, we may define meaningfulness by means of the following pair of definitions.

DEFINITION 1: *An atomic formula S of L_M is empirically meaningful if and only if the closure of the formula*

$$\alpha > 0 \rightarrow (S \leftrightarrow S(\alpha))$$

is arithmetically true for every numerical term α, where $S(\alpha)$ results from S by replacing any occurrence of 'm' in S by the term α, followed by the multiplication sign, followed by 'm'.[9]

If, for example,

$$S = '\,m(a) > m(b)'$$
$$\alpha = '(2 + 1)',$$

then

$$S(\alpha) = '(2 + 1)\cdot m(a) > (2 + 1)\cdot m(b)'.$$

DEFINITION 2: *A formula S of L_M is empirically meaningful in sense A if and only if each atomic formula occurring in S is itself empirically meaningful in the sense of Definition* 1.

It is clear on the basis of Definitions 1 and 2 that Sentence 5 is not empirically meaningful in sense A.

On the other hand, there is a certain logical difficulty, within ordinary two-valued logic, besetting the set of true formulas which are meaningful in sense A. Following Tarski (1930), a set of formulas is a *deductive system* if and only if the set is closed under the relation of logical consequence, that is, a formula which is a logical consequence of any subset of formulas in the given set must also be in the set. Clearly it is most

desirable to have the set of meaningful true formulas about any phenomenon be a deductive system, but we have for the present case the following negative result.

THEOREM 1: *The set of formulas of L_M which are meaningful in sense A and whose closures are true is not a deductive system.*

Proof: The true sentence:

$$(\forall x)\,(x > 2 \rightarrow x > 1)$$

is meaningful in sense A, but the following logical consequence of it is not:

$$m(o_1) > 2 \rightarrow m(o_1) > 1\,,$$

for the two atomic sentences '$m(o_1)>2$' and '$m(o_1)>1$' are both meaningless in the sense of Definition 1.

To be sure, there are some grounds for maintaining that formulas that are empirically meaningless may play an essential deductive role in empirical science, but *prima facie* it is certainly desirable to eliminate them if possible.

A second objection to Definition 1 is that, by considering numerical terms α rather than similarity transformations, we have in effect restricted ourselves to a denumerable number of similarity transformations because the number of such terms in L_M is denumerable. The intuitive idea of invariance with respect to *all* similarity transformations may be caught by a definition of meaningfulness which uses the concept of two models of L_M being related by a similarity transformation. (The operation $\circ$ referred to in the definition is that of functional composition.)

DEFINITION 3: *Let $\mathfrak{M}_1 = \langle \mathfrak{S}, A_1, \mathbf{m}_1 \rangle$ and $\mathfrak{M}_2 \langle \mathfrak{S}, A_2, \mathbf{m}_2 \rangle$ be two models of L_M. Then $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are related by a similarity transformation if and only if:*

(i) $A_1 = A_2$.

(ii) *There is a similarity transformation ϕ such that*

$$\phi \circ \mathbf{m}_1 = \mathbf{m}_2\,.$$

Using these notions, we may replace Definitions 1 and 2 by the following:

DEFINITION 4: *A formula S of L_M is empirically meaningful in sense B if and only if S is satisfied in a model $\mathfrak{M}$ of L_M when and only when it is*

satisfied in every model of L_M related to $\mathfrak{M}$ by a similarity transformation.

Unfortunately, we have for meaningfulness in sense B a result analogous to Theorem 1.

THEOREM 2: *Let $\mathfrak{M}$ be a model of L_M. Then the set of all formulas which are meaningful in sense B and which are satisfied in $\mathfrak{M}$ is not a deductive system.*

Proof: Consider the two sentences:

$$(1) \qquad (\forall a)(\forall b)(a = b \rightarrow (m(a) = 2 \rightarrow m(b) = 2))$$

$$(2) \qquad (\forall a)(\forall b)(a = b).$$

It is easy to verify that Sentences 1 and 2 are satisfied in any model whose set A has exactly one element, and are meaningful in sense B, yet they have as a logical consequence the sentence:

$$(3) \qquad (\forall a)(\forall b)(m(a) = 2 \rightarrow m(b) = 2),$$

which is not meaningful in sense B. That this is so may be seen by considering a model with at least two objects with different masses. Let $A = \{o_1, o_2\}$, and let $\mathfrak{M}_1 = \langle \mathfrak{S}, A, \mathbf{m}_1 \rangle$ be such that $\mathbf{m}_1(o_1) = 2$ and $\mathbf{m}_2(o_2) = 3$, and let $\mathfrak{M}_2 = \langle \mathfrak{S}, A, \mathbf{m}_2 \rangle$ be related to $\mathfrak{M}_1$ by the similarity transformation $\phi(x) = 2x$. Thus $\mathbf{m}_2(o_1) = 4$ and $\mathbf{m}_2(o_2) = 6$. It is then easily checked that Sentence 3 is satisfied in $\mathfrak{M}_2$ but not in $\mathfrak{M}_1$.

The negative result of these two theorems indicates the difficulties of eliminating the appearance of empirically meaningless statements in valid arguments with meaningful premises. We return to this point in the next section in connection with consideration of a three-valued logic.

On the other hand, we do have the positive result for both senses of meaningfulness that the set of meaningful formulas is a Boolean algebra; more exactly, the set of such formulas under the appropriate equivalence relation is such an algebra. Here we carry out the construction only for sense B. We consider the theory of Boolean algebras as based on six primitive notions: the nonempty set B of elements; the operation $+$ of addition which corresponds to the sentential connective 'or'; the operation $\cdot$ of multiplication which corresponds to the connective 'and'; the operation $\bar{x}$ of complementation which corresponds to negation; the zero element 0, which corresponds to the set of logically invalid formulas; and the unit element 1, which corresponds to the set of logically valid

formulas. We omit stating familiar postulates on these notions which a Boolean algebra must satisfy.

Let E be the set of formulas which are empirically meaningful in sense B. We define the equivalence class of a formula S in E as follows: $[S]$ *is the set of all formulas* S' *in* E *which are satisfied in exactly the same models* $\mathfrak{M}$ *of* L_M *as* S *is.* Let $\mathbf{E}$ be the set of all such equivalence classes; obviously $\mathbf{E}$ is a partition of E. The zero element $\mathbf{0}$ is the set of formulas in E which are satisfied in no model of L_M; the unit element $\mathbf{1}$ is the set of formulas in E which are satisfied in all models of L_M. If S and T are in E, then $[S]+[T]$ is the set of all formulas in E which are satisfied in the models of L_M in which either S or T is satisfied. If S and T are in E, then $[S]\cdot[T]$ is the set of all formulas in E which are satisfied in those models in which both S and T are satisfied. Finally if S is in E, then $[\bar{S}]$ is the set of formulas which are satisfied in a model if and only if S is not satisfied in the model. On the basis of these definitions, it is straightforward but tedious to prove the following:

THEOREM 3: *The system* $\langle \mathbf{E}, +, \cdot, -, \mathbf{0}, \mathbf{1} \rangle$ *is a Boolean algebra.*
(The proof is omitted.)

I interpret this theorem as showing that the set of meaningful formulas in sense B of L_M has a logical structure identical with that of classical logic. In connection with other systems of measurement for which the set of transformations referred to in the analogue of Definition 3 is not a group, this classical Boolean structure does not necessarily result.

Exponentiation was introduced into L_M deliberately to illustrate the sensitivity of the decidability of meaningfulness to the strength of L_M. The problem of decidability for the arithmetical language of Tarski's monograph mentioned earlier is open when his language is augmented by notation for exponentiation to a fixed base. It seems unlikely that the decidability of meaningfulness in L_M can be solved without solving this more general problem. If L_M is weakened by deleting exponentiation to the base 2, then it easily follows from Tarski's well-known result that meaningfulness is decidable. On the other hand, if L_M is strengthened to include sufficient elementary number theory to yield undecidability of whether, for instance, a given term designates zero, then meaningfulness is not decidable, for the meaningfulness of formulas of the form $m(a)=\mathbf{t}$, where $\mathbf{t}$ is a numerical term, would not be decidable.

IV. A THREE-VALUED LOGIC FOR L_M

Since sentences like '$(\forall a)(m(a)>2)$' of L_M cannot be determined as true or false on the basis either of logical argument or of empirical observation, it is natural to ask what are the consequences of assigning them the truth value *meaningless*, which we designate by 'μ', and reserving the values *truth* and *falsity* for meaningful sentences, which we designate by 'T' and 'F', respectively. The first thing to be noticed is that meaningfulness in sense B does not lead to a truth-functional logic in these three values. This may be seen by considering two examples. The component sentences of the sentence:

$$(\exists a)\,(m(a)=1)\vee -(\exists a)\,(m(a)=1)$$

have the value μ but the whole sentence is meaningful in sense B and has the value T. On the other hand, the component sentences of:

$$(\exists a)\,(m(a)=1)\vee(\exists b)\,(m(b)=2)$$

have the value μ and so does the whole sentence. Thus these two examples taken together show that disjunction is not truth-functional for a three-value logic of meaningfulness in sense B.

The state of affairs for meaningfulness in sense A is much better; it does lead to a truth-functional logic in the three values, T, F, and μ. The appropriate truth tables are easily found by using the simple observation that a formula has the value μ if any well-formed part of it has that value. Thus as the tables for negation and conjunction we have:

S	$-S$
T	F
F	T
μ	μ

$\&$	T	F	μ
T	T	F	μ
F	F	F	μ
μ	μ	μ	μ

Tables for the sentential connectives of disjunction, implication, and equivalence follow at once from the standard definitions of these connectives in terms of negation and conjunction. On the other hand, it is obvious that this three-valued logic is not functionally complete with respect to negation and conjunction. For example, we cannot define in terms of these two connectives a unary connective which assigns the value μ to formulas having the value T.

Besetting meaningfulness in sense A is the negative result of Theorem 1. This difficulty we shall meet head on by proposing a revision of the definition of the semantical notion of logical consequence. However, before turning to this definition, it will be advantageous to give a model-theoretic definition of meaningfulness which combines the virtues of sense A and sense B.

DEFINITION 5: *A formula S of L_M is empirically meaningful in sense C if and only if every atomic formula occurring in S is meaningful in sense B.*

It is easily verified that the truth tables just given are satisfied when the value μ signifies meaninglessness in sense C. Moreover, the exact analogue of the Boolean structure theorem for sense B (Theorem 3) can be proved for sense C.

To meet the difficulty of having formulas which are meaningless in sense C be logical consequences of formulas which are meaningful in sense C, a revision of the standard definition of *logical consequence* is proposed. For this purpose we need to widen the notion of a model to that of a *possible realization* of L_M. A model of L_M requires that the arithmetical symbols be interpreted in terms of the usual system of real numbers, but no such restriction is imposed on a possible realization. For example, any domain of individuals and any two binary operations on this domain provide a possible realization of the operation symbols of addition and multiplication. Details of the exact definition of a possible realization are familiar from the literature and will not be given here. This notion is used to define that of logical consequence, namely, a formula S of L_M is a *logical consequence* of a set A of formulas of L_M if S is satisfied in every possible realization in which all formulas in A are satisfied. We may then define:

DEFINITION 6: *Let S be a formula and A a set of formulas of L_M. Then S is a meaningful logical consequence of A if and only if S is a logical consequence of A and S is meaningful in sense C whenever every formula in A is meaningful in sense C.*

The central problem in connection with this definition is to give rules of inference for which it may be established that if A is a set of formulas meaningful in sense C, then S is a meaningful logical consequence of A if and only if S is derivable from A by use of the rules of inference.[10] For this purpose, we may consider any one of several systems of natural deduction. The eight essential rules are: rule for introducing premises;

rule for tautological implications; rule of conditional proof (the deduction theorem); rule of universal specification (or instantiation); rule of universal generalization; rule of existential specification; rule of existential generalization; and rule governing identities.[11] To these eight rules we add the *general restriction* that every line of a derivation must be a formula meaningful in sense *C*. This means, for instance, that in deriving a formula by universal specification from another formula we must check that the result of the specification is meaningful. This restriction entails that the modified rules of inference are finitary in character only if there is a decision procedure for meaningfulness in sense *C*. Remarks on this problem were made at the end of the previous section. Because we have modified the rules of inference only by restricting them to meaningful formulas, it follows easily from results in the literature on the soundness of standard rules of inference that:

THEOREM 4: *Let A be a set of formulas meaningful in sense C. If a formula S is derivable from A by use of the rules of inference subject to the general restriction just stated, then S is a meaningful logical consequence of A.*

Of considerable more difficulty is the converse question of completeness, namely, does being a meaningful logical consequence of a set of meaningful formulas imply derivability by the restricted rules? The following considerations suggest that the answer may be affirmative to this question. Let L_M* be a second language which differs from L_M in the following single respect: the one-place function symbol 'm' is replaced by the two-place function symbol 'r', where both argument places are filled by individual variables or constants. The intuitive interpretation of the formula '$r(a, b)=x$' is that the numerical ratio of the mass of a to the mass of b is the real number x, that is,

$$r(a, b) = m(a)/m(b).$$

Clearly every formula in L_M* is meaningful with respect to our intuitive criterion of invariance. (The practical objection to L_M* is that such a ratio language is tedious to work with and does not conform to ordinary practice in theoretical physics.) No restrictions on the rules of inference are required for L_M* and, consequently, the usual completeness result holds. The suggestion is to use translatability of meaningful formulas of L_M into L_M* to prove completeness of inferences from meaningful

formulas of L_M. The possible pitfall of this line of reasoning is that translatability requires certain arithmetical operations which are preserved in every model but not necessarily in every possible realization of L_M.

Certain aspects of this construction of a three-valued logic for L_M seem worthy of remark. In the first place, the construction has assumed throughout use of a two-valued logic in the informal metalanguage of L_M. In particular, ordinary two-valued logic is used in deciding if a given sentence of L_M is satisfied in a given model of L_M. On the other hand, the relation between sets of empirical data on mass measurements and models of L_M is one–many. The empirical content of the data is expressed not by a particular model but by an appropriate equivalence class of models. Consequently, sentences of L_M which are not invariant in truth value (in the two-valued sense) over these equivalence classes do not have any clear empirical meaning even though they have a perfectly definite meaning relative to any one model. Thus it seems to me that to call a formula like '$m(a)=5$' empirically meaningless is no abuse of ordinary ideas of meaningfulness, and in this particular situation accords well with our physical intuitions. If this is granted, the important conclusion to be drawn is that, for the language L_M, the three-valued logic constructed is intuitively more natural than the ordinary two-valued one.

NOTES

[1] I am indebted to Georg Kreisel for several helpful comments on an earlier draft of this article.

[2] A real-valued function ϕ is a similarity transformation if there is a positive number α such that for every real number χ

$$\phi(x) = \alpha x.$$

In transforming from pounds to grams, for instance, the multiplicative factor α is 453.6.

[3] A real-valued function ϕ is a linear transformation if there are numbers α and β with $\alpha > 0$ such that for every number x

$$\phi(x) = \alpha x + \beta.$$

In transforming from Centrigrade to Fahrenheit degrees of temperature, for instance, $\alpha = \frac{9}{5}$ and $\beta = 32$.

[4] A real-valued function ϕ is a monotone increasing transformation if, for any two numbers x and y, if $x < y$, then $\phi(x) < \phi(y)$. Such transformations are also called *order-preserving*.

[5] For simplicity we shall consider here only the arithmetical relation $<$. There is no other reason for excluding $>$.

[6] Weyl's original statement is with respect to Galileo's principle of relativity, "Only the motions of bodies (point-masses) relative to one another have an objective meaning" [1922, p. 152].

[7†] An axiomatic analysis in terms of these ideas may be found in Suppes (1951; Article 3 in this volume). However, the analysis given there may be criticized on several empirical counts; for example, the set A must be infinite.

[8] Technical details about $\mathfrak{S}$ are omitted. Characterization of models of the purely arithmetical part of L_M are familiar from the literature.

[9] In this definition and subsequently we follow, without explicit discussion, certain use-mention conventions. It would be diversionary to go into these conventions, and it seems unlikely any serious confusion will result from not being completely explicit on this rather minor point.

[10] Although two kinds of variables are used in L_M, we may easily modify L_M to become a theory with standard formalization in first-order predicate logic and thus consider only modification of standard rules of inference for first-order predicate logic.

[11] By various devices this list can be reduced, but that is not important for our present purposes. Exposition of systems of natural deduction which essentially use these eight rules is to be found in Copi (1954), Quine (1950), and Suppes (1957).

PART II

METHODOLOGY: PROBABILITY AND UTILITY

The six articles in this part represent over a decade of work on subjective probability and utility, primarily in the context of investigations that fall within the general area of decision theory. Articles 6 and 8 are closely related to the theory of measurement. Because of doubts about the possibility of measuring either subjective probability or utility, much of the theory of these subjects has been devoted to an explicit working out of the theory of measurement. Article 9 on the behavioristic foundations of utility is related closely to the articles in Part IV on the foundations of psychology. The discussion in this article of learning theory overlaps the more detailed analyses given in Articles 16 and 23. To those readers who want a quick survey of decision theory without confronting the technical problems, I would recommend Article 7 on the philosophical relevance of decision theory. Duncan Luce and I (1965) have attempted a much more substantial and technical survey in an article not reprinted here.

The last article in this part, Article 11 on probabilistic inference, makes the closest connection of any of the articles with much of the recent philosophical literature on induction. I think the line of attack begun in this article can be considerably extended, particularly in areas of experience and those parts of science not yet well organized from a theoretical standpoint. Above all, however, the problems raised about rational behavior at the end of this article seem to me the most important open problems that I have raised in any of the six articles in this part, and in this respect, the article is closely related to the tradition of analysis in decision theory exemplified by the first article of this part, Article 6.

From a general philosophical standpoint, the central theme of these six articles is the problem of characterizing and analyzing the elusive concept of rationality. I suppose it is clear to everyone who thinks about the matter very much that we are still only in the beginning stages of a satisfactory analysis, and there are many people who are skeptical of ever giving a systematic characterization that is intuitively satisfactory. I do not think we should yet aim or hope for anything that is complete, but,

as in the case of work in the foundations of mathematics over the past century, there is now some ground at least for believing that progress of a definite and objectively agreed upon sort is possible. The work that originates with the theory of games is turning out to be one of the most useful general lines of approach, even though the classical article by Milnor (1954) shows how treacherous and difficult it is to give an intuitively complete, but consistent list of attributes of a rational strategy in an uncertain situation, even when that situation is highly restricted. In many respects, the great classical tradition in economics, going back to Adam Smith, can be viewed as an attempt to work out a normative theory of rational behavior in economic contexts. The recent literature in normative economics has generalized the relatively narrow economic context to a wider context of decision or action, as exemplified, for example, in Arrow's classical book (1951).

I turn now to a more detailed consideration of the last two articles in this part. In an as yet unpublished book on welfare economics, Dr. Amartya K. Sen of the Delhi School of Economics, University of Delhi, India has made a number of acute comments on the grading principle of justice introduced in Article 10. The fundamental point he makes is that some possible relations J_i of *more just than* can violate Pareto optimality. The relation J_i is person i's preference ordering of the possible consequences accruing to him and the possible consequences accruing to the other person as well (in Article 10 I restricted the number of persons to two, but the generalization to n is straightforward and has been carried out by Dr. Sen). Here is a simple instance of Sen's demonstration of incompatibility with Pareto optimality. Consider two vectors of consequences $x=\langle x_1, x_2\rangle$ and $y=\langle y_1, y_2\rangle$. Let person 1 order these four consequences

$$x_2 \, P_1 \, y_1, y_1 \, P \, x_1 \quad \text{and} \quad x_1 \, P \, y_2,$$

and let person 2 order them thus

$$x_1 \, P_2 \, y_2, \quad y_2 \, P \, x_2, \quad \text{and} \quad x_2 \, P \, y_1.$$

Then according to Definition 5 (p. 159), we have $x \, J_1 \, y$ and $x \, J_2 \, y$, but on the other hand, for person 1, $y_1 \, P_1 \, x_1$ and for person 2, $y_2 \, P \, x_2$, whence by Pareto optimality, the appropriate social choice is y over x. This undesirable result follows whenever each man presumes to know his

neighbor's preferences better than the neighbor does himself. Thus in Sen's example, person 1 thinks that for 2, x_2 is better than y_2, even though 2 thinks the opposite. Person 2 judges similarly the ranking of x, and y_1 for 1.

I accept Dr. Sen's criticism and believe that it calls for a change. Fortunately, one is already implicit in his analysis. This is to require that in ordering the set C_2 of consequences for person 2, person 1, with ordering relation R_1, agree on C_2 with R_2, i.e., with person 2's own ordering on C_2; a similar constraint is placed on R_2 with respect to R_1 on C_1. Formally, we need to add to Definition 4 (p. 158) the condition that on the subsets C_1 and C_2 of $C_1 \cup C_2$ the ordering relations R_1 and R_2 agree.

Dr. Sen's criticism leads to an emendation in the right direction, because it forces more structure on the concept of justice being set forth. I am, however, still far from satisfied with matters as they now stand. Far stronger structural principles are required to rule out other counter-intuitive examples, such as the one given at the end of the article.

The issues concerning probabilistic inference, its nature and its justification, have received extensive discussion in recent philosophical literature. I originally intended to relate what I had to say about these matters in Article 11 to what other people have said in the past couple of years. A wide-ranging and informative discussion of many of the central issues in inductive logic is to be found in the volume edited by Lakatos (1968), and the 1968 volume of *Philosophy of Science* contains useful papers by Hempel and others. When I attempted a preliminary review of the rapidly increasing literature, however, it soon became apparent that it would not be possible to deal with it briefly and in a way that was limited to trying to extend my own work to meet it. For example, a good part of the Lakatos volume is taken up by discussions by Salmon and others of rules of acceptance. In my judgment, the issues raised need to be analyzed in the context of modern statistical decision theory, not in terms of extending the theory of inference and the theory of explanation. In other words, to take the idea of acceptance seriously, we must proceed to an analysis of behavior and a theory of decisions.

The lottery paradox, which has been so much discussed in relation to rules of acceptance, seems to me an example of the sort of artificial puzzle generated by considering rules of acceptance apart from a theory of

decisions. In a way, perhaps, the St. Petersburg paradox of utility theory is similar in spirit to the lottery paradox, but in terms of the concepts of acceptance and certainty, there is a total lack of similarity. From still another standpoint, the law of large numbers, the central limit theorem, and other asymptotic results in probability theory are related both to probabilistic inference and rules of acceptance, because they describe what, under rather general assumptions, may be predicted to happen with near certainty in the long run. But to examine these relations is not possible here.

Ian Hacking's criticisms of Salmon and Reichenbach in the Lakatos volume are also pertinent. Hacking presses his remarks from the standpoint of de Finetti's ideas on the foundations of probability and induction. Hacking argues well for the Bayesian conception of learning by experience, especially in criticizing relative-frequency theories of induction. I share his skepticism of the ability of Bayesian ideas to deal with large parts of our cognitive experience. In another article (Suppes, 1966) published at the same time as Article 11, I tried to show in some detail why Bayesian ideas are not adequate to that part of learning by experience which requires the learning of a new concept. Some brief remarks about these matters are made at the end of Article 11. The learning-theoretic account of finite automata in the very last article of the present volume says as much as I can sharply formulate at the present time about the manner in which a learning mechanism might operate in learning a new concept.

In a detailed critique of Article 11 given in Levi's review (1967), I am accused, probably rightly, of adopting a radical psychologism toward the problems of induction. I am increasingly prepared to defend this general way of looking at both deductive and inductive logic. I suppose I feel the real test of a theory of concept formation or a theory of induction is its ability to generate the drawings for a machine, or more specifically, a computer that can form concepts and make inductions. Theories of this kind will not answer many sorts of Humean puzzles about predicting the future from knowledge of the past. Nevertheless, the contribution of such theories, once developed, to the philosophy of induction should be as substantial as have been the contributions of explicitly formulated set theories to the philosophy of mathematics.

6. THE ROLE OF SUBJECTIVE PROBABILITY AND UTILITY IN DECISION-MAKING*[1]

I. INTRODUCTION

Although many philosophers and statisticians believe that only an objectivistic theory of probability can have serious application in the sciences, there is a growing number of physicists and statisticians, if not philosophers, who advocate a subjective theory of probability. The increasing advocacy of subjective probability is surely due to the increasing awareness that the foundations of statistics are most properly constructed on the basis of a general theory of decision-making. In a given decision situation subjective elements seem to enter in three ways: (i) in the determination of a utility function (or its negative, a loss function) on the set of possible consequences, the actual consequence being determined by the true state of nature and the decision taken; (ii) in the determination of an *a priori* probability distribution on the states of nature; (iii) in the determination of other probability distributions in the decision situation.

These subjective factors may be illustrated by a simple example. A field general knows he is faced with opposing forces which consist of either (s_1) three infantry divisions and one armored division, or (s_2) two infantry divisions and two armored divisions. Thus the possible states of nature are s_1 and s_2. The possible consequences are a tactical victory (v), a stalemate (t), and a defeat (d). He subjectively estimates utilities as follows: $u(v)=3$, $u(t)=2$, $u(d)=-1$. On the basis of his intelligence he subjectively estimates the probability of s_1 as $\frac{1}{3}$, and of s_2 as $\frac{2}{3}$. Also in his view there are two major possible dispositions of his forces (f_1 and f_2). Using military experience and knowledge he now estimates the probability of victory, stalemate or defeat if he decides for disposition f_1 and s_1 is the true state of nature. Corresponding estimates are made for the pairs (f_1, s_2), (f_2, s_1) and (f_2, s_2). He then presumably decides on f_1 or f_2

* Reprinted from *The Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–55* **5** (1956), 61–73.

depending on which yields the greater expected utility with respect to his estimated *a priori* distribution on s_1 and s_2.

In connection with this example, it may properly be asked why probabilities and utilities play such a prominent role in the analysis of the general's problem. The most appropriate initial answer, it seems to me, is that we expect the general's decision to be rational in some definite sense. The probabilities are measures of degree of belief, and the utilities measures of value. To be rational he should try to maximize expected value or utility with respect to his beliefs concerning the facts of the situation. The crucial problem is: what basis is there for introducing numerical probabilities and utilities? Clearly methods of measurement and a theory which will properly sustain the methods are needed. Our intuitive experience is that at least in certain limited situations, like games of chance, such measurement is possible. The task for the decision theorist is to find unobjectionable postulates which will yield similar results in broader situations. It would be most unusual if any set of postulates which guaranteed formally satisfactory measures of probability and utility also was unequivocally intuitively rational. As we shall see in Section III, compromises of some sort must be reached.

Because of the many controversies concerning the nature of probability and its measurement, those most concerned with the general foundations of decision theory have abstained from using any unanalyzed numerical probabilities, and have insisted that quantitative probabilities be inferred from a pattern of qualitative decisions. A most elaborate and careful analysis of these problems is to be found in L. J. Savage's recent book, *Foundations of Statistics* (1954). The present paper gives an axiomatization of decision theory which is similar to Savage's. The summary result concerning the role of subjective probability and utility is the same: one decision is preferred to a second if and only if the expected value of the first is greater than that of the second.

The theory presented here differs from Savage's in two important respects: (i) the number of states of nature is arbitrary rather than infinite; (ii) a fifty-fifty randomization of two pure decisions is permitted; this does not presuppose a quantitative theory of probability. More detailed differences are discussed in Section III. Since the present scheme is offered as an alternative to Savage's it is perhaps worth emphasizing that the intuitive ideas at its basis were developed in collaboration with

Professor Donald Davidson in the process of designing experiments to measure subjective probability and utility (1955, 1956). I suspect that experimental application of Savage's approach may be more difficult. It should also be mentioned that the approach developed here goes back to the early important, unduly neglected work of Ramsey (1931).

The proof of adequacy of the axioms in Section IV depends on previous work by Mrs. Muriel Winet and me (1955)[2†], and unpublished results by Professor Herman Rubin (1954); it is unfortunate that Rubin's important results are still unpublished. His work differs from the present in that he assumes a quantitative theory of probability.

Finally it should be remarked that the theory developed in the present paper is presumed susceptible of either prescriptive or descriptive use.

II. PRIMITIVE AND DEFINED NOTIONS

The four primitive notions on which our axiomatic analysis of decision-making is based are very similar to the four used by Savage (1954). Our first primitive is a set S of states of nature; the second, a set C of consequences; and the third, a set D of decision functions mapping S into C. Savage's first three primitive notions are identical. His fourth primitive is a binary relation of preference on D. In contradistinction, our fourth primitive $\geqslant$ is a binary relation of preference on the *Cartesian product* $D \times D$. ($D \times D$ is the set of all ordered couples (f, g) such that f and g are in D.) This apparently slight technical difference reflects the introduction of a restricted notion of randomization which does not require a quantitative concept of probability. Thus if f, g, f' and g' are in D, the intended interpretation of $(fg) \geqslant (f', g')$ is that the decision-maker (weakly) prefers a half chance on f and a half chance on g to the mixed decision consisting of a half chance on f' and a half chance on g'. For application of the apparatus developed here it must be possible to find a chance event which is independent of the state of nature and which has a subjective probability of $\frac{1}{2}$ for the decision-maker.[3] In most applications of decision theory it should be relatively easy to find such a chance event, since we are usually dealing with what Savage calls small-world situations, and not the fate of the whole universe.

To illustrate the intended interpretation of our primitive notions we may consider the following example. A certain independent distributor

of bread must place his order for a given day by ten o'clock of the preceding evening. His sales to independent grocers are affected by whether or not it is raining at the time of delivery, for if it is raining, the grocers tend to buy less on the reasonably well-documented evidence that they have fewer customers. On a rainy day the maximum the distributor can sell is 700 loaves; on such a day he makes less money if he has ordered more than 700 loaves. On the other hand, when the weather is fair, he can sell 900 loaves. If the simplifying assumption is made that the consequences to him of a given decision with a given state of nature (rainy or not) may be summarized simply in terms of his net profits, the situation facing him is represented in Table I.

TABLE I

	d_1–buy 700 loaves	d_2–buy 800 loaves	d_3–buy 900 loaves
s_1–rain	\$21.00	\$19.00	\$17.00
s_2–no rain	21.00	24.00	26.50

The distributor's problem is to make a decision. Decision d_2 is a kind of hedge. We also permit him the hedge of randomizing fifty-fifty between two pure decisions. He may own a coin which he believes is fair, and he does not believe that flipping this coin has an effect on the weather. Thus he may choose the mixture (d_1, d_3) over d_2. On a particular morning he might prefer the possible course of action open to him as follows:

$$(1) \qquad d_1 > (d_1, d_2) > (d_1, d_3) > d_2 > (d_2, d_3) > d_3 .$$

The use of the relation $>$ in this example is made precise by two definitions. Since the mixture (f, f) in the intended interpretation just means decision (or action) f, it is natural to extend the field of $\geqslant$ to D.

DEFINITION 1: *$(f, g) \geqslant h$ if and only if $(f, g) \geqslant (h, h)$; $h \geqslant (f, g)$ if and only if $(h, h) \geqslant (f, g)$; and $h \geqslant g$ if and only if $(h, h) \geqslant (g, g)$.*

For α and β either mixtures or pure decisions we now define the relation $>$ of strong preference.

DEFINITION 2: *$\alpha > \beta$ if and only if $\alpha \geqslant \beta$ and not $\beta \geqslant \alpha$.*

For later work we also need the definition of equivalence in preference (that is, indifference).

DEFINITION 3: *$\alpha \sim \beta$ if and only if $\alpha \geqslant \beta$ and $\beta \geqslant \alpha$.*

For the statement of our axioms on decision-making two further definitions are needed. The first is the definition of a notion we need for the statement of the Archimedean axiom (A.7).

DEFINITION 4: *$(f, g)\ L\ (f', g')$ if and only if $f \sim f'$ and $(f, g) \sim g'$.*

The Archimedean axiom makes use of *powers* of L. We have that $(f, g)\ L^2\ (f', g')$ if and only if there exist decisions f'' and g'' such that $(f, g)\ L\ (f'', g'')$ and $(f'', g'')\ L\ (f', g')$, which situation is represented in Figure 1. (Note that f, f' and f'' all occupy the same position.)

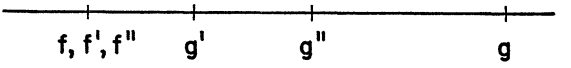


Fig. 1.

The nth power of L is defined recursively:

(1) $(f, g)\ L^1\ (f', g')$ if and only if $(f, g)\ L\ (f', g')$;

(2) $(f, g)\ L^n\ (f', g')$ if and only if there are elements f'' and g'' in D such that $(f, g)\ L^{n-1}\ (f'', g'')$ and $(f'', g'')\ L\ (f', g')$. The numerical interpretation of the relationship $(f, g)\ L^n\ (f', g')$ is that $f = f'$ and

$$\frac{2^n - 1}{2^n} f + \frac{1}{2^n} g = g'.$$

Finally, we need the notion of a *constant* decision function, that is, a function which yields the same consequence independent of the state of nature.

DEFINITION 5: *If $x \in C$ then x^* is the function mapping S into C such that for every $s \in S$, $x^*(s) = x$.*

As we shall see, the constant decisions play an all too important role in the theory developed in this paper.

III. AXIOMS

Using the primitive and defined notions just considered we now state our axioms for what we shall call *rational subjective choice structures.*

A system $\langle S, C, D \geqslant \rangle$ is a RATIONAL SUBJECTIVE CHOICE STRUCTURE *if and only if the following axioms* 1–11 *are satisfied for every f, g, h, f', g',*

h', f'' and g'' in D:

A1. $(f, g) \geqslant (f', g')$ *or* $(f', g') \geqslant (f, g)$;

A2. If $(f, g) \geqslant (f', g')$ *and* $(f', g') \geqslant (f'', g'')$ *then* $(f, g) \geqslant (f'', g'')$;

A3. $(f, g) \sim (g, f)$;

A4. $f \geqslant g$ *if and only if* $(f, h) \geqslant (g, h)$;

A5. If $(f, g) \geqslant (f', g')$ and $(h, g') \geqslant (h', g)$ *then* $(f, h) \geqslant (f', h')$;

A6. If $(f, g) > (f', g')$ *and* $g > g'$ *then there is an h in D such that* $g > h$ *and* $h > g'$ *and* $(f, g) \geqslant (f', h)$;

A7. If $f > g$ *and* $f' > g'$, *then there is an h in D and a natural number n such that* $(f, g)\ L^n\ (f, h)$ *and* $(f', h) \geqslant (f, g')$;

A8. For every x in C, $x^* \in D$;

A9. If for every s in S, $(f(s)^*, g(s)^*) \geqslant (f'(s)^*, g'(s)^*)$, *then* $(f, g) \geqslant (f', g')$;

A10. There is an h in D such that for every s in S, $h(s)^* \geqslant f(s)^*$ *and* $h(s)^* \geqslant g(s)^*$;

A11. There is an h in D such that for every s in S, $(f(s)^*, g(s)^*) \sim h(s)^*$.

The interpretation of the first two axioms is clear: they require a simple ordering of decisions. The third axiom guarantees that our special chance event independent of the state of nature has subjective probability $\frac{1}{2}$. To see this, let $f > g$, and let E^* be our special chance event. The interpretation of (f, g) is that decision f is taken if E^* occurs and g if $\tilde{E}^*$ occurs (that is, if E^* does not occur). If the subjective probability of E^* (in symbols: $s(E^*)$) is greater than that of $\tilde{E}^*$, (f, g) will be preferred to (g, f). On the other hand, if $s(E^*) < s(\tilde{E}^*)$, then (g, f) will be preferred to (f, g). Hence, A.3 corresponds to saying that $s(E^*) = s(\tilde{E}^*) = \frac{1}{2}$. (For further discussion of this, see Davidson and Suppes, 1956.)

Axiom A.4 states an obvious substitution property. It is a special case ($\alpha = \frac{1}{2}$) of an axiom introduced by Friedman and Savage (1952, p. 468, axiom P3). It also is essentially a special case of Samuelson's strong independence axiom (1952). A kind of domination property is expressed by A.5. If the mixture (f, g) is at least as desirable as the mixture (f', g'), and h is sufficiently preferred over h' to reverse this preference in the sense that (h, g') is weakly preferred to (h', g), then it is reasonable to expect that (f, h) is weakly preferred to (f', h'). The content of this axiom is made clearer by considering particular cases among the possible orderings of the decisions. An example which brings out the implications of the axiom is given by the supposition that we have the following ordering:

$f' \geqslant f \geqslant g \geqslant g'$. Now we must then have $h \geqslant h'$ since $(h, g') \geqslant (h', g)$; furthermore, the latter implies that the difference between h and h' is greater than between f' and f, since when h is coupled with the least desirable decision g', the mixture (h, g') is preferred to (h', g), but in the case of f', the mixing with g' leads to (f, g) being preferred. Hence, we expect to find that (f, h) is weakly preferred to (f', h'), which is what the axiom requires.

Axiom A.6 I regard as a blemish which should be eliminated or changed in form. It says nothing essentially new about the structure of any model of our axioms; just that if (f, g) is preferred to (f', g'), then we may find a decision h slightly better than g' such that we will have (f, g) preferred to the new mixture (f, h). Axiom A.7 is an Archimedean axiom of the sort necessary to get measurability. Its existence requirements are not unreasonable in view of the plenitude of decisions guaranteed by A.10 and A.11. The meaning of A.7 is very simple. No matter how great the interval between f and g, the interval may be subdivided sufficiently to find an h closer to f than g' is to f'. The axiom could be weakened by adding to the hypothesis the condition that $(f, g') \geqslant (f', g)$.

Axiom A.8 requires that all constant decisions, that is, decisions whose consequences are independent of the state of nature, be in D. The inclusion of such constant decisions, or of something essentially as strong, is necessary to obtain the summary result we want: f is preferred to g if and only if the expected value of f with respect to a utility function on consequences and an *a priori* distribution on states of nature is greater than the corresponding expected value of g. The inclusion of these constant decisions is not peculiar to the theory of decision-making developed here, but is also essential to Rubin's (1954) and Savage's (1954) theories.[4] The difficulties surrounding the inclusion of these decisions may be illustrated by considering one of Savage's colorful examples (see Savage, 1954, p. 14). We have before us an egg. One of two states of nature obtains: the egg is good (s_1) or the egg is rotten (s_2). We are making an omelet and five good eggs have already been broken into the bowl. We may take one of three actions: break the egg in the bowl (f), break the egg in a saucer and inspect it (g), simply throw the egg away (h). The various consequences are easy to describe: $f(s_1)$=six-egg omelet, $g(s_1)$=six-egg omelet and saucer to wash, etc. But now suppose we add the constant decisions. How are we to think about the decision which

guarantees us a six-egg omelet? If the true state of nature is s_2, it is not clear that we are considering an action which makes any kind of sense. Certainly we are in no position to push the ultrabehavioristic interpretation of decision-making favored by Savage when we consider the constant decisions. I can, for instance, imagine no behavioristic evidence which would persuade me that an individual in the situation just described had chosen the constant decision guaranteeing a six-egg omelet. As far as I can see, about the most reasonable way to analyze a preference involving a constant decision such as the above one is to regard it as a non-behavioristic subjective evaluation of consequences. Axioms A.8–A.11 have the effect intuitively of requiring such direct evaluations of consequences.

Axiom A.9 corresponds closely to Savage's seventh postulate and to Rubin's sixth axiom (1954). If for every state of nature the consequences of the mixture of decisions f and g are preferred to the consequences of the mixture of f' and g', then the mixture of f and g should be preferred to that of f' and g'. As Savage remarks, the kind of sure-thing principle expressed by this axiom is one of the most acceptable postulates of rational behavior. Axiom A.10 asserts that given any two decisions there is a third at least as good as either of the two with respect to every state of nature. This axiom is weaker than the assumption that the set of consequences of any decision f has an upper bound, that is, there is an x in C such that for every s in S, $x^* \geqslant f(s)^*$. It is possible that the main theorem of Section IV can be proved without this axiom, but I have not succeeded in finding such a proof.

Axiom A.11 should probably be regarded as the strongest axiom of the group. Given any two decisions f and g, A.11 asserts there is another decision h with the property that for each state of nature the consequence of h is halfway between the consequence of f and the consequence of g. This axiom may be regarded as a very strong form of Marschak's continuity axiom (1950). His axiom is that if $f > g$ and $g > h$ then there is a numerical probability α such that the mixture of f and h with probability α and $1-\alpha$ respectively is equivalent to g. The significance of A.11 is discussed in more detail below.

Now that the analysis of individual axioms is complete, some general remarks are pertinent. Compared to Savage's axiomatization (1954), we may say of the present theory that there are more axioms but perhaps less

complicated definitions. A more important kind of comparison between Savage's and the present analysis is the rather radical difference in what I like to call the *structure* axioms (as opposed to the *rationality* axioms). By and large, a structure axiom is an existential assertion.[5] Axiom A.11 is the main structure axiom in the present axiomatization. If we consider the situation facing the independent distributor of bread, which was discussed in the last section, it is clear that A.11 is not satisfied. In fact, it is easy to show that if there are two decisions, one of which is strictly preferred to the other, then A.11 and certain of the other axioms imply that there is an infinity of decisions. However, I for one am reluctant to call the distributor irrational because an insufficient number of decisions is available to him. I prefer to say that the situation the distributor is in does not permit the structure axioms to be satisfied, and hence the present theory is inapplicable; we cannot use it to decide if the distributor is regularly choosing an action or decision solely in terms of its expected value. In a given axiomatic analysis of decision-making it is not always easy or even possible clearly to separate the axioms into the two categories of rationality axioms and structure axioms. Of the eleven axioms used in this paper, I would say that A.1–A.5 and A.9 are "pure" rationality axioms which should be satisfied by any rational, reflective man in a decision-making situation. On the other hand, A.8, A.10 and A.11 are "pure" structure axioms which have little directly to do with the intuitive notion of rationality. They are to be considered as axioms which impose limitations on the kind of situations to which our analysis may be applied. Axiom A.6 is a technical structure axiom which tells us little intuitively about restrictions on applicability of the theory. Without A.11, the Archimedean axiom, A.7, would need to be considered a structure axiom, but in the presence of A.11, I regard it as a rationality axiom.

Of Savage's seven postulates, two are structure axioms (P5 and P6), and the rest are rationality axioms. His P5 excludes the trivial case where all consequences are equivalent in utility and thus every decision is equivalent to every other. Postulate P6 is his powerful structure axiom corresponding to my A.11. Essentially his P6 says that if event B is less probable than event C (B and C are subsets of S, the set of states of nature), then there is a partition of S such that the union of each element of the partition with B is less probable than C. As Savage remarks, this postulate is slightly stronger than the axiom of de Finetti and Koopman which

requires the existence of a partition of S into arbitrarily many events which are equivalent in probability. Thus the consequence of his P6 is that there must be an infinity of states of nature, and as a consequence, an infinity of decisions; whereas the consequence of A.11 is that there must be an infinity of decisions, with the number of states of nature wholly arbitrary. Such infinite sets, either of decisions or states of nature, can be eliminated by various kinds of special structure axioms. Davidson and I (1956) eliminated them by requiring that all consequences be equally spaced in utility – an assumption which has proved manageable in some controlled experiments on decision-making at Stanford (Davidson *et al.*, 1955), but is not realistic in general.

Savage defends his P6 by holding it is workable if there is a coin which the decision-maker believes is fair for any finite sequence of flips (1954, p. 33). However, if the decision-maker does not believe the flipping of the coin affects what is ordinarily thought of as the state of nature, such as raining or not raining in the case of the bread distributor, then it seems to me that it is misleading to construct the states of nature around the fair coin. Once *repeated* flips of a fair coin are admitted, we can extend the single act of randomization admitted in the interpretation of the axiomatization given here, and directly introduce all numerical probabilities of the form $k/2^n$. With this apparatus available we can give an axiomatization very similar to Rubin's (1954) and drop any strong structure axioms on the number of states of nature or the number of decisions.

To illustrate further the nature of the structure axiom A.11, and at the same time to argue by way of example that it does not make our theory impossible of application, I would like to modify one of Savage's finite examples (1954, pp. 107–108) which does not, even as modified, satisfy his P6. A man is considering buying some grapes in a grocery store. The grapes are in one of three conditions (the three states of nature): green, ripe, or rotten. The man may decide to buy any rational number of pounds between 0 and 3. If, for example, the state of nature is that the grapes are rotten and he makes the decision to buy two pounds, then the immediate consequence is possession of two pounds of rotten grapes and the loss of a certain small amount of capital. If the man is at all intuitively rational in his preferences concerning the amount of grapes to buy, it will not be hard for him to satisfy A.1–A.11 – provided, of course, that he has at hand some simple random mechanism, such as a coin he believes to be

fair for single tosses (he need not believe that any finite *sequence* of outcomes is as likely as any other). This example is discussed further in Section V.

By way of summary my own feeling is that Savage's postulates are perhaps esthetically more appealing than mine, but this fact is balanced by two other considerations: my axioms do not require an infinite number of states of nature, and their intuitive basis derives from ideas which have proved experimentally workable.

IV. ADEQUACY OF AXIOMS

We now turn to the proof that our axioms for decision-making are adequate in the sense that decision f is weakly preferred to decision g if and only if the expected value of f is at least as great as the expected value of g. The actual result is not quite this strong. As might be expected, the theorem holds only for bounded decisions (precisely what is meant by a bounded decision is made clear in the statement of the theorem). On the basis of A.1–A.11 uniqueness of the *a priori* distribution on the states of nature cannot be proved, since the constant decisions alone constitute a realization of the axioms. If S is assumed finite, various conditions which guarantee uniqueness are easy to give. In stating the theorem, we use the notation: $U \circ f$ for the *composition* of the functions U and f.

THEOREM: *If* $\langle S, C, D, \geqslant \rangle$ *is a rational subjective choice structure, then there exists a real-value function* ϕ *on* D *such that*

(i) *for every* f, g, f' *and* g' *in* D $(f, g) \geqslant (f', g')$ *if and only if* $\phi(f) + \phi(g) \geqslant \phi(f') + \phi(g')$,

(ii) ϕ *is unique up to a linear transformation, and*

(iii) *if* U *is the function defined on* C *such that for every* x *in* C

$$U(x) = \phi(x^*), \tag{1}$$

then there exists a finitely additive probability measure P *on* S *such that for every* f *in* D *if* $U \circ f$ *is bounded, then*

$$\phi(f) = \int_S (U \circ f)(s)\, dP(s). \tag{2}$$

Proof: The proof of (i) and (ii) follows rather easily from some previous

results obtained by Mrs. Muriel Winet and me. Using a notion R of utility differences and a notion Q of preference, we established (1955)[6†] that, on the basis of axioms similar to A.1–A.7 and A.11 of this paper, there exists a real-valued function ψ unique up to a linear transformation such that

(3) $f\ Q\ g$ if and only if $\psi(f) \geqslant \psi(g)$,
f, gRf', g' if and only if $|\psi(f) - \psi(g)| \geqslant |\psi(f') - \psi(g')|$.

If we introduce the two defining equivalences[7]

(4) $f\ Q\ g$ if and only if $(f, f) \geqslant (g, g)$;

(5) f, gRf', g' if and only if either (i) $f \geqslant g, f' \geqslant g'$ and $(f, g') \geqslant (f', g)$, or (ii) $g \geqslant f$, $f' \geqslant g'$ and $(g, g') \geqslant (f, f')$, or (iii) $f \geqslant g, g' \geqslant f'$ and $(f, f') \geqslant (g, g')$, or (iv) $g \geqslant f, g' \geqslant f'$ and $(g, f') \geqslant (f, g')$,

then on the basis of A.1–A.7, A.9 and A.11 we may prove the axioms of Suppes and Winet (1955)[8] on Q and R as theorems, as well as the equivalence

(6) $(f, g') \geqslant (f', g)$ if and only if either (i) $f \geqslant g, f' \geqslant g'$ and $f, g\ R\ f', g'$, or (ii) $f \geqslant g$ and $g' \geqslant f'$, or (iii) $g \geqslant f, g' \geqslant f'$ and $f', g'\ R\ f, g$.

Parts (i) and (ii) of our theorem then follow immediately from the main theorem in Suppes and Winet (1955).

The proof of (iii), concerning the existence of an *a priori* distribution on S, essentially uses Rubin's results (1954). However, certain extensions of D are required in order to apply his main theorem.

By means of the utility function U on the set of consequences C, as defined in the hypothesis of (iii), we define the set F of all numerical income functions

(7) $F = \{\rho$: there exists f in D such that $\rho = U \circ f\}$,

and we define the functional η on F

(8) $\eta(U \circ f) = \phi(f)$.

We observe first that if $\rho, \sigma \in F$, then

$$\frac{1}{2}\rho + \frac{1}{2}\sigma \in F, \tag{9}$$

for let $\rho = U \circ f$ and $\sigma = U \circ g$, then by (i) and (ii) of our theorem, A.9, and A.11, there exists an h in D such that for every s in S

$$\tfrac{1}{2}(U \circ f)(s) + \tfrac{1}{2}(U \circ g)(s) = (U \circ h)(s). \tag{10}$$

Hence,

$$\frac{1}{2}\rho + \frac{1}{2}\sigma = U \circ h, \tag{11}$$

and $U \circ h$ is in F. Also, since

$$\eta(U \circ h) = \phi(h) = \tfrac{1}{2}\phi(f) + \tfrac{1}{2}\phi(g) = \tfrac{1}{2}\eta(U \circ f) + \tfrac{1}{2}\eta(U \circ g), \tag{12}$$

we have

$$\eta\left(\frac{1}{2}\rho + \frac{1}{2}\sigma\right) = \frac{1}{2}\eta(\rho) + \frac{1}{2}\eta(\sigma). \tag{13}$$

From (9) and (13) it easily follows that if $\rho, \sigma \in F$ and k and n are positive integers such that $k \leqslant 2^n$, then

$$\frac{k}{2^n}\rho + \left(1 - \frac{k}{2^n}\right)\sigma \in F, \tag{14}$$

and

$$\eta\left[\frac{k}{2^n}\rho + \left(1 - \frac{k}{2^n}\right)\sigma\right] = \frac{k}{2^n}\eta(\rho) + \left(1 - \frac{k}{2^n}\right)\eta(\sigma). \tag{15}$$

We now extend F by the following definition: $\rho \in \bar{F}$ if and only if there is a finite sequence $\langle a_1, \ldots, a_n \rangle$ of real numbers and a finite sequence $\langle \rho_1, \ldots, \rho_n \rangle$ of elements of F such that

$$\rho = \sum_n a_i \rho_i. \tag{16}$$

(It is clear from (16) that $\bar{F}$ is a linear space.)

In order to extend η in a well-defined manner to $\bar{F}$, we need to prove that if

$$\sum_n a_i \rho_i = \sum_m b_j \sigma_j \tag{17}$$

then

$$(18) \qquad \sum_n a_i \eta(\rho_i) = \sum_m b_j \eta(\sigma_j).$$

Clearly without loss of generality we may assume

$$(19) \qquad a_i, b_j > 0 \quad \text{for} \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m.$$

We shall first establish (18) under the restriction that

$$(20) \qquad \sum_n a_i = \sum_m b_j.$$

If a_i and b_j are rational numbers of the form $k/2^n$ with $k \leqslant 2^n$, then (18) follows from (15) by a straightforward inductive argument (which we omit), provided

$$(21) \qquad \sum_n a_i = \sum_m b_j = 1.$$

But the requirement of (21) is easily weakened to

$$(22) \qquad \sum_n a_i = \sum_m b_j < 1,$$

for we may add $c\rho$ with $c = 1 - \sum_n a_i$ to both sides of (17), and then (21) will be satisfied. Furthermore, (22) is readily extended to arbitrary positive rationals, since two finite sequences of positive rationals can be reduced to (22) by multiplying through and dividing by a sufficiently high power of 2.

We are now ready to consider the case where the a_i's and b_j's are arbitrary positive real numbers. There are rational numbers r_i and s_j such that

$$(23) \qquad r_i < a_i \quad \text{and} \quad s_j > b_j.$$

It is an immediate consequence of A.10 that there is a τ in F such that

$$(24) \qquad \tau \geqslant \rho_i \quad \text{and} \quad \tau \geqslant \sigma_j.$$

From (23) and (24) we have, by a regrouping of coefficients

$$(25) \qquad \sum_n r_i \rho_i + [\sum_n (a_i - r_i) + \sum_m (s_j - b_j)]\, r \geqslant \sum_m s_j \sigma_j.$$

Since the coefficient of τ is rational, we obtain by our previous results

$$\sum_n r_i\eta(\rho_i) + \lambda\eta(\tau) \geqslant \sum_m s_j\eta(\sigma_j), \tag{26}$$

where

$$\lambda = \sum_n (a_i - r_i) + \sum_m (s_j - \beta_j). \tag{27}$$

By suitable choice of the r_i's and s_j's, we may make λ arbitrarily small, and we thus infer from (23) and (26)

$$\sum_n a_i\eta(\rho_i) \geqslant \sum_m b_j\eta(\sigma_j). \tag{28}$$

By an exactly similar argument, we get

$$\sum_m b_j\eta(\sigma_j) \geqslant \sum_n a_i\eta(\rho_i). \tag{29}$$

To establish (18) in full generality it remains only to consider the case where

$$\sum_n a_i \neq \sum_m b_j. \tag{30}$$

Suppose, for definiteness, that

$$\sum_n a_i > \sum_m b_j. \tag{31}$$

There are elements x and y in C such that $U(x) > U(y)$ (if there are no two such elements, the proof of the whole theorem is trivial). Furthermore, in view of A.11, we may choose x and y such that $U(x) > 0$ and $U(y) > 0$, or $U(x) < 0$ and $U(y) < 0$. Let $\mu = U \circ x^*$ and $\nu = U \circ y^*$. Then μ and ν are in F, and there are nonnegative numbers a_0 and b_0 such that

$$a_0 + \sum_n a_i = b_0 + \sum_m b_j \tag{32}$$

and

$$a_0\mu = b_0\nu. \tag{33}$$

Then by our previous result under the restriction (20), we have

$$a_0\eta(\mu) + \sum_n a_i\eta(\rho_i) = b_0\eta(\nu) + \sum_m b_j\eta(\sigma_j), \tag{34}$$

but from (33), (8) and the definition of U

$$a_0\eta(\mu) = b_0\eta(\nu), \tag{35}$$

and thus

$$\sum_n a_i\eta(\rho_i) = \sum_m b_j\eta(\sigma_j), \tag{36}$$

which establishes (18) in full generality.

On the basis of (18) we extend η to $\bar{F}$. The argument from (30) on has closely followed Rubin's proof (1954). His proof may now be used to complete the proof of (iii). We sketch the main steps. Clearly η is a linear functional on $\bar{F}$, and it is easily shown that η is nonnegative, and hence, that $\eta(\rho)$ is between $\inf_{s\in S}\rho(s)$ and $\sup_{s\in S}\rho(s)$. Let G be the space of all functions on S bounded by elements of $\bar{F}$. Then by the Hahn-Banach theorem (Banach, 1932, pp. 27–28) η can be extended to G. Finally, it can be shown (Rubin, 1949a, b) that such a linear functional on G is, for bounded functions in F, their expected value with respect to an *a priori* distribution on S which is in general finitely additive. (A result closely related to the existence of such a distribution is established in Theorem 2.3, Yosida and Hewitt, 1952.)

V. CRITICAL REMARKS

The theory of decision developed in the previous sections is no doubt defective in a number of ways, some of which I am well aware of. In this final section I briefly examine what I consider to be its gravest weakness, at least for normative applications. It is laudable to wish to base a theory of decision on behaviorally observable choices, but the decision-maker is interested in something more. He wants advice on how to choose among alternative courses of action. He wants to have at hand a theory which tells him how to use initial information. The result of the analysis in this paper and in Savage's book is that if certain structure axioms are satisfied, any rational man acts as if he had an *a priori* distribution on the states of nature. But what the rational man wants is a method for selecting that *a priori* distribution which *best* uses his *a priori* information. The present theory or Savage's offers little help on this point. The importance of this problem is testified to by the over-all situation in statistical decision

theory: we have clear ideas of optimality only when given an *a priori* distribution on the states of nature. Bayesian principles of choice seem naturally to dominate the scene. (For some penetrating reasons, see Blackwell and Girshick, 1954, Chap. 4.)

In recent years a serious attempt has been made by philosophical logicians to develop a theory of confirmation which is closely related to the problem under discussion. The theory of confirmation is concerned with precisely characterizing the degree to which a given hypothesis is supported by given evidence. The confirmation function which is usually introduced is very similar in its formal properties to the standard notion of conditional probability. Perhaps because the theory of confirmation has usually been stated in logical or linguistic terms, its connections with decision theory have not been made as clear as they could. Thus viewed, the purpose of confirmation theory is to develop methods for codifying prior information to yield an *a priori* distribution on the states of nature. The available evidence is our prior information, and a hypothesis corresponds to asserting that a given state of nature is the true one.

For concreteness we may consider the grape example of Section III. In Savage's discussion of this example (1954, p. 108) he assigns subjective probabilities to the three states of nature, and then goes on to consider what action the decision-maker should take after observing a sample of one grape. But the point at issue here is: given certain prior information is one *a priori* distribution as reasonable as any other? As far as I can see, there is nothing in my or Savage's axioms which prevents an affirmative answer to this question. Yet if a man had bought grapes at this store on fifteen previous occasions and had always got green or ripe, but never rotten grapes, and if he had no other information prior to sampling the grapes, I for one would regard as unreasonable an *a priori* distribution which assigned a probability of $\frac{2}{3}$ to the rotten state. Unfortunately, though I am prepared to reject this one distribution as unreasonable, I am not prepared to say what I think is optimal.

The most thoroughgoing analysis of confirmation theory has been made by Carnap (1950), but his chosen confirmation function c^* is beset with many technical difficulties which give rise to counterintuitive examples (see, for example, Kemeny, 1951, 1953; Rubin and Suppes, 1955). Here I am not concerned to scrutinize the current problems of confirmation theory, but merely to argue for the relevance of the theory to decision

theory.[9] An adequate confirmation theory would not discredit the kind of axiomatization of decision-making given in this paper; it would not disturb the central role of subjective probability and utility.[10] It would stand to the theory of this paper more as statistical mechanics stands to macroscopic thermodynamics: a decision theory which included a confirmation function would have the axioms of the present paper (or of a similar theory such as Savage's) forthcoming as theorems. Such an enlarged decision theory would remain subjective, but an important element of counterintuitive arbitrariness would have been eliminated.

In conclusion, I should like to acknowledge my indebtedness to Herman Rubin for a number of helpful suggestions, as well as to Donald Davidson, Robert McNaughton and Jean Rubin for their useful comments.

NOTES

[1] I am indebted to Herman Rubin for a number of helpful suggestions.

[2†] Article 8 in this volume.

[3] The term 'mixed decision' is used here in the very restricted sense of referring to gambles involving just this special chance event independent of the state of nature; formally such gambles are the elements of $D \times D$.

[4] An analogue of our A8 is not included among Savage's seven axioms unless his set F of acts (corresponding to our set D of decisions) is meant to be the set of *all* functions mapping S into C, which is of course a stronger assumption than A8. In any case it is essential to his formal developments to have such decisions at hand (see Savage, 1954, from p. 25 on).

[5] This is certainly not always the case. The strong structure axiom in Davidson and Suppes (1956), which asserts that consequences are equally spared in utility, is not existential in character.

[6†] Article 8 in this volume.

[7] In Suppes and Winet (1955; Article 8 in this volume) the inequalities of (3) are actually reversed, but trivial changes in the axioms given there yield (3) as a consequence.

[8†] Article 8 in this volume.

[9] A central problem in confirmation theory is what *a priori* distribution to choose when there is no information whatsoever. Chernoff (1954) has shown that if certain reasonable postulates are accepted and if the number of states of nature is finite, then the distribution to choose is that one which makes each state equally probable.

[10] This remark is controversial. In the opinion of many competent investigators an adequate confirmation theory would dispense with any need for subjective probability. I cannot here state my reasons for disagreeing with this view.

7. THE PHILOSOPHICAL RELEVANCE OF DECISION THEORY*

I. INTRODUCTION

There is, I am sure, a sense in which any developed scientific theory has philosophical significance. It is equally clear that some scientific theories are of considerably more philosophical importance than others. For philosophy, quantum mechanics is more important than hydrodynamics, learning theory than social psychology, the theory of sets than topology, and so on. It is the primary point of the present paper to discuss the philosophical relevance or importance of decision theory, a theory I classify as a new branch of mathematical statistics and economics, with certain ramifications in psychology. I hope to be able to show that in its own way decision theory has the kind of primary relevance for philosophy that we associate with quantum mechanics or the theory of sets.

To begin with, we may characterize the fundamental problem of decision theory in the following way. A person, or group of persons, is faced with several alternative courses of action. In most cases the decision-maker will have only incomplete information about the true state of affairs and the consequences of each possible act. The problem is to choose an act that is optimal relative to the information available and according to some definite criteria of optimality.

In a very natural way, the most important branches of decision theory may be characterized by a 2×2 table as illustrated in Table I. The left column is for the category of individual decisions and the right column for the category of group decisions. The first row is for the category of normative theory and the second for the category of descriptive theory.

Rather than comment on the philosophical relevance of decision theory in general, I shall attempt to indicate what I think are the most interesting ramifications for philosophy of each of the quadrants shown. The emphasis will be on normative theory.

* Reprinted from *The Journal of Philosophy* **58** (1961) 605–614.

TABLE I

	Individual decisions	*Group decisions*
Normative theory	Classical economics Statistical decision theory Moral philosophy	Game theory Welfare economics Political theory
Descriptive theory	Experimental decision studies Learning theory Survey studies of voting behavior	Social psychology Political science

II. INDIVIDUAL NORMATIVE THEORY

A common problem besetting both the theory of induction and moral philosophy is that of giving an adequate account of the concept of rationality. The normative theory of individual decision making has been concerned to explicate the notion of rationality in what is, in some respects, a very thorough fashion. It would be rather absurd in this general paper to attempt to survey even a small fraction of the many substantial results, primarily of a technical nature, that have been achieved in the last two decades. There is, however, a central kind of difficulty that has arisen and that I think is of great philosophical importance.

Let me begin in an indirect fashion by a comparison with the situation in the foundations of mathematics. As work on the arithmetization of analysis and the development of the theory of sets progressed in the 19th century, it seemed possible, at least for a short period, that the whole of mathematics could be derived from three simple postulates: the principle of abstraction, that is, that for any property there exists a set of elements having this property; the principle of extensionality, that is, that two sets are identical just when they have the same members; and the axiom of choice. Mathematical work in the 19th century indicated that the theory of sets was a natural framework within which to construct the rest of mathematics.[1] The intuitive and simple reasonableness of these three principles (with the possible exception of the axiom of choice) seemed overwhelming.

The discovery of paradoxes derivable from the first two assumptions by Burali-Forti, Russell, and others shook the foundations of mathematics

to such an extent that a full recovery has not yet been made. But the discovery of the paradoxes has had an effect that has not been entirely negative. The many attempts, ranging from intuitionism to formalism, to put the foundations of mathematics on a secure basis have brought our understanding of the nature of mathematics to a new level of sophistication. Related negative results like the incompleteness theorems of Gödel have shown that what seemed to be immediate and obvious intuitions about even so restricted a branch of mathematics as elementary number theory are unreliable.

Recent work in decision theory has shown in similar fashion that there is no simple coherent set of principles capable of precise statement that corresponds to naive ideas of rationality. Just as research in this century in the foundations of mathematics has shown that we do not yet know exactly what mathematics is, so the work in decision theory shows that we do not yet understand what we mean by rationality. I mean by this not merely that we have no adequate general definition of rationality, but that, even for highly restricted circumstances, it turns out to be extremely difficult to characterize what we intuitively would want to mean by a rational choice among alternative courses of action. To focus on a collection of special situations and to attempt to characterize a rational strategy of choice for them is very similar and, in fact, closely related to an attempt to solve particular problems of induction without necessarily resolving "the" problem of induction.

The formidable problems besetting the rational decision-maker may be illustrated by considering the difficulty of formulating an adequate principle of choice for finite games against "nature". Such games are special cases of the description given above of the general decision situation, although a wide variety of decision situations may be mathematically represented as such finite games. The game may be represented by a matrix in which a player must choose a row and a column is chosen by nature. The entries in the matrix represent the payoff to the decision-maker, when he chooses a row and nature chooses a column, but the metaphorical talk about nature should not mislead anyone. This is merely a way of referring to a situation in which a course of action must be taken against an opponent or in an objective situation about which there is no information. The restriction to no information is not so severe as it may seem, for a situation in which partial information is incorporated

may be redefined to yield a game with no information – a familiar practice in the mathematical theory of decisions. The happy situation when the entries in some one row are for every column better than those for any other row presents no difficulties. We simply apply the sure-thing principle; i.e., we choose that clearly preferred row. Unfortunately this is the unusual situation, and this principle in itself is seldom selective of a unique course of action.

The classical principle of indifference of Laplace states that when we have no information about the different possible states of nature (the columns of the game) we should assume that the probabilities of each are equal and then choose that row whose expectation is maximal. The much more conservative minimax principle states that we should choose according to the hypothesis that we may expect the worst and thereby minimize our maximum loss. The minimax principle was first proposed by von Neumann as a sound principle of strategy in games against an intelligent opponent. Its extension to games against nature as a fundamental principle of statistical decision theory was first made by Wald. A variety of other principles have been proposed.

In view of the numerous suggestions that have been made, particular interest may be attached to John Milnor's analysis (1954) of what would seem to be the desirable characteristics of any rational principle of selection. His results, like those of Russell's paradox for the foundations of set theory, yield an impossibility theorem. Briefly, Milnor proposes the following nine axioms for any fully acceptable principle of choice. First, the principle must order the alternative courses of action. Second, this ordering must be independent of the arbitrary ordering of the rows and columns of the matrix. Third, the principle must be compatible with the sure-thing principle mentioned earlier; that is, if one row dominates another row in every column, that first row must be preferred. Fourth, the principle must satisfy an obvious condition of continuity. Fifth, the preference ordering of the courses of action must be unaffected by a linear change in all entries in the matrix. (This principle reflects the general result that the utility of consequences is measured only on an interval scale.) Sixth, the principle of choice must satisfy the condition of independence from irrelevant alternatives, namely that the ordering between old rows must not be changed by the addition to the game of a new row. The seventh axiom asserts that the principle must be invariant

under a linear change of any column; in other words, the ordering among courses of action must not change if a constant is added to some column. Eighth, the ordering generated by the principle of choice must be indifferent to a column duplication; that is, the ordering must not change if a new column identical with some old column is added to the game. The ninth axiom asserts the requirement of convexity, which means simply that, if a row is equal to the average of two other rows judged equivalent in the preference ordering, then the first row must not be judged worse than the two equivalent rows of which it is an average. This axiom is sometimes described as asserting that the decision-maker should not be prejudiced against randomization.

None of the familiar criteria of rationality for the decision-maker is compatible with all nine of these axioms. These nine axioms are obtained by extracting desirable properties of various criteria of rationality that have been proposed. Milnor shows, for example, that the Laplacean criterion is characterized by axioms 1, 2, 3, 6, and 7 and the Wald minimax criterion by axioms 1, 2, 3, 4, 6, 8, and 9. The undesirable negative result is that no criterion satisfies all nine together.

Closely related paradoxical results in confirmation theory suggest that the naive theory of rationality, like the naive theory of sets, cannot easily be systematically reconstructed in any simple and consistent fashion.

Certain moral philosophers will undoubtedly be inclined to dismiss the kind of results I have been discussing as applicable only to the technical problems that arise in the theory of induction. In their minds, the concept of rationality I have been discussing has little if any relevance to the concept of rationality that arises in moral theory. This I think is clearly a mistake, as the discussion in the next section is meant to show.

Elsewhere (Suppes, 1960a) I have discussed the difficulties of characterizing the pure theory of rationality even when we accept a principle of choice, in this case the Bayesian principle that enjoins the decision-maker to maximize his expected utility (the expectation being relative to a subjective probability distribution on the possible states or strategies of nature). Roughly speaking, the pure theory permits no structural assumptions about the environment, but is intended rather to hold always and everywhere. An example of an axiom of the pure theory is the postulate that the preference relation on the set of acts is transitive. A typical structural axiom is the assumption that the decision-maker can partition

the set of states of nature as fine as he pleases in terms of probability, an assumption which seems to me irrelevant to the pure concept of making a rational decision. I shall not try to summarize the technical results obtained, but only remark that they are all essentially negative and indicate the difficulties of axiomatizing even the simplest part of the pure theory of rationality.

By concentrating on the difficulties facing the development of any adequate concept of rationality, I do not mean to imply that the philosophical significance of decision theory under the heading of individual normative theory is restricted entirely to negative results. The revival of subjective probability, for instance, within the framework of decision theory by L. J. Savage and others has already had and will undoubtedly have further repercussions in the foundations of the theory of probability. The simple relative-frequency theories of von Mises and Reichenbach are already beginning to seem old-fashioned.

The development of utility theory within the general framework of decision theory has brought the kind of calculus envisioned long ago by Bentham to a high degree of technical perfection. It is unfortunate that the word 'utility' is connected in most philosophers' minds with hedonism. The formal calculus of utility developed in recent decades is no more committed to a calculus of pleasure than to one of duty. The important intellectual contribution of the hedonistic tradition has been the recognition that some principle of calculation is required for rational action in the face of partial or incomplete information. Some years ago Davidson, McKinsey, and I (1955) attempted to show the close relations that exist between the measurement of utility and a formal theory of value. I shall not attempt here to restate the arguments we gave, but only to reiterate my conviction that the recent formal theories of utility are as important for moral philosophy as are recent theories of subjective probability for the philosophical problems of induction.

III. GROUP NORMATIVE THEORY

In the table shown above three disciplines are listed in this quadrant: game theory, welfare economics, and political theory. The relations of game theory to the concept of rationality as discussed in the preceding section are apparent and will not be considered in any detail here. Suffice

it to say that one of the most satisfactory analyses of the concept of rationality exists for competitive games. This is the classical minimax theory of von Neumann already mentioned.

In this section I should like to turn from the concept of rationality to the more specifically moral concept of justice.

I have the impression that for many years the consideration of political theory by either philosophers or political scientists has been nearly identical with consideration of the history of political theory. It is my feeling that the newer welfare economics has broken out of its specific economic context and has already laid the foundations of a new approach to political theory. The central problem of welfare economics has been the Benthamite one of devising and analyzing a variety of schemes for the distribution of economic goods in the society. It has gradually come to be realized that the restriction to economic goods can be dropped and that the problem may be regarded as the more general one of setting social and political policy. It is assumed in this theoretical work that in some sense decisions reached reflect in an equitable and just manner the values and tastes of the members of the society. In fact, the preferences of the individuals making up the society are usually and somewhat unrealistically taken as given data.

To indicate the kind of results that may be obtained by the methods of welfare economics, I should like to sketch another impossibility theorem. This is due to Kenneth J. Arrow (1951) and is concerned with the existence of a just or equitable method of social decision. We suppose that there are a number of possible social states and that each member of the society has a preference ordering for these states. The problem is to construct an intuitively reasonable social preference ordering from the given individual orderings. One simple proposal is, of course, the method of majority decision. Social state A is preferred to social state B by the group as a whole if a majority of the members of the group prefer A to B, otherwise not. Just as in the case of the Laplacean principle of indifference or the minimax principle for choosing a strategy in games against nature, there are intuitively desirable axioms that are violated by the method of majority decision. Perhaps the easiest way to illustrate the difficulties is to describe the so-called paradox of voting. Suppose there are three issues A, B, C and three people voting on these issues. Let us assume that the first person prefers A to B to C, the second person prefers B to C to A,

and the third person prefers C to A to B. The issues are voted on in pairs. It is easily checked that if the first choice is between A and B the selected issue will be C. If it is between A and C the outcome will be B, and if it is between B and C the outcome will be A. In other words, the outcome chosen is completely dependent in this symmetrical situation on the arbitrary choice of which issues are to be voted first.

What Arrow has done is to proceed as Milnor did, namely, to write down reasonable axioms that any social decision method should satisfy and then to ask if indeed there exist any methods satisfying the axioms. In essence his four axioms are the following. The first axiom postulates a positive association of social and individual values. In particular, if one alternative social state rises in the ordering of every individual without any other change in the orderings, it is natural to postulate that it rises or at least does not fall in the social ordering. The second axiom states the independence of irrelevant alternatives. The meaning of this postulate is that if, for instance, a set of candidates is being considered for an office and the preferences of voters for these candidates are known, then the deletion of one candidate from the list will not affect the relative preferences for the other candidates. In thinking about this postulate it is important to emphasize that strategic considerations are not being considered. We are concerned with the actual preferences of the group members and not with their behavioral use of a strategy in those situations where they feel their first choice could not possibly be elected. For instance, in the 1948 presidential election a strong states' rights advocate might have preferred J. Strom Thurmond to Thomas E. Dewey, but, because he felt that Thurmond did not have a chance, he may have voted for Dewey. It is this kind of strategic consideration that is being ignored in this postulate. The meaning of the postulate is that, if the states' rights conservative preferred Thurmond to Dewey to Truman to Wallace, then if Dewey were no longer a candidate he would retain the same ordering of preferring Thurmond to Truman to Wallace, and similarly the deletion of any one of the four candidates would not disturb the preference ordering for the remaining three, even though the deletion of one of the four might affect his voting behavior. The third axiom asserts that the social decision method is not to be imposed. A decision method is said to be imposed when there is some pair of alternative social states X and Y such that the community can never express its preference for Y over X

no matter what the preferences of all the individuals concerned may be. The existence of outmoded taboos furnishes examples violating this condition. The fourth axiom asserts that the social decision method shall not be dictatorial; that is, the preferences shall not simply correspond to those of some one individual in the social group. Unfortunately, Arrow is able to prove that, if there is any degree of variety in the individual preference orderings, then there exists no social decision method satisfying the four axioms stated.

The point I have been concerned to make here is that the Milnor and Arrow impossibility theorems are as philosophically relevant to the foundations of the concepts of rationality and justice as are the paradoxes of set theory for the foundations of mathematics. These impossibility theorems demonstrate that our naive intuitions about rationality and justice cannot be counted upon to yield a coherent and consistent theory.

IV. DESCRIPTIVE THEORY

Perforce the descriptive or behavioristic theory has less direct relevance to philosophical problems than the normative theory, but there is one particularly important point I should like to describe in the brief space remaining. There are a rapidly increasing number of experimental studies of the actual decision behavior of human beings. Studies have been made of value choices, of actual inductive behavior, and of behavior in competitive or cooperative game contexts. The literature is too large to review here, but an important general conclusion seems to be that actual behavior deviates rather sharply from the normative models, even in the case of strictly competitive games. On the other hand, the behavior of experimental subjects in many cases corresponds well with quantitative predictions derived from learning theory formulated in terms of stimulus sampling and conditioning (Suppes and Atkinson, 1960).

Stimulus-sampling learning theory was first given a quantitative formulation in 1950 by the psychologist W. K. Estes. It has since been developed by a number of investigators. In a highly simplified form, the basic ideas run as follows. The organism is presented with a sequence of trials, on each of which he makes a response that is one of several possible choices. In any particular setup it is assumed that there is a set of stimuli from which the organism draws a sample at the beginning of

each trial. It is assumed that on each trial each stimulus is conditioned to at most one response. The probability of making a given response on any trial is postulated to be simply the proportion of sampled stimuli conditioned to that response, unless there are no conditioned stimuli in the sample, in which case there is a "guessing" probability for each response. Learning takes place by the following mechanism. At the end of a trial a reinforcing event occurs which identifies that one of the possible responses which was correct. With some fixed probability the sampled stimuli become conditioned to this response if they are not already, and the organism begins another trial in a new state of conditioning.

Independent of the question of empirical adequacy in predicting actual choice behavior, behavioral scientists have more general reasons for preferring a learning theory (like the stimulus-sampling variety just sketched) to decision theory. To the experimental psychologist, the static character of the concepts of subjective probability and utility is suspect, and these two concepts are the central concepts of decision theory. The psychologist resists accepting them as basic or primitive concepts of behavior. Ideally, what he desires is a dynamic theory of the inherent or environmental factors determining the acquisition of a particular set of beliefs or values. If these factors can be identified and their theory developed, the concepts of probability and utility become otiose in one sense. I have recently tried to show how stimulus-sampling learning theory provides the beginnings of such a development (Suppes, 1961a).[2†] The philosophical interest of the behavioristic approach lies in the possibility of constructing a more realistic framework than the static one of decision theory for discussing the normative theory of choice. This is not to say that the distinction between normative and descriptive questions is to be abolished. It is rather that the more detailed and fundamental behavioral theory opens up the possibility of analyzing normative questions at a deeper conceptual level, but this is a matter that cannot be explored here.

NOTES

[1] This sketch is not meant to be historically exact.

[2†] Article 9 in this volume.

8. AN AXIOMATIZATION OF UTILITY BASED ON THE NOTION OF UTILITY DIFFERENCES*

I. INTRODUCTION

In the literature of economics (e.g. Allais, 1952; Frisch, 1937; Lange-1934) the notion of utility differences has been much discussed in con, nection with the theory of measurement of utility.[1] However, to the best of our knowledge, no adequate axiomatization for this difference notion has yet been given at a level of generality and precision comparable to the von Neumann and Morgenstern construction of a probabilistic scheme for measuring utility. (The early study of Wiener, 1919–1920, is not axiomatically oriented.) The purpose of this paper is to present an axiomatization of this notion and to establish the expected representation theorem guaranteeing measurement unique up to a linear transformation.

Recent experimental work by economists and psychologists (see the bibliography in Edwards, 1954) suggests there are cogent reasons for reviving the notion of utility differences in order clearly to separate utility and subjective probability. The interaction between probability and utility makes it difficult to make unequivocal measurements of either one or the other. The recent Mosteller and Nogee experiments (1951) may be interpreted as measuring utility if objective probabilities are assumed or as measuring subjective probabilities if utility is assumed linear in money.

In Davidson *et al.* (1955) and Davidson and Suppes (1955) a detailed description is given of how utility may be experimentally measured by use of utility differences and a *single* chance event with subjective probability $\frac{1}{2}$.

The scheme may be briefly described as follows.[2] Let E^* be a chance event with subjective probability $\frac{1}{2}$, and suppose that the individual we are testing prefers outcome x to y, and outcome z to w. We present him with two alternative gambles, one of which he must choose. Gamble 1

* Reprinted from *Management Science* **1** (1955), 259–270. Written jointly with Muriel Winet.

is that if E^* occurs he gets x, and if E^* does not occur he gets w; Gamble 2 is that if E^* occurs he gets z, and if E^* does not occur he gets y. It seems intuitively reasonable to say that the individual should prefer Gamble 2 if and only if the utility difference between x and y is less than that between z and w. Once utility is measured by a procedure of this kind, we may measure subjective probabilities. (To some extent, this approach was anticipated in Ramsey, 1931.)

Since the chance event E^* is fixed throughout the discussion, it does not play any formal role in our axiomatization and enters only via one particular empirical interpretation of the notion of utility differences. Consequently, interpretations of our primitive notions, completely divorced from any probability questions, are available for analyzing other approaches to utility theory. A justification for considering alternative schemes is the limited applicability of the probabilistic approach just described. It can and has been used in some laboratory experiments at Stanford (Davidson *et al.*, 1955), but it is far from clear that it can be seriously applied to market behavior. An interpretation of utility differences in terms of amounts of money is an obvious alternative. We present such a scheme in the form of a *reduction* sentence (the general character of reduction sentences is discussed in Carnap, 1936, 1937). For simplicity we consider a fixed individual, say, Jones, and we assume that a prior satisfactory analysis of preference (as opposed to preference differences) has already been given.

(1) IF: (i) *Jones prefers commodity x to commodity y, and commodity u to commodity v*, (ii) *Jones has in his possession commodities y and v, and* (iii) *Jones is presented with the opportunity of paying money to replace y by x and v by u*, THEN: *the utility difference between x and y is at least as great as that between u and v if and only if Jones will pay at least as much money to replace y by x as to replace v by u.*

An obvious objection to (1) is that it has the effect, so often argued against, of *measuring* utility in terms of money. However, the only assumption needed for (1) is that the relation between amounts of money and utility differences is monotonically increasing. A linear relation is *not* required. In our opinion such a monotonicity assumption is very reasonable for a wide variety of persons and situations.

An alternative reduction may easily be stated in terms of work. It should be clear that the choice of money or work is not meant to entail

any special status for these two commodities. What is needed as a basis for constructing other reductions is simply the existence of a commodity flexible enough to serve in different situations and such that its marginal utility is either always positive or always negative in the situations under consideration.

In view of the many complex issues involved in assessing the workability, even in principle, of such reductions, it may be more useful to describe a particular experimental set-up which could be used to measure utility differences. For reasons which will become obvious, this scheme would not be directly applicable to market behavior, but on the other hand, it does not presuppose any fixed relations between money and other commodities.

For definiteness, we consider six household appliances of approximately the same monetary value, for instance, a mixer, a deluxe toaster, an electric broiler, a blender, a waffle iron and a waxer. A housewife who does not own any of the six is chosen as subject. Two of the appliances are selected at random and presented to the housewife, say, the toaster and the waxer. She is then confronted with the choice of trading the toaster for the waffle iron, or the waxer for the blender. Presumably she will exchange the toaster for the waffle iron if and only if the utility difference between the waffle iron and the toaster is at least as great as the difference between the blender and the waxer (due account being taken of the algebraic sign of the difference). A sequence of such exchanges (repetitions permitted) can easily be devised such that every utility difference is compared to every other. Our axioms specify for the set of choices sufficient ideal properties to guarantee the existence of a cardinal utility function.[3]

From another conceptual standpoint (as pointed out to us by our colleague, Professor Davidson), we may think of the housewife as expressing a simple preference between *pairs* of appliances. Thus if she trades the toaster for the waffle iron she has decided that she would rather have the pair (waffle iron, waxer) than the pair (toaster, blender). Put in these terms we are asking for a utility function φ of the Frisch (1932) and Fisher (1927) type such that one pair (x, y) is preferred to another (u, v) if and only if

$$\varphi(x) + \varphi(y) > \varphi(u) + \varphi(v).$$

The existence of such a function is taken to mean that 'utilities are independent', that is, the commodities involved are neither complementary nor competitive with respect to each other. Viewed in this light, our axioms analyze the special conditions required for the existence of a cardinal utility function on a set of *independent* commodities. Whatever one's *a priori* feelings about the plausibility of the independence hypothesis there can be little doubt that the experiment just described would provide a means of empirically testing the hypothesis[4], and thus would satisfy Samuelson's methodological demand (1947, p. 183):

> It may be argued that regarded purely as a working hypothesis the facts do not sharply contradict the independence assumption. A little investigation reveals that such a hypothesis has not been tested from this point of view. On the contrary, it is implicitly assumed from the beginning in the manipulation of the statistical data. Hence, one would have to go back to examine the original empirical data.

It is interesting to note that the problem of complementarity occupies a position in this interpretation analogous to the position occupied by the problem of a specific utility of gambling in a probabilistic interpretation.

It is also our opinion that many areas of economic and modern statistical theory do not warrant a behavioristic analysis of utility. In these domains, there seems little reason to be ashamed of direct appeals to introspection. For example, in welfare economics there are sound arguments for adopting a subjective view which would justify the determination of utility differences by introspective methods. Some psychological experiments on utility differences which essentially use introspective methods are reported in Coombs and Beardslee (1954).

It is to be emphasized that the formal results presented in the remainder of this paper do not depend on any of the particular interpretations here proposed.

II. PRIMITIVE AND DEFINED NOTIONS

Our axiomatization is based on three primitive notions. The primitive K is a nonempty set, to be interpreted as a set of alternatives (objects, experiences, events, or decisions) available to a given individual at a given time. The primitive Q is a binary relation whose field is K; the interpretation of Q is that $x\,Q\,y$ if and only if the individual does not prefer y to x. The third primitive is a quaternary relation R whose field

is also K. In the intended interpretation $x, y\, R\, z, w$ if and only if the difference in preference between x and y is not greater than the difference in preference between z and w.

Our axiomatization assumes a rather complicated form if it is given only in terms of our three primitives. It is intuitively desirable to use some defined notions whose interpretation follows directly from that of the primitives.

DEFINITION D1: $x\, I\, y$ *if and only if* $x\, Q\, y$ *and* $y\, Q\, x$. Obviously, I is the relation of *indifference*.

DEFINITION D2: $x\, P\, y$ *if and only if not* $y\, Q\, x$. The relation P is the relation of *strict preference*.

DEFINITION D3: $x, y\, E\, z, w$ *if and only if* $x, y\, R\, z, w$ *and* $z, w\, R\, x, y$. The interpretation of the quaternary relation E is that if x, y, z and w are alternatives, then $x, y\, E\, z, w$ if and only if the difference in preference between x and y is *equivalent* to the difference in preference between z and w.

DEFINITION D4: $x, y\, S\, z, w$ *if and only if not* $z, w\, R\, x, y$. Clearly, $x, y\, S\, z, w$ if and only if the difference in preference between x and y is *strictly less* than the difference between z and w.

DEFINITION D5: $B(y, x, z)$ *if and only if either* $x\, P\, y$ *and* $y\, P\, z$, *or* $z\, P\, y$ *and* $y\, P\, x$. The intuitive idea of *betweenness* is expressed by the relation B.

The above notions suffice for the statement of all but the last axiom, the Archimedean axiom. For the latter, one further quaternary relation is needed.

DEFINITION D6: $x, y\, M\, z, w$ *if and only if* $y\, I\, z$ *and* $B(y, x, w)$ *and* $x, y\, E\, z, w$. The quaternary relation M appears to be a trivial specialization of the relation E. To clarify this situation, we introduce the notion of powers of M. The second power of M, for example, is the relation M^2 such that $x, y\, M^2\, z, w$ if and only if there exist elements u and v such that $x, y\, M\, u, v$ and $u, v\, M\, z, w$. The nth power of M is defined recursively:

$$x, y\, M^1\, z, w \quad \text{if and only if,} \quad x, y\, M\, z, w;$$
$$x, y\, M^n\, z, w \quad \text{if and only if there exist elements } u \text{ and } v$$
$$\text{such that} \quad x, y\, M^{n-1}\, u, v \quad \text{and} \quad u, v\, M\, z, w.$$

The difference between powers of E and of M may be brought out by interpreting x, y, z, and w as points on a line. The interpretation of x,

$y\ M^3\ z, w$, for instance, is that the intervals (x, y) and (z, w) are of the same length, and there are two intervals of this length between y and z. Of special significance is the fact that the interval (x, w) is four times the length of (x, y). On the other hand, in the case of the relation E^3 no specific length relation may be inferred for intervals (x, w) and (x, y).

As we shall see in Section V, the proof of our representation theorem essentially depends on exploiting the properties of the powers of M.

III. AXIOMS

Using our primitive and defined notions, we now state our axioms for difference structures.

A system $\mathscr{K} = \langle K, Q, R\rangle$ *will be said to be a* DIFFERENCE STRUCTURE *if the following eleven axioms are satisfied for every* $x, y, z, w, u,$ *and* v, in K:

Axiom A1: $x\ Q\ y$ *or* $y\ Q\ x$;

Axiom A2: *If* $x\ Q\ y$ *and* $y\ Q\ z$ *then* $x\ Q\ z$;

Axiom A3: $x, y\ R\ z, w$ *or* $z, w\ R\ x, y$;

Axiom A4: *If* $x, y\ R\ z, w$ *and* $z, w\ R\ u, v$ *then* $x, y\ R\ u, v$;

Axiom A5: $x, y\ R\ y, x$;

Axiom A6: *There is a* t *in* K *such that* $x, t\ E\ t, y$;

Axiom A7: *If* $x\ I\ y$ *and* $x, z\ R\ u, v$, *then* $y, z\ R\ u, v$;

Axiom A8: *If* $B(y, x, z)$ *then* $x, y\ S\ x, z$;

Axiom A9: *If* $B(y, x, z)$ *and* $B(w, u, v)$ *and* $x, y\ R\ u, w$ *and* $y, z\ R\ w, v$, *then* $x, z\ R\ u, v$;

Axiom A10: *If* $x, y\ S\ u, v$ *then there is a* t *in* K *such that* $B(t, u, v)$ *and* $x, y\ R\ u, t$;

Axiom A11: *If* $x, y\ R\ u, v$ *and not* $x\ I\ y$, *then there are elements* s *and* t *in* K *and a positive integer* n *such that* $u, s\ M^n\ t, v$ *and* $u, s\ R\ x, y$.

The interpretation of Axioms A1–A4 is obvious. Axiom A5 expresses a commutativity property of R and means essentially that for pairs of elements to stand in the relation R only their differences matter and not their relative order.

Axiom A6 means intuitively that between any two elements of K, there is a midpoint. This axiom represents a more reasonable assumption than, for instance, a formulation requiring that between any two elements there exist an element some arbitrary part, say $\frac{1}{17}$th, of the distance between them. Indeed, the axiom as here stated, receives empirical corroboration

in the field of psychology from the practice of 'fractionation' and 'bisection' experiments requiring the subject to select the tones in just the way described, and from the existence of laboratory equipment designed for such experimental use. (See, e.g., Stevens, 1936, and Stevens and Volkmann, 1940.) Also, the probabilistic experiments (Davidson *et al.*, 1955) described in the first section have demonstrated the practicality of finding such midpoints.

Axiom A10 means that if the difference between x and y is less than that between u and v, then there is an element t of K between u and v and the difference between x and y is not greater than the difference between u and t.

Axiom A11, the Archimedean axiom, means that if the difference between x and y is not greater than that between u and v, and if x is not indifferent to y, then there are n elements of K equally spaced in utility between u and v such that the difference between any consecutive two of these elements is not greater than the difference between x and y.

IV. ELEMENTARY THEOREMS

A rather large number of elementary theorems is required for the complete proof of our representation theorem for difference structures. In the present paper, however, we are concerned merely to sketch the main outlines of such a proof; and, for this purpose, it will be sufficient in this section to present definitions of certain relations, not needed for stating the axioms, but used in a key way to develop the required proof; and to state without proof several elementary theorems which describe typical properties of the relations defined, or which figure centrally in the sketched proof of the representation theorem. In particular, we omit completely a large group of theorems which develops the expected properties of Q and R and of the other simple 'qualitative' relations (I, P, E, S, B) described in Section II.

We first introduce the notion of the quaternary relation $N(a)$.

DEFINITION D7: *$N(a)$ is the quaternary relation defined as follows*

(i) *if $a=1$, then $x, y\ N(a)\ u, v$ if and only if $x\ I\ u$ and $y\ I\ v$*

(ii) *if $a\neq 1$, then $x, y\ N(a)\ u, v$ if and only if $x\ I\ u$ and there exists a z such that $x, y\ M^{a-1}\ z, v$.*

The interpretation of $N(1)$, of course, is obvious. To say for $a\neq 1$, that

$x, y\ N(a)\ u, v$ means that x and u coincide, and that there are $a-1$ equally spaced elements of K between u and v such that the difference between any two of them equals the difference between x and y. If x, y, u and v are interpreted as points on a line, this notion obviously corresponds to the intuitive notion of 'laying off' an interval on another interval; that is, we interpret $x, y\ N(a)\ u, v$ intuitively as meaning that if we start from u, and 'lay off' an interval of the length (x, y) a times in the appropriate direction, we obtain the interval (u, v). By means of the $N(a)$ relation, therefore, we are able to express the quantitative fact that the length of an interval (u, v) is a times the length of a subinterval (x, y).

The sort of 'multiplication' of intervals characterized by the $N(a)$ relation possesses the expected properties; for example, we have the following theorem concerning ratios of intervals.

THEOREM 1: *If* $x, z\ N(a)\ x, y$ *and* $x, z\ N(ab)\ x, w$ *then* $x, y\ N(b)\ x, w$.

Another theorem involving the $N(a)$ relation generalizes A6 and may be justified along similar lines. Characteristic of our system, it asserts that appropriate elements exist for dividing any interval into powers of 2.

THEOREM 2: *If not* $x\ I\ y$ *then there is a* z *such that* $x, z\ N(2^m)\ x, y$.

Further $N(a)$-theorems state properties of 'N-multiplication' for powers of 2. We have, for example, the usual law for addition of exponents:

THEOREM 3: *If* $x, w\ N(2^m)\ x, z$ *and* $x, z\ N(2^n)\ x, y$ *then* $x, w\ N(2^{m+n})\ x, y$.

A crucial, but less obvious property is stated in the following theorem.

THEOREM 4: *If* $B(y, x, z)$ *and* $x, t\ N(2^m)\ x, y$ *and* $y, s\ N(2^m)\ y, z$ *and* $x, r\ N(2^m)\ x, z$ *then* $t, r\ E\ y, s$.

We now define a relation in terms of which most of the proof of the representation theorem is carried through.

DEFINITION D9: $H(m, a; n, b)$ *is the quaternary relation such that* $x, y\ H(m, a; n, b)\ u, v$ *if and only if there are elements* z_1, z_2, w_1 *and* w_2 *such that* $x, z_1\ N(2^m)\ x, y$ *and* $u, w_1\ N(2^n)\ u, v$ *and* $x, z_1\ N(a)\ x, z_2$ *and* $u, w_1\ N(b)\ u, w_2$ *and* $x, z_2\ R\ u, w_2$.

To say that $x, y\ H(m, a; n, b)\ u, v$ means intuitively that an $(a/2^m)$th part of the interval (x, y) is not greater than a $(b/2^n)$th part of the interval (u, v).

We may view our first theorem on this notion as enabling us to specify a partial bound for the values of arguments satisfying the H-relation between two intervals.

THEOREM 5: *If not* $x\,I\,y$ *and* $x, y\ H(m, a; n, b)\ u, v$, *then not* $u, v\ H(n, b; m+1, a)\ x, y$.

Since the H-relation can be thought of intuitively as a special sort of inequality, we would expect to be able to prove many of the laws governing inequalities. Thus Theorem 6 expresses a kind of transitivity property and Theorem 7 an intuitively simple conservation property. Theorems 8, 9, 10 and 11 assert cancellation and multiplication laws.

THEOREM 6: *If* $x, y\ H(m, a; n, b)\ u, v$ *and* $u, v\ H(n, b; p, c)\ r, s$ *then* $x, y\ H(m, a; p, c)\ r, s$.

THEOREM 7: *If not* $x, y\ H(m, a; n, b)\ u, v$ *and* $w, z\ H(p, c; n, b)\ u, v$ *and* $a \leqq 2^m$ *and not* $x\,I\,y$, *then not* $x, y\ H(m, a; p, c)\ w, z$.

THEOREM 8: *If* $x, y\ H(m, a; n, b)\ u, v$ *and* $ac \leqq 2^m$ *and* $bc \leqq 2^n$, *then* $x, y\ H(m, ac; n, bc)\ u, v$.

THEOREM 9: *If* $x, y\ H(m, ac; n, bc)\ u, v$, *then* $x, y\ H(m, a; n, b)\ u, v$.

THEOREM 10: *If* $x, y\ H(m, a; n, b)\ u, v$ *and either* $m \neq 0$ *or not* $x\,I\,y$, *then* $x, y\ H(m+c, a; n+c, b)\ u, v$.

THEOREM 11: *If* $x, y\ H(m+c, a; n+c, b)\ u, v$ *and* $a \leqq 2^m$ *and* $b \leqq 2^n$ *then* $x, y\ H(m, a; n, b)\ u, v$.

Theorem 12 states an addition property for the arguments of the H-relation in the case of adjacent intervals.

THEOREM 12: *If* $B(y, x, z)$ *and* $a+b \leqq 2^n$ *and* $x, y\ H(m, 1; n, a)\ u, v$ *and* $y, z\ H(m, 1; n, b)\ u, v$ *then* $x, z\ H(m, 1; n, a+b)\ u, v$.[5]

Finally, we state two existence theorems for arguments of the H-relation. These theorems are the form in which we make use of our purely qualitative continuity axiom (A10) and our Archimedean axiom (A11) respectively.

THEOREM 13: *If* $x, y\ S\ u, v$, *then there are integers* b *and* n *such that* $b < 2^n$ *and* $x, y\ H(0, 1; n, b)\ u, v$.

THEOREM 14: *If not* $u\,I\,v$, *then there is an integer* m *such that* $x, y\ H(m, 1; 0, 1)\ u, v$.

V. REPRESENTATION THEOREM

Our desired representation theorem is an immediate consequence of the following lemma. (As a matter of fact, it is rather customary in the theory of measurement to label a lemma of this sort the "theorem of adequacy" and not to state explicitly a representation theorem. Cf., e.g., von Neumann and Morgenstern 1947, pp. 24–29.)

Fundamental Lemma: Let $\mathscr{K} = \langle K, Q, R \rangle$ *be a difference structure. Then:*

(A) *There exists a real-valued function* ϕ *defined on* K *such that for every* x, y, z, w *in* K,

(i) $x\ Q\ y$ *if and only if* $\phi(x) \leqq \phi(y)$, *and*

(ii) $x, y\ R\ z, w$ *if and only if* $|\phi(x) - \phi(y)| \leqq |\phi(z) - \phi(w)|$.

(B) *if* ϕ_1 *and* ϕ_2 *are any two functions satisfying* (A), *then there exist real numbers* α *and* β *with* $\alpha > 0$ *such that for every* x *in* K, $\phi_1(x) = \alpha\phi_2(x) + \beta$.

Proof of Part A: We begin by choosing two elements u and v in K such that $u\ P\ v$ (if no such two elements exist, the proof is trivial). We next define for x and y in K the set of numbers $\mathscr{S}(x, y; u, v)$. A rational number r is in $\mathscr{S}(x, y; u, v)$ if and only if there are non-negative integers m and n and a positive integer b such that $b \leqq 2^n$ and $r = (b2^m)/2^n$ and $x, y\ H(m, 1; n, b)\ u, v$.

Let r and r' be positive rational numbers. Using Theorems 8, 10, 6, 9 and 11, in that order, we may easily prove that

(1) If $r \in \mathscr{S}(x, y; u, v)$ and $r < r'$ then $r' \in \mathscr{S}(x, y; u, v)$.

Using now principally Theorem 14 and Theorem 5 we may show that if not $x\ I\ y$ then the set $\mathscr{S}(x, y; u, v)$ has a positive number as a lower bound. Since by Theorem 14 $\mathscr{S}(x, y; u, v)$ is not empty, we conclude that it has a greatest lower bound. We use this fact to define the function $f_{(u, v)}$:

$$f_{(u, v)}(x, y) \quad \text{is the greatest lower bound of} \quad \mathscr{S}(x, y; u, v).$$

Obvious arguments prove that

$$f_{(u, v)}(x, y) = 0 \quad \text{if and only if} \quad x\ I\ y$$

and

$$f_{(u, v)}(u, v) = 1,$$

the choice of (u, v) thus corresponding to choice of a unit of length.

We obtain by an indirect argument from (1) that for any rational number r

(2) If $f_{(u, v)}(x, y) < r$ then $r \in \mathscr{S}(x, y; u, v)$,

and we are in a position to establish:

(3) If $x, y\ R\ z, w$ then $f_{(u, v)}(x, y) \leqq f_{(u, v)}(z, w)$.

(The proof is trivial in case $x\,I\,y$; hence we assume: not $x\,I\,y$.) Suppose, if possible, that $f_{(u,v)}(z, w) < f_{(u,v)}(x, y)$. Then there are integers m, n, b such that

$$f_{(u,v)}(z, w) < b2^m/2^n < f_{(u,v)}(x, y).$$

From (2) we then obtain:

$$z, w\ H(m, 1; n, b)\ u, v \quad \text{and not} \quad x, y\ H(m, 1; n, b)\ u, v.$$

Hence by Theorem 7, not $x, y\ H(m, 1; m, 1)\ z, w$, and thus by Theorem 10 and D9, not $x, y\ R\ z, w$, which contradicts the hypothesis of (3).

We next prove:

(4) If $f_{(u,v)}(x, y) \leqq f_{(u,v)}(z, w)$ then $x, y\ R\ z, w$.

Let

$$r = b_1 2^{m_1}/2^{n_1} \quad \text{be in} \quad \mathscr{S}(x, y; u, v)$$

and let

$$q = b_2 2^{m_2}/2^{n_2} \quad \text{be in} \quad \mathscr{S}(z, w; x, z).$$

Then we have: $b_1 \leqq 2^{n_1}$, $b_2 \leqq 2^{n_2}$, $x, y\ H(m_1, 1; n_1, b_1)\ u, v$, and $z, w\ H(m_2, 1; n_2, b_2)\ x, y$. Hence by Theorem 10, Theorem 8, and Theorem 6, $z, w\ H(m_1+m_2, 1; n_1+n_2, b_1b_2)\ u, v$. We conclude that

(5) rq is in $\mathscr{S}(z, w; u, v)$.

Now for the moment let

$$\alpha = f_{(u,v)}(x, y)$$
$$\beta = f_{(x,y)}(z, w)$$
$$\gamma = f_{(u,v)}(z, w).$$

Suppose, if possible, that $\alpha\beta < \gamma$. Then there is a positive ε such that $(\alpha+\varepsilon)\cdot(\beta+\varepsilon) = \gamma$. Clearly we may choose a number r in the open interval $(\alpha, \alpha+\varepsilon)$ and a number q in the open interval $(\beta, \beta+\varepsilon)$ such that r is in $\mathscr{S}(x, y; u, v)$ and q is in $\mathscr{S}(z, w; x, y)$. Since $rq < \gamma$, rq is not in $\mathscr{S}(z, w; u, v)$, but this contradicts (5), and we conclude that

(6) $f_{(u,v)}(x, y)\cdot f_{(x,y)}(z, w) \geqq f_{(u,v)}(z, w)$.

Suppose now that not $x, y\ R\ z, w$. By Theorem 13 it follows that there is an n and a b with $b/2^n < 1$ such that $z, w\ H(0, 1; n, b)\ x, y$, and we con-

clude that $f_{(x,y)}(z, w)<1$. Combined with (6), this result gives us: $f_{(u,v)}(x, y)>f_{(u,v)}(z, w)$, which contradicts our hypothesis, completing the proof of (4).

We now define the function $\phi_{(u,v)}$ as follows. For every x in K,

$$\phi_{(u,v)}(x)=\begin{cases} f_{(u,v)}(u, x), & \text{if } u\,Q\,x. \\ -f_{(u,v)}(u, x), & \text{if } x\,Q\,u. \end{cases}$$

We see at once that $\phi_{(u,v)}(u)=0$, and thus our choice of u corresponds to the choice of an origin. (3) and (4) provide the basis for an obvious proof that

(7) $\quad x\,Q\,y$ if and only if $\phi_{(u,v)}(x) \leqq \phi_{(u,v)}(y)$.

To complete the proof of Part A we need to show that

(8) $\quad x, y\,R\,z, w$ if and only if $|\phi_{(u,v)}(x)-\phi_{(u,v)}(y)| \leqq |\phi_{(u,v)}(z)-\phi_{(u,v)}(w)|$.

From (3) and (4) we see at once that it will be sufficient to prove

(9) $\quad f_{(u,v)}(x, y)=|\phi_{(u,v)}(x)-\phi_{(u,v)}(y)|$.

Of the five possible cases that need to be considered for (9) we consider only the typical one where $x\,P\,y$ and $y\,P\,u$. For this case we must prove:

(10) $\quad f_{(u,v)}(x, y)+f_{(u,v)}(u, y)=f_{(u,v)}(u, x)$.

Suppose, if possible, that

$$f_{(u,v)}(x, y)+f_{(u,v)}(u, y)<f_{(u,v)}(u, x).$$

Then clearly there are integers m, n, b, b_1, b_2 such that

$$\begin{aligned} &f_{(u,v)}(x, y)+f_{(u,v)}(u, y)<b2^m/2^n<f_{(u,v)}(u, x), \\ (11)\quad &f_{(u,v)}(x, y)<b_1 2^m/2^n \\ &f_{(u,v)}(u, y)<b_2 2^m/2^n, \end{aligned}$$

and

$$b=b_1+b_2 \leqq 2^n.$$

By (2) we have: $x, y\,H(m, 1; n, b_1)\,u, v$ and $y, u\,H(m, 1; n, b_2)\,u, v$. Hence by Theorem 12, $x, u\,H(m, 1; n, b_1+b_2)\,u, v$, but from (11), we infer: not $x, u\,H(m, 1; n, b_1+b_2)\,u, v$. On the basis of this contradiction,

we conclude that

$$f_{(u,v)}(x, y) + f_{(u,v)}(u, y) \geqq f_{(u,v)}(u, x), \tag{12}$$

and by an argument similar to the above we may show that equality holds in (12), thus establishing (9) for a typical case, and completing the proof of Part A.

Proof of Part B: Using elements u and v in K as in the proof of (A), we define functions h_1 and h_2 for every x in K by the equations:

$$h_1(x) = \frac{\phi_1(x) - \phi_1(u)}{\phi_1(v) - \phi_1(u)}$$
$$h_2(x) = \frac{\phi_2(x) - \phi_2(u)}{\phi_2(v) - \phi_2(u)},$$

where ϕ_1 and ϕ_2 are functions satisfying (A). Since $u\,P\,v$, we see at once that

$$h_1(u) = h_2(u) = 0$$
$$h_1(v) = h_2(v) = 1,$$

and that h_1 and h_2 satisfy (A). Thus in order to establish (B) it will be sufficient to prove that

$$h_1 = h_2. \tag{1}$$

We give the proof for the case where $u\,P\,x$ and $x\,P\,v$. Suppose, if possible, that $h_1(x) \neq h_2(x)$. For definiteness, let $h_1(x) < h_2(x)$. Then there is a positive ε such that

$$h_2(x) = h_1(x) + \varepsilon. \tag{2}$$

We now consider the smallest integer, say, n^*, such that $\frac{1}{2}^{n^*} < \varepsilon$. (Since $h_1(x)$ and $h_2(x)$ are both between 0 and 1, $n^* \neq 0$.) By Theorem 2 there exists an element, say, z^*, such that $u,\ z^*\ N(2^{n^*})\ u,\ v$. A simple argument shows that we must have: $z^*\,P\,x$.

Suppose now that there is an integer a such that $u,\ z^*\ N(a)\ u,\ x$. It is easy to prove by induction that we must then be able to infer:

$$h_1(x) = h_2(x) = a/2^{n^*}, \tag{3}$$

which contradicts (2).

Since on the supposition of (2) there is no such integer a, there must be

an integer b and elements z_1 and z_2 such that

$$(4) \qquad \begin{cases} u, z^* N(b) u, z_1 \\ u, z^* N(b+1) u, z_2 \\ z_1 P x \\ x P z_2 . \end{cases}$$

Using the induction which yielded (3), we have from (4),

$$h_2(z_2) - h_1(z_1) = (b+1)/2^{n^*} - b/2^{n^*} < \varepsilon ,$$

and we also obtain from (4):

$$h_1(z_1) < h_1(x) < h_1(z_2)$$
$$h_2(z_1) < h_2(x) < h_2(z_2).$$

Combining inequalities we conclude:

$$h_2(x) - h_1(x) < h_2(z_2) - h_1(z_1) < \varepsilon ,$$

which contradicts (2).

The proof of (1) is completed by a consideration of the four other possible cases for the position of x with respect to u and v. (Two of the cases are trivial: $u \, I \, x$ and $v \, I \, x$.) Since (1) establishes (B), the proof of our lemma is finished.

We would not expect to have a strict isomorphism between an arbitrary difference structure $\mathscr{K} = \langle K, Q, R \rangle$ and some numerical structure, since distinct elements which stand in the relation I are assigned the same number. However, by considering the coset algebra $\mathscr{K}/I = \langle K/I, Q/I, R/I \rangle$ of $\mathscr{K}$ under I, we may easily establish such an isomorphism. (Since I is obviously a congruence relation on K with respect to Q and R, it should be clear that K/I is the set of all I-equivalence classes and that Q/I and R/I are the relations between equivalence classes corresponding to Q and R.)

We define the quaternary relation T for real numbers as follows:

$$\text{if } \alpha, \beta, \gamma, \text{ and } \delta \text{ are real numbers, then } \alpha, \beta \, T \, \gamma, \delta$$
$$\text{if and only if } |\alpha - \beta| \leqq |\gamma - \delta| .$$

Let N be a set of real numbers. Then we call an ordered triple $\langle N, \leqq, T \rangle$ a *numerical difference structure* if N is closed under the formation of

midpoints, i.e., if α, β, are in N, then $(\alpha+\beta/2)$ is in N. We then obtain the following representation theorem as an immediate consequence of our lemma.

REPRESENTATION THEOREM: *If* $\mathscr{K}=\langle K, Q, R\rangle$ *is a difference structure, then* $\mathscr{K}/I=\langle K/I, Q/I, R/I\rangle$ *is isomorphic to a numerical difference structure. Moreover any two numerical difference structures isomorphic to* $\mathscr{K}/I$ *are related by a linear transformation.*

NOTES

[1] The formally similar notion of sensation differences is important in the literature of psychology (e.g., Coombs, 1950; Guilford, 1936; Hanes, 1949; Stevens, 1936; Stevens and Volkmann, 1940).

[2] The intuitive idea of this approach was primarily due to Donald Davidson. It was suggested in Davidson *et al.* (1954) and has been the basis for the experiments reported in Davidson *et al.* (1955).

[3] By considering just six items, we cannot get a realization of the axioms given in Section III. However, by increasing the number of items, we would presumably be able to get a successively closer approximation.

[4] Some experiments are planned in collaboration with Professor Davidson.

[5] We are indebted to Herman Rubin for the proof of this theorem.

9. BEHAVIORISTIC FOUNDATIONS OF UTILITY*[1]

In the past two decades there has been an intensive development of the subject of decision making. A variety of objectives and viewpoints has dominated the constructive as well as the critical work on the subject. Nonetheless a pervasive goal of nearly all contributors has been the elucidation of a theory of rationality for purposive behavior in situations of risk and uncertainty. Intuitively we expect every considered judgment or decision of a serious person to be rational in some definite sense. Certain authorities would maintain even that every considered decision of any mammalian organism is rational in the sense of representing the attempt to maximize some significant quantity. The most prominent "maximization" analysis of rationality is the thesis that the decision-maker should maximize expected utility or value with respect to his beliefs concerning the facts of the situation. To perform this maximization, he needs to have, or to act as if he had, a subjective probability function measuring his degrees of belief and a utility function measuring the relative value to him of the various possible outcomes of his actions or decisions.

It is not my purpose here to expound the expected utility theory of behavior. An excellent detailed and leisurely analysis is Savage (1954). Rather, my concern is to explore the extent to which behavioristic foundations can be supplied for utility. And I am using the term 'behavioristic' in the rather narrow sense of the experimental psychologist. The static character of the concepts of subjective probability and utility is suspect to the psychologist and he resists accepting them as basic concepts of behavior. Ideally, what is desired is a dynamic theory of the inherent or environmental factors determining the acquisition of a particular set of beliefs or values. Moreover, in the notions of stimulus, response, and reinforcement the experimental psychologist has a triad of concepts which have proved adequate to explain much simple choice behavior. It is,

* Reprinted from *Econometrica* **29** (1961), 186–202.

therefore, a scientific problem of some interest to try to use just these behavioristic notions to derive a theory of subjective probability and utility.

In the first section I set forth the fundamental assumptions of stimulus-sampling learning theory, which is the most formally sophisticated theory yet stated in terms of the concepts of stimulus, response, and reinforcement. In the second section I attempt to show how this theory may be used to derive a utility function for various simple choice situations. This derived utility function is for stochastic choice behavior of the kind studied by Davidson and Marschak (1959), Luce (1959), Papandreou (1957), and others. In the third and final section, the earlier results are related to Shannon's concept of entropy and Luce's choice axiom.

I. STIMULUS-SAMPLING LEARNING THEORY

The theory to be used in this paper is a modification of stimulus sampling theory as first formulated by Estes (1950), Estes and Burke (1953), and Burke and Estes (1957). It is most closely connected with a formulation given by Suppes and Atkinson (1960), but it also differs, in ways indicated below, from the latter. The concepts of stimulus, response, and reinforcement and the processes of stimulus sampling and conditioning are the basic notions of the theory. In an economic situation a typical stimulus might be the price set by a competitor for a given product during the past quarter; the response, the firm's own price decision for the current quarter; and the reinforcement or reward, the quarterly gross profits. In a simple two-choice experiment the single stimulus might be the light signaling onset of the trial; the response, the pressing of one of two keys used to predict which one of two lights will flash; and the reinforcement, the actual flashing of one of these two lights.

The axioms are formulated verbally here, and although there is no attempt in this paper to give a mathematically exact statement of the theory, it is hoped that the relation between the fundamental axioms and the results derived later will be reasonably clear, even to the reader without prior familiarity with the literature. The first group of axioms deals with the conditioning of sampled stimuli, the second with the sampling of stimuli, and the third with responses.

Conditioning Axioms

C1. *On every trial each stimulus element is conditioned to exactly one response.*

C2. *If a stimulus element is sampled on a trial it becomes conditioned with probability θ to the response (if any) which is reinforced on that trial.*

C3. *If no reinforcement occurs on a trial there is a probability that the sampled stimulus becomes conditioned to some other response.*

C4. *Stimulus elements which are not sampled on a given trial do not change their conditioning on that trial.*

C5. *The probability of a sampled stimulus element being conditioned is independent of the trial number and the outcome of preceding trials.*

Sampling Axioms

S1. *Exactly one stimulus element is sampled on each trial.*

S2. *If on a given trial it is known what stimuli are available for sampling, then no further knowledge of the subject's past behavior or of the past pattern of reinforcement will change the probability of sampling a given element.*

Response Axiom

R1. *On any trial that response is made to which the sampled stimulus element is conditioned.*

Detailed remarks about these axioms are to be found in Suppes and Atkinson (1960). The major change from the version in Suppes and Atkinson (1960) is to be found in Axiom C3. There this axiom reads: "If no reinforcement occurs on a trial there is no change in conditioning on that trial." For the kind of experimental situation to be considered below it is natural to adopt the modified axiom given above as C3. A slight change in Axiom C5 has been made to accommodate the major change in C3; otherwise the axioms given here are those of Suppes and Atkinson (1960).

Many readers may be particularly critical of the first sampling axiom, S1. There are at least two different kinds of remarks to be made in defense of the assumption that exactly one stimulus element is sampled on each trial. In the first place, this assumption is mathematically extremely convenient and it is scarcely possible to distinguish, for the kind of experiments to be described here, between it and more "liberal" sampling axioms, as for example the assumption that all stimulus elements in the basic stimulus set are sampled with independent probabilities.

Secondly, S1 may be made more intuitively plausible by interpreting 'stimulus element' to mean *pattern of stimuli*, for it may be maintained that in any given situation an organism, at any given moment, is sampling exactly one pattern of stimuli. (For a more detailed discussion of the pattern concept, see Estes, 1957.)

We may consider two simple applications, which will be integrated into our discussion of utility in the next section. These two examples should serve adequately to illustrate how the basic axioms of stimulus-sampling theory are related to particular experimental situations in order to make predictions about response behavior.

Suppose the task presented a subject is to predict on each trial exactly which one of two lights will come on. Thus on each trial exactly one of two reinforcing events, E_1 or E_2, occurs. The subject indicates his prediction at the beginning of each trial by pressing one of two keys, response A_1 or A_2, where A_i is the key under light E_i. The sequence of events on a given trial may be described thus:

trial begins with stimuli conditioned to A_1 or A_2	→ stimulus sampled	→ response A_1 or A_2	→ reinforcement E_1 or E_2	→ possible change in conditioning of sampled stimulus.

Using the "independence of path" assumptions represented by Axioms C5 and S2, it may be shown that if we assume that the stimulus set S consists of exactly one element then the sequence of response random variables $\langle A_1, A_2, \ldots, A_n, \ldots \rangle$ is a Markov chain for many schedules of reinforcement satisfying the experimental conditions just described. (Here the value for each n of the random variable A_n is 1 or 2, according to whether the A_1 or A_2 response is made on trial n.) Using this result about Markov chains and the description of events on a trial, we may, upon imposition of a particular schedule of reinforcement, derive the transition matrix of the Markov chain. For consideration at this point we introduce the *simple contingent* case of reinforcement, namely, the probability of an E_1 or E_2 reinforcement on trial n depends only on the response made on trial n. Thus, using notation common in the literature:

$$P(E_1 \mid A_1) = \pi_1,$$
$$P(E_1 \mid A_2) = \pi_2.$$

The states of the Markov process are A_1 and A_2. Being in state A_1, for in-

stance, means that the single stimulus element is conditioned to A_1. The trees of the process are as shown in Figure 1.

The probabilities θ and $1-\theta$ occurring in the final branches of the trees are derived from Axiom C2, which is concerned with the con-

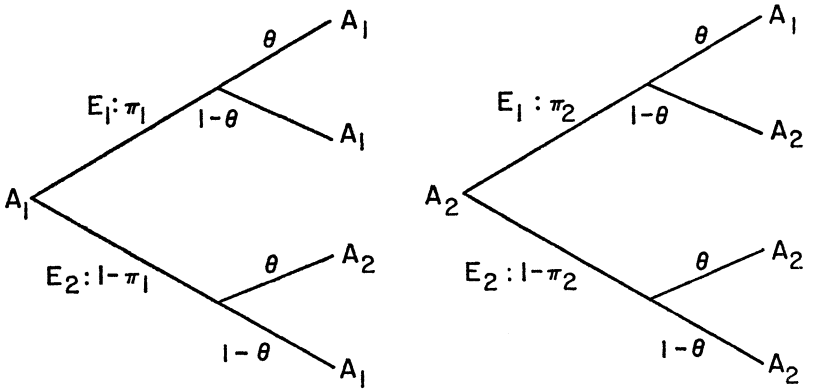


Fig. 1.

ditioning of stimulus elements. For example, in the lower half of the first tree, an E_2 reinforcement occurs with probability $1-\pi_1$ after the initial response A_1. This initial response means that the single stimulus element is connected (or conditioned) to A_1. However, an E_2 reinforcement occurs. With probability θ this reinforcement is effective in changing the connection or conditioning of the single stimulus element to the A_2 response.

We immediately derive from the two trees the following transition matrix for the Markov chain:

$$\begin{array}{c|cc} & A_1 & A_2 \\ \hline A_1 & 1-\theta(1-\pi_1) & \theta(1-\pi_1) \\ A_2 & \theta\pi_2 & 1-\theta\pi_2 \end{array}.$$

The asymptotic probability p_∞ of an A_1 response is easily computed from this matrix. The probability p_{n+1} of being in state A_1 is just

$$p_{n+1} = p_{11}p_n + p_{21}(1-p_n),$$

where p_{ij} is the transition probability of going from A_i to A_j in one trial.

(Thus p_{ij} is just the entry for the ith row and jth column of the transition matrix.) Now at asymptote

$$p_{n+1} = p_n = p_\infty,$$

whence

$$p_\infty = (1 - \theta(1 - \pi_1))\, p_\infty + \theta\pi_2(1 - p_\infty),$$

and this simple linear equation has as its solution

$$(1) \qquad p_\infty = \frac{\pi_2}{1 - \pi_1 + \pi_2}.$$

It is worth noting that the asymptotic probability p_∞ is independent of the conditioning parameter θ. Experimental evidence supporting Equation (1) is to be found in Estes (1954).

Rather than derive further predictions for the simple contingent case of reinforcement, I now turn to the second example, which I shall call the *two-arm bandit* case of reinforcement. The name stems from the resemblance of the experimental situation to that of playing a slot machine with two arms or levers rather than one; on each trial a choice between the levers is made. (Mathematical statisticians have, during the past few years, considered in detail what is the optimal way to play a two-arm bandit for a finite number of trials when the probabilities of pay-off of the two arms are unknown.)

The experimental situation, then, consists of choosing on each trial between two levers. In the experiment to be described in somewhat more detail in the next section, lever 1 is given a probability π_1 of paying, and lever 2 a probability π_2. Unlike the simple contingent case there is no "correction" procedure, i.e., the subject is not told, or led to believe, that on each trial exactly one of the arms of the "bandit" will pay off. If he chooses lever 1, say, then either it pays off or it does not, without reference to the possible choice of lever 2. Such an analysis of reinforcement leads to an application of Axiom C3: if lever i is chosen (i.e., response A_i occurs) and no reward or reinforcement follows (event E_0 occurs), then there is a probability ε_i that the sampled stimulus will become conditioned to the other response, i.e., choosing the other lever. Application of C3 to the present situation seems natural and intuitively sound, but it is to be emphasized that any uniform method, applicable to many other experiments, for handling nonreinforcement trials would be

premature in view of the highly conflicting experimental evidence obtained by various investigators, particularly in connection with the extinction of learning. The trees for the one-element model may be drawn as in Figure 2 (we have eliminated the θ and $1-\theta$ branches in case of reward, for they lead to the same result, namely, retention in the same state with which the trial began).

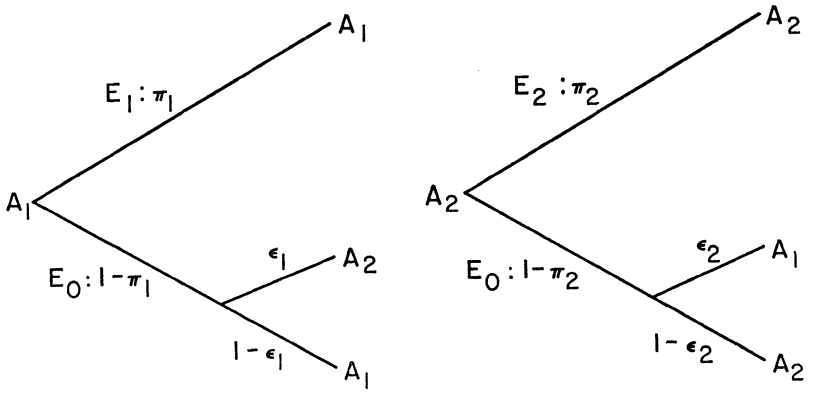


Fig. 2.

(Note that we use E_0 to designate the event of no reinforcement.)

The trees yield as the transition matrix of the Markov chain:

$$\begin{array}{c|cc} & A_1 & A_2 \\ \hline A_1 & 1-\varepsilon_1(1-\pi_1) & \varepsilon_1(1-\pi_1) \\ A_2 & \varepsilon_2(1-\pi_2) & 1-\varepsilon_2(1-\pi_2). \end{array}$$

And by the same line of argument which led to Equation (1) we obtain as the asymptotic probability p_∞ of the A_1 response for the two-arm bandit:

$$(2) \qquad p_\infty = \frac{\varepsilon_2(1-\pi_2)}{\varepsilon_1(1-\pi_1)+\varepsilon_2(1-\pi_2)}.$$

If $\varepsilon_1 = \varepsilon_2$, Equation (2) simplifies to:

$$(3) \qquad p_\infty = \frac{1-\pi_2}{(1-\pi_1)+(1-\pi_2)}.$$

In connection with these two applications of stimulus-sampling theory,

it is important to emphasize that the asymptotic probabilities (1) and (2) do not in any way depend on the assumption that there is exactly one stimulus element. In fact, the results (1) and (2) hold on the assumption of any finite number of stimulus elements. To illustrate the methods of working with more than one stimulus element, we may write down some

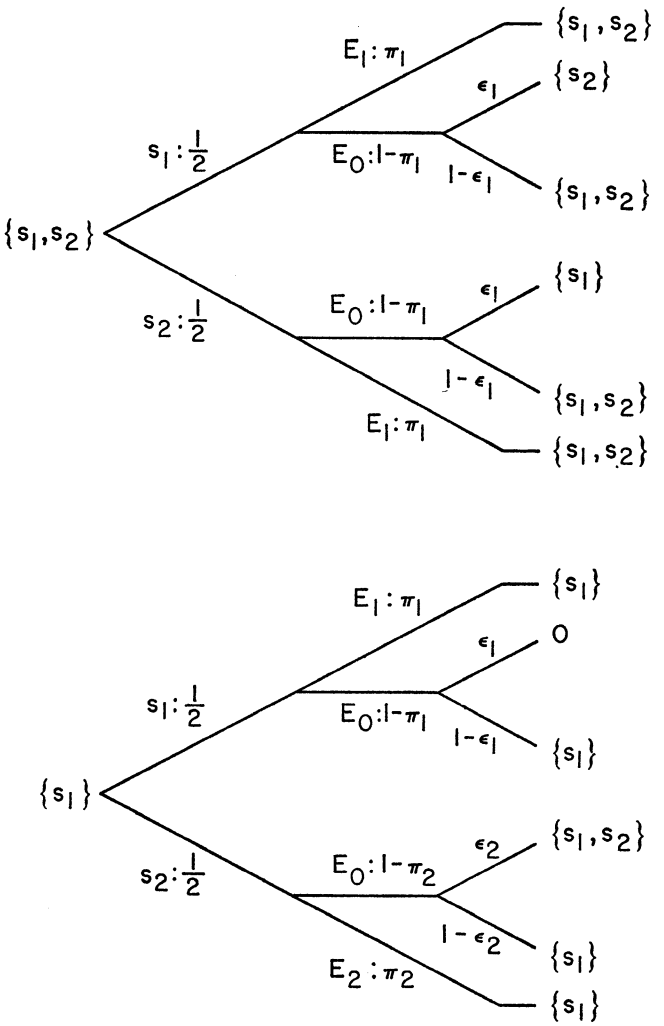


Fig. 3.

of the trees and the transition matrix for the two-element model as applied to the case of the two-arm bandit (see Figure 3). The states of the Markov chain are no longer the responses A_1 and A_2, but the possible partitions representing the conditioning of the two stimulus elements. Let s_1 and s_2 be the two elements. We may indicate any partition of the set $\{s_1, s_2\}$ between the two responses A_1 and A_2 simply by indicating which elements are conditioned to A_1. Thus the four states of the process may be denoted by $\{s_1, s_2\}$, $\{s_1\}$, $\{s_2\}$, and 0, where 0 is the empty set (meaning here that neither s_1 nor s_2 is conditioned to A_1 if the subject is in state 0). We give the trees when the subject begins in either state $\{s_1, s_2\}$ or $\{s_1\}$; the other two trees are similar to these. The one assumption needed, and not given in our fundamental axioms, is the probability of sampling s_1 as against that of sampling s_2. Here we assume there is an equal chance of sampling either, although this is not very crucial to any of our results.

Note that in the first tree either E_1 or E_0 must occur since both stimulus elements are conditioned to A_1, and thus only the A_1 response occurs regardless of which element is sampled. This is not the case for the second tree; if s_1 is sampled A_1 occurs and then either E_1 or E_0, but if s_2 is sampled A_2 occurs and then either E_2 or E_0. The transition matrix to be derived from these two trees and the other two not shown here is the following:

	$\{s_1, s_2\}$	$\{s_1\}$	$\{s_2\}$	0
$\{s_1, s_2\}$	$1-\varepsilon_1(1-\pi_1)$	$\frac{1}{2}\varepsilon_1(1-\pi_1)$	$\frac{1}{2}\varepsilon_1(1-\pi_1)$	0
$\{s_1\}$	$\frac{1}{2}\varepsilon_2(1-\pi_2)$	$1-\frac{1}{2}\varepsilon_1(1-\pi_1)-\frac{1}{2}\varepsilon_2(1-\pi_2)$	0	$\frac{1}{2}\varepsilon_1(1-\pi_1)$
$\{s_2\}$	$\frac{1}{2}\varepsilon_2(1-\pi_2)$	0	$1-\frac{1}{2}\varepsilon_1(1-\pi_1)-\frac{1}{2}\varepsilon_2(1-\pi_2)$	$\frac{1}{2}\varepsilon_1(1-\pi_1)$
0	0	$\frac{1}{2}\varepsilon_2(1-\pi_2)$	$\frac{1}{2}\varepsilon_2(1-\pi_2)$	$1-\varepsilon_2(1-\pi_2)$.

Note that the probability of an A_1 response when in state $\{s_1, s_2\}$ is one, when in states $\{s_1\}$ or $\{s_2\}$ is $\frac{1}{2}$, and when in state 0 is zero. Whence from computation of the asymptotic probabilities for each state we may at once determine the asymptotic probability of an A_1 response. As already remarked, the result is again Equation (2). We shall not consider the details of these computations here. In fact, at this point we end the consideration of stimulus-sampling theory in order to turn to utility theory proper.

II. UTILITY

As indicated in the introductory section, in this paper I am mainly con-

cerned with a utility function for the kind of choice behavior which has come to be labeled, not entirely happily, 'stochastic'. Roughly speaking, the central character of stochastic choice behavior is that upon presentation of two alternatives a and b, with a choice of one required, under essentially identical circumstances sometimes a will be chosen by a subject and sometimes b. Let $p(a, b)$, then, be the probability that a is chosen over b. A (*stochastic*) *utility function* for a set of alternatives A is a real-valued function u defined on A such that for every a, b, c and d in A

$$(4) \qquad p(a, b) \geqslant p(c, d) \text{ if and only if } u(a) - u(b) \geqslant u(c) - u(d).$$

Combining results in Suppes and Winet (1955)[2†], Suppes (1956a), and Davidson and Marschak (1959), it may be shown that if the set A and the probabilities $p(a, b)$ satisfy the following axioms, then there exists a stochastic utility function for A, and moreover this function is unique up to a positive linear transformation.

Axiom U1: $p(a, b)+p(b, a)=1$.

Axiom U2: $0<p(a, b)<1$.

Axiom U3: *If* $p(a, b)\geqslant p(c, d)$ *then* $p(a, c)\geqslant p(b, d)$.

Axiom U4: *There is a* c *in* A *such that* $p(a, c)=p(c, b)$.

Axiom U5: *If* $p(c, d)>p(a, b)>\frac{1}{2}$ *then there is an* e *in* A *such that* $p(c, e)>\frac{1}{2}$ *and* $p(e, d)\geqslant p(a, b)$.

Axiom U6: *(Archimedean Axiom)*: *If* $p(a, b)>\frac{1}{2}$ *then for every probability* q *such that* $p(a, b)>q>\frac{1}{2}$ *there is a positive integer* n *such that* $q\geqslant p(a, c_1)=p(c_1, c_2)=\cdots=p(c_n, b)>\frac{1}{2}$.

Now one implication of these six axioms is that A must be an infinite set if for at least two members a and b of A, $p(a, b)\neq\frac{1}{2}$. Simple and natural conditions, which are not unduly restricted and which will guarantee existence of a stochastic utility function for a finite set A, are not easily found. An unworkable recursive, but not finite, axiomatization can be given by enumerating for each n all isomorphism types. Some of the fundamental difficulties of finite axiomatization are brought out in Scott and Suppes (1958).[3†] The upshot of these axiomatic problems, it seems to me, is that for finite sets of alternatives we have no clear and intuitively natural ideas in terms only of probabilities of choice of the notion of utility, and thus of the notion of rationality for such situations.[4]

On the other hand, we may apply the results of the preceding section to indicate how from the axioms of stimulus-sampling theory a utility

function may be derived for finite sets of alternatives. To begin with, let us consider the second example of the application of stimulus-sampling theory, namely, the two-arm bandit. On each trial the subject must choose between two alternatives, but now, to make the utility considerations interesting, we assume there is a set of alternatives available, with choice restricted on each trial to one of a pair. Clearly alternative a does not in and of itself have more value than alternative b; the value of a is determined by the probability of pay-off, as is that of b. Thus the experimenter may manipulate the value of any alternative according to his determination of its pay-off function. We seek a function u satisfying (4). Now according to (2) of the last section, at asymptote,

$$p(a, b) = \frac{\varepsilon_b(1 - \pi_b)}{\varepsilon_a(1 - \pi_a) + \varepsilon_b(1 - \pi_b)} \tag{5}$$

where π_a is the probability of pay-off of alternative a when it is chosen, ε_a is the probability the sampled stimulus will become conditioned to the other alternative when the choice of a is not rewarded, and similar definitions hold for π_b and ε_b. In view of (5) to satisfy (4), we need to find a function u such that

$$\frac{\varepsilon_b(1 - \pi_b)}{\varepsilon_a(1 - \pi_a) + \varepsilon_b(1 - \pi_b)} \geqslant \frac{\varepsilon_d(1 - \pi_d)}{\varepsilon_c(1 - \pi_c) + \varepsilon_d(1 - \pi_d)} \tag{6}$$

$$\text{if and only if} \quad u(a) - u(b) \geqslant u(c) - u(d).$$

Let $\rho_a = \varepsilon_a(1 - \pi_a)$ for every a in A.[5] The right-hand inequality of (6) may then be written:

$$\frac{\rho_b}{\rho_a + \rho_b} \geqslant \frac{\rho_d}{\rho_c + \rho_d}, \tag{7}$$

but (7) holds, if and only if

$$\rho_b/\rho_c \geqslant \rho_a/\rho_d,$$

which holds, if and only if

$$\rho_b/\rho_a \geqslant \rho_d/\rho_c,$$

which again holds, if and only if

$$\frac{1/\rho_a}{1/\rho_b} \geqslant \frac{1/\rho_c}{1/\rho_c},$$

which, finally, holds, if and only if

$$\log 1/\rho_a - \log 1/\rho_b \geqslant \log 1/\rho_c - \log 1/\rho_d. \tag{8}$$

From (6), (7) and (8) we conclude that an appropriate utility function is, for a in the set A of alternatives:

$$u(a) = \log \frac{1}{\varepsilon_a(1-\pi_a)}. \tag{9}$$

If $\varepsilon_a = \varepsilon_b$ for every a and b in A, we may take the simpler function

$$u'(a) = \log \frac{1}{1-\pi_a}.$$

It is straightforward to show that the utility function defined by (9) is unique up to a positive linear transformation if the reasonable restriction is made that any acceptable utility function must be continuous in ε_a and π_a. Moreover, from the existence of a function u satisfying (4), it immediately follows that the asymptotic choice behavior predicted by stimulus-sampling theory satisfies all the various conditions of weak and strong stochastic transitivity discussed in the literature, as well as the quadruple condition expressed by Axiom U3 above. It should be mentioned that these results do not necessarily hold during the course of learning; in particular the utility function defined by (9) does not satisfy (4) during the course of learning. This fact, it seems to me, accords well with the widespread assumption, albeit often tacit, that the utility function of a person is an equilibrium concept. It may also be noted that the numerical-valued utility function defined by (9) may be replaced by a function whose values are probability distributions if the basic theory is formulated so that $p(a, b)$ is a random variable rather than a number. Unfortunately the derivation of the distribution of this random variable is tedious and difficult. As formulated here, the number $p(a, b)$ is the asymptotic expectation of the response random variable that has the value 1 for choice of a and 0 for choice of b. The utility function of (9) is defined in terms of this expectation and is not sensitive to the trial-by-

trial fluctuations in the values of the response random variable itself, which is another facet of its equilibrium character.

It is, of course, to be emphasized that the utility function defined by (9) is not that of the mathematical statistician bent on maximizing his monetary pay-off in the long run. It should be abundantly clear that the whole theory of probabilistic choice behavior is not meant to apply to such a person. For under the pay-off conditions defined here, if $\pi_a > \pi_b$ the statistician should have at asymptote $p(a, b) = 1$. The point of (9) is rather to define a utility function which may be used to predict the actual behavior of all but the statistically sophisticated few. Numerous empirical studies (Mosteller and Nogee, 1951; Davidson *et al.*, 1957; Papandreou, 1957; Atkinson and Suppes, 1958; Davidson and Marschak, 1959) have clearly shown that naive subjects do not behave like mathematical statisticians. Experimental data on utility functions as defined by (9) for the two-arm bandit situation will be reported elsewhere.

The preceding analysis also has direct application to the first example of simple contingent reinforcement discussed in the preceding section. By replacing π_2 by $1 - \pi_2$, for purposes of symmetry, thus having as reinforcement probabilities $P(E_1 \mid A_1) = \pi_1$ and $P(E_2 \mid A_2) = \pi_2$, we may, obviously, get a utility function satisfying (4) by taking

$$u(a) = \log \frac{1}{1 - \pi_a}.$$

Further remarks on this case do not seem necessary.

The interesting question of generalization, it seems to me, is that of considering situations in which choice is made from one of n alternatives. In classical economic theory, the resolution of this choice problem is immediate: simply choose the most preferred item. But, as far as I know, with the notable exception of Luce (1959) there has been little if any analysis of stochastic choice behavior when the choice set has more than two alternatives. To describe this situation, let us use the notation $p(a, A)$ to mean the probability a is chosen in preference to any member of A, with the understanding that $\{a\} \cup A$ is the full choice set available, i.e., $p(a, A) + p(A, a) = 1$, where $p(A, a)$ means the probability an element of A is chosen in preference to a.[6] Beginning simply with $p(a, A)$, it is far from clear to me what axioms of rational behavior one might expect an organism to satisfy, in order to guarantee the existence of a utility

function. In fact, it is not completely obvious what should be the defining characteristic of a utility function. In analogy to (4) I suggest:

(10) $p(a, A) \geqslant p(b, B)$ if and only if $u(a) - u(A) \geqslant u(b) - u(B)$.

Condition (10) requires the utility of a *set* of alternatives to be defined, but it by no means implies that this set function need be additive, i.e., we need *not* have if A and B are disjoint sets that

$$u(A \cup B) = u(A) + u(B).$$

On the other hand, the intuitive interpretation of $p(a, A)$ suggests that if A is a subset of B then the utility of A is equal to or less than that of B, for in some sense the utility of A is the overall value weighting assigned to the set in deciding to choose a rather than any member of A. Also, it seems reasonable to require that if the utility of A is equal to or greater than that of B and a set C is added to both A and B, with C disjoint from both A and B, then the utility of $A \cup C$ is equal to or greater than that of $B \cup C$. These two principles may be summarized:

(11) if $A \subseteq B$ then $u(A) \leqslant u(B)$,

(12) if $A \cap C = B \cap C = 0$ and $u(A) \leqslant u(B)$ then $u(A \cup C) \leqslant u(B \cup C)$.

(Evidently (11) and (12) would not be acceptable if some of the alternatives had negative pay-offs, a possibility which we exclude here.)

What I now want to show is that for this multi-choice case a utility function satisfying (10), (11), and (12) may be derived from the axioms of stimulus-sampling theory by generalizing the approach to the two-arm bandit problem. For simplicity I shall again consider only the model with one stimulus element, although the results given here may easily be extended to a finite number of stimulus elements. The axioms given in the preceding section do need to be supplemented in one important respect, namely, we shall make Axiom C3 more definite by assuming that when a chosen response is not reinforced, the probability of the stimulus element becoming conditioned to some other response is uniformly distributed over the remaining set of available responses. Thus, in the notation of Section II, if there are n other available responses and total probability ε_i that the stimulus element will become conditioned to some other

response than A_i after A_i is not reinforced, then ε_i/n is the probability it will become conditioned to A_j, for $j \neq i$ and A_j in the available set. Keeping this notation in mind, it is easy to see that the transition matrix for $n+1$ possible responses (i.e., $n+1$ possible choices) has the following form:

$$
(13) \qquad \begin{array}{c|cccc}
 & A_1 & A_2 & \cdots & A_{n+1} \\
\hline
A_1 & 1-\varepsilon_1(1-\pi_1) & \frac{\varepsilon_1}{n}(1-\pi_1) & & \frac{\varepsilon_1}{n}(1-\pi_1) \\
A_2 & \frac{\varepsilon_2}{n}(1-\pi_2) & 1-\varepsilon_2(1-\pi_2) & \cdots & \frac{\varepsilon_2}{n}(1-\pi_2) \\
\vdots & \cdots & \cdots & \cdots & \cdots \\
A_{n+1} & \frac{\varepsilon_{n+1}}{n}(1-\pi_{n+1}) & \frac{\varepsilon_{n+1}}{n}(1-\pi_{n+1}) & \cdots & 1-\varepsilon_{n+1}(1-\pi_{n+1}).
\end{array}
$$

Following standard notation, let u_j be the asymptotic probability of response A_j. Then, as is well known, the asymptotic probabilities u_j may be obtained as the solution of the system of linear equations

$$
(14) \qquad \begin{cases} u_j = (1-\varepsilon_j(1-\pi_j))\,u_j + \displaystyle\sum_{i \neq j} \frac{\varepsilon_i(1-\pi_i)\,u_i}{n}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for} \quad j = 1, \ldots, n+1, \\ \sum u_j = 1, \end{cases}
$$

provided the matrix (13) satisfies certain regularity conditions, which are indeed satisfied here because every entry in the matrix is strictly positive.

It is not difficult to show that the solution of (14) is:

$$
(15) \qquad u_j = \frac{\prod\limits_{i \neq j} \varepsilon_i(1-\pi_i)}{\sum\limits_j \prod\limits_{i \neq j} \varepsilon_i(1-\pi_i)}.
$$

Now $p(a, A) = u_a$, and if we divide the numerator and denominator of the right-hand side of (15) by $\prod_{j \in X} \rho_j$, where as before $\rho_j = \varepsilon_j(1-\pi_j)$ and the set of alternatives is $X = A \cup \{a\}$, then

$$
(16) \qquad p(a, A) = \frac{1/\rho_a}{\sum\limits_{j \in X} 1/\rho_j}.
$$

On the basis of (16) we have a simple chain of equivalences like that leading from (7) to (8), which yields that $p(a, A) \geqslant p(b, B)$ if and only if

$$\log 1/\rho_a - \log \sum_{j \in A} 1/\rho_j \geqslant \log 1/\rho_b - \log \sum_{j \in B} 1/\rho_j, \tag{17}$$

and thus to satisfy (10), we define a utility function u for any nonempty finite set A of alternatives as:

$$u(A) = \log \sum_{j \in A} 1/\rho_j. \tag{18}$$

Moreover, we may use (18) to generalize (10) immediately to the probabilities $p(A, \tilde{A})$, where $\tilde{A}$ is the complement of the set A with respect to the total set of alternatives, i.e., $A \cup \tilde{A} = X$. The interpretation of $p(A, \tilde{A})$ is that this is the probability of choosing an alternative from A rather than from its complement $\tilde{A}$. We observe first that (16) yields:

$$p(A, \tilde{A}) = \frac{\sum_A 1/\rho_j}{\sum_A 1/\rho_j + \sum_{\tilde{A}} 1/\rho_j}. \tag{19}$$

Manipulations similar to those already carried out then result in:

$$\begin{aligned} &p(A, \tilde{A}) \geqslant p(B, \tilde{B}) \quad \text{if and only if} \\ &u(A) - u(\tilde{A}) \geqslant u(B) - u(\tilde{B}). \end{aligned} \tag{20}$$

It is easily verified that the utility function u defined by (18) satisfies (11) and (12) as well as (10) and (20). If u were also an additive set function it would be more appropriate to call it a subjective probability function. It seems to me that its logarithmic rather than additive character is intuitively sound. In particular, the marginal utility of adding another alternative to a set of such is appropriately a decreasing function of the size of the set. In other words, the utility function defined by (18) has the classical property that as wealth increases each additional unit has decreasing marginal utility.

III. RELATIONS TO OTHER THEORIES

To begin with, I want to show that the entropy of any set of alternatives X, probability distribution p, and partition $\prod$ of X is a negative linear transformation of the expected utility of $(X, p, \prod)$.[7] Following the well-

known work of Shannon (see, e.g., Shannon and Weaver, 1949) on the theory of information, the entropy H of $(X, p, \prod)$ is defined as:

$$(20) \qquad H(\prod) = -\sum_{A \in \Pi} p(A, \tilde{A}) \log_2 p(A, \tilde{A}).$$

And the expected utility $\mathscr{E}(u, \prod)$ is defined in the standard manner as:

$$(21) \qquad \mathscr{E}(u, \prod) = \sum_{A \in \Pi} p(A, \tilde{A})\, u(A).$$

Now

$$u(A) = \log \sum_{A} 1/\rho_j = \log \frac{\sum_{A} 1/\rho_j}{\sum_{X} 1/\rho_j} + \log \sum_{X} 1/\rho_j$$

$$= \alpha \log_2 p(A, \tilde{A}) + \beta,$$

where $\alpha = \log 2$ and $\beta = \log \sum_X 1/\rho_j$, and it is clear α and β are both independent of $\prod$.[8]

Substituting this last result for $u(A)$ into (21) we have

$$\mathscr{E}(u, \prod) = \sum_{A \in \Pi} p(A, \tilde{A}) [\alpha \log_2 p(A, \tilde{A}) + \beta]$$

$$= -\alpha H(\prod) + \beta,$$

the desired conclusion. It is to be noticed that the finest partition of X maximizes entropy, whereas the coarsest one maximizes expected utility (with respect to the set of all partitions of X).

I now turn to consideration of Luce's choice axiom (1959, p. 6) which we may formulate as follows: if $A \subseteq B \subseteq X$ then

$$(22) \qquad p_X(A) = p_B(A)\, p_X(B),$$

where $p_X(A)$ is the probability that an element of A is selected from the total choice set X. Thus if $A \cup \tilde{A} = X$, then in the notation used earlier, $p_X(A) = p(A, \tilde{A})$. The purpose of the subscript usage is to indicate an explicit change in the total set of available alternatives.

Without further assumptions (22) cannot be derived from the postulates for learning theory given at the beginning, because they include no assertions about the constancy or continuity of behavior when the number of available responses is changed. To derive (22), however, we need add only the postulate that the conditioning parameter ε_i of response A_i for every i is independent of what subset of the alternatives X is available.

Granted this additional assumption about conditioning, derivation of Luce's axiom is a simple matter, for

$$p_X(A) = \frac{\sum_A 1/\rho_j}{\sum_X 1/\rho_j} = \frac{\sum_A 1/\rho_j}{\sum_B 1/\rho_j} \cdot \frac{\sum_B 1/\rho_j}{\sum_X 1/\rho_j} = p_B(A)\, p_X(B).$$

Using his choice axiom Luce proves the existence of a ratio scale $v(j)$ (1959, pp. 20–28) with the property that

$$p_X(A) = \frac{\sum_A v(j)}{\sum_X v(j)}.$$

The relation of this additive ratio scale to the utility function u defined by (18) is simply

$$v(A) = ke^{u(A)},$$

where k is a positive real number.

NOTES

[1] I have benefited from conversations with several people on the topic of this paper, but most particularly from those with Donald Davidson, William K. Estes, and R. Duncan Luce.

[2†] Article 8 in this volume.

[3†] Article 4 in this volume.

[4] Under a rather natural continuity assumption, which is however stronger than U4–U6, Debreu (1958) has shown that the quadruple condition (U3) is necessary and sufficient for the existence of a utility function satisfying (4). Of course, granted U4–U6, and the "technical axioms" U1 and U2, it is obvious that the quadruple condition is also necessary and sufficient in this context. It may also be remarked that to give necessary and sufficient conditions on the set A and the function p, without continuity or finiteness restrictions, is the extremely difficult mathematical problem of classifying all isomorphism types representable by a real-valued function u satisfying (4).

[5] I assume throughout that $0 < \pi_a, \varepsilon_a < 1$, for every a in A.

[6] From this point on, X rather than A will represent the total set of available alternatives.

[7] A *partition* of a set X is a family of nonempty, pairwise disjoint subsets of X such that the union of all sets in the family is X.

[8] When no base of a logarithm is indicated, it is understood to be e.

10. SOME FORMAL MODELS OF GRADING PRINCIPLES*[1]

I. INTRODUCTION

The present paper offers an analysis of grading principles from the viewpoint of statistical decision theory and game theory. The mistaken notion is widely held that the plain man is really clear about practical ethical and moral issues and that philosophers need only tidy up certain wayward corners of the subject.[2] Personally I find difficult the problem of devising any general ethical rules of behavior for simple two-person games; the ethical complexities of progressive taxation, tariff barriers, or treatment of sexual psychopaths are beyond any exact conceptual analysis. That decisions are and must be made about these issues no more proves that their ethical aspects are completely understood than does the fact that the Romans built bridges prove that they had any quantitative grasp of the mechanical theory of stress.

It is pertinent to remark that the first model used in this paper is at the basis of much recent foundational work in statistics (see Blackwell and Girshick, 1954; Savage, 1954). The considerations in the last two sections are within the more general framework of the theory of games as developed by von Neumann and others. My particular concern is the embedding in this framework of a theory of two-person justice.

II. INDIVIDUAL DECISION MODEL

The structure of the first model to be considered is simple. We shall call an ordered triple $\mathscr{S} = \langle S, C, D \rangle$ an *individual decision* situation when S and C are sets and D is a set of functions mapping S into C. The intended interpretation is:

S = set of states of nature,
C = set of consequences,
D = set of decisions or actions.

* Reprinted from *Synthese* **16** (1966), 284–306.

Since the terms 'states of nature', 'consequences', 'decisions' and 'actions' are used here in a somewhat special manner, an example may help to make clearer their intended meaning.

Example 1: Suppose I come home and find a bottle of ink spilt on the rug, and also suppose I know immediately that it could have been spilt either by my four-year-old daughter or by my cat. These two possibilities correspond to the two states of nature. I can take one of two actions, let us say: spank the child or do not spank the child. And the possible consequences are four in number, as illustrated in Table I. The rows correspond to the two states of nature, the columns to the two actions, and the entries in the table to possible consequences.

TABLE I

actions / states of nature	a_1 – spank the child	a_2 – do not spank the child
s_1 – child spilt the ink	c_1 – ink spilt by child and child spanked	c_2 – ink spilt by child and child not spanked
s_2 – cat spilt the ink	c_3 – ink spilt by cat and child spanked	c_4 – ink spilt by cat and child not spanked

Since the term 'states of nature' is not much used in philosophy there should be little objection to its special use here; the term 'action' is used in a way that is consonant with at least one of its major uses in ordinary contexts. But my use of 'consequence' is probably at variance with its primary use in the writings of moral philosophers. The consequence c_1 above, for instance, ink spilt by child and child spanked, would be regarded by many as the bare beginning of consequences. It is to avoid exactly the vagueness of the consequences flowing from c_1, c_2, c_3 or c_4, that I have adopted the restricted use. The longer term 'immediate consequence' could be used. Yet in ordinary usage there is much to defend the use adopted here. When a quarterback throws an intercepted pass in the last two minutes of play it might be appropriate to remark "The

consequence of that is obvious. We lose the game." It would seem pedantic to insist on saying "The *immediate* consequence of that is obvious. We lose the game." And it would be a classroom gambit to object that the use of the definite article is wrong, because the action could have other important consequences for the quarterback: he quarrels with his girl that night, the coach decides not to start him in the next game.

Apart from any questions of ordinary usage there is a technical device which may be used to meet the difficulty that it is almost always impossible to characterize the full set of consequences which may flow from an action. Given an individual decision situation $\langle S, C, D \rangle$, let C' be the set of all consequences which result from some state of nature in S and some decision in D. Then C is a *partition* of C', that is C is a family of non-empty pairwise disjoint sets whose union is C'. In this analysis, each c_i in our example is a set of consequences. It is practically impossible to say exactly what the members of c_1, say, are, but in rough terms they are the possible consequences, proximate and remote, which would wholly or in part result from the immediate consequence of the ink's being spilt by the child and the child's being spanked.[3]

The still more complicated question of what kind of language is appropriate for describing either consequences or states of nature cannot be examined here. Certainly in most situations it is difficult to avoid evaluative or normative terms, but the use of non-factual language does not directly disturb or vitiate the analysis given here.

One of the basic problems of statistical decision theory is to introduce a preference ordering on the set of decisions of an individual decision situation and to consider what postulates the preference ordering of a reasonable man should satisfy. (For such an analysis see Savage, 1954 or Suppes, 1956b.[4†]) The notion of reasonableness or rationality used here is an informal, intuitive one, and its application in defense of any particular postulate consists of analyzing particular examples. The problem is presumed solved if reasonable postulates can be found which are strong enough to guarantee the existence of a (subjective) probability measure on the states of nature and a utility function on the set of consequences such that one decision is to be preferred to another if and only if the expected utility of the first decision with respect to the probability measure is greater than that of the second. Once such a probability measure and utility function are constructed no further principles of action are needed.

The sole maxim to be followed by the rational man is: maximize expected utility.

Historically the idea of maximizing utility is closely connected with the hedonistic ideas of Bentham, Mill, Sidgwick, and their followers. However, it is an unequivocal mistake to think that the maxim: maximize expected utility, in any respect involves a commitment to hedonism. As I hope to make clear in the sequel, if the utility function on consequences were guided by an ethic of duty rather than pleasure, it would still be good advice to maximize expected utility. In this case a calculus of duty would replace a calculus of pleasure. To my mind the most important aspect of the hedonistic tradition in ethics has been the clear recognition that *some* principle of calculation is required for rational action in the face of other than trivial situations. The main point of this paper is to defend a thesis as to how grading principles should enter into these calculations.

Before developing these ideas further I want to say something about a major criticism that is usually made of the general maximization viewpoint adopted here. To wit, as one philosopher scornfully put it to me, whoever heard of a man making such calculations prior to making any actual decision. Naturally this philosopher had in mind the "ordinary" man in "ordinary" situations like that of buying a pint of whiskey or selecting a new tobacco. One might as well reject a whole discipline such as the physical theory of the strength of materials by remarking that no carpenter computes the load capacity of a joist before sawing and nailing it. There are situations where elaborate calculations are made in order to maximize utility; the new disciplines of management science and operations research provide numerous examples.[5] Moreover, I maintain that in many ordinary situations it is not the impossibility of detailed calculation that is relevant but rather the superfluity of it. For instance, in the simple situation schematized by Table I, if it is definitely known that the ink was spilt by the child and not the cat then to take appropriate action I need only order in preference two consequences: c_1 and c_2, according to my principles of childrearing. I need no numerical utility function. And this situation is characteristic: whenever uncertainty regarding the true state of nature is eliminated, the pertinence of a numerical utility function disappears, and the principle of maximizing expected utility assumes a very simple form: choose that action whose consequence is most preferred (for reasons of pleasure, duty, justice, or what have you).

III. DEFINITION OF GRADING PRINCIPLES

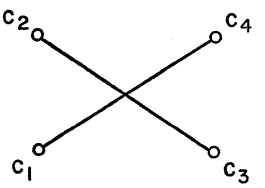
Traditionally in ethics, actions are said to be right or wrong, and consequences good or bad. If we carried over this distinction to individual decision situations then we would need moral principles of grading governing acts and value principles of grading arranging consequences in order of preference. But I am proposing here that the one controlling moral principle of action is the maxim: maximize expected utility, hereafter referred to as the M.E.U. maxim. On this view it is a mistake to hold that grading principles aid us directly in distinguishing between the quality of acts. The function of grading principles is rather to aid the individual in constructing his preference relation on the set of consequences.[6] There is a simple reason why this position is not in conflict with most of the standard examples purporting to show how grading principles should regulate actions; namely, if the state of nature is known, there is an effective one-one correspondence between the set D of acts and the set C of consequences, and any relation on C defines a corresponding relation on D. This point is further amplified below.

To put it baldly then, I am claiming that the proper logical status of a grading principle in an individual decision situation is as a binary relation on the set C of consequences, in fact, an asymmetric, transitive relation on C, i.e., a strict partial ordering of C.

DEFINITION 1: *Let $\mathscr{S} = \langle S, C, D \rangle$ be an individual decision situation. Then a grading principle with respect to $\mathscr{S}$ is a strict partial ordering of C.*

I have insisted that a grading principle have at least the properties of a strict partial ordering, for otherwise it would scarcely be a guide to fixing the preference relation.

Example 2: *A principle of childrearing.* Referring to Example 1, a tenable grading principle held by some modern parents is: never punish a child. This leads to the following strict partial ordering of C[7], which we may represent by a Hasse diagram:



It should be noted that this principle of childrearing is sufficient to determine action although the set of consequences is not completely ordered by it. For, whatever the true state of nature, the consequence of taking action a_2 is preferred to the consequence of taking action a_1, that is, c_2 is preferred to c_1 and c_4 is preferred to c_3.

As the following example drawn from welfare economics shows, most grading principles are not sufficient to determine action.

Example 3: *Principle of unanimity.* Suppose that the decision situation consists of an arbitrary set S of states of nature, and C is a set of ordered n-tuples (n-dimensional vectors) representing the distribution of some desired commodity to a group of n individuals. Administrator A, a member of the group, is to decide in a just manner which distribution vector is to be used in allotting the quantities of the commodity. The grading principle of unanimity asserts that vector $x=\langle x_1, \ldots, x_n\rangle$ is to be preferred to vector $y=\langle y_1, \ldots, y_n\rangle$ if for every $i=1, \ldots, n, x_i \geqslant y_i$ and for some $i, x_i > y_i$. This principle, also known as the principle of efficiency or Pareto optimality, is a very weak grading principle and surely any administrator who did not satisfy it would be stoned out of office.[8] It is obvious that in general the principle of unanimity does not uniquely determine the optimal action even when there is only one state of nature.

More troublesome, at least from a psychological standpoint, is the decision situation in which two grading principles are in conflict. This state of affairs is reflected formally in our model by the fact that the union of two strict partial orderings is not always a strict partial ordering.

DEFINITION 2: *Let* $\mathscr{S}=\langle S, C, D\rangle$ *be an individual situation, and let* G_1 *and* G_2 *be grading principles with respect to* $\mathscr{S}$. *Then* G_1 *and* G_2 *are compatible if, and only if,* $G_1 \cup G_2$ *is a grading principle with respect to* $\mathscr{S}$.[9]

Two simple conditions with reasonable interpretations which will insure compatibility of grading principles are the following.

THEOREM 1: *If* (i) G_1 *is a subrelation of* G_2 *or* G_2 *is a subrelation of* G_1, *or* (ii) *if the fields of* G_1 *and* G_2 *are mutually exclusive, then* G_1 *and* G_2 *are compatible.*

The proof of this theorem is trivial. Some examples illustrating it may be drawn from welfare economics, where S and C are defined as in Example 3.

Example 4: Let

G_1 = principle of unanimity,
G_2 = principle of gross aggregation,
G_3 = principle of social weights $a = \langle a_1, \ldots, a_n \rangle$,

where

$$x \; G_2 \; y \quad \text{if, and only if,} \quad \sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i$$

and

$$x \; G_3 \; y \quad \text{if, and only if,} \quad \sum_{i=1}^{n} a_i x_i > \sum_{i=1}^{n} a_i y_i .$$

Principle G_1 has already been discussed; Principle G_2 says that one distribution x is to be preferred to another y if x results in a greater total quantity of the commodity for the social group; Principle G_3 corresponds to the assignment of weights to each individual by Administrator A; presumably A would use some further principle of need or merit to aid in determining the weights.[10] As application of Theorem 1, we have that G_1 and G_2 are compatible, since G_1 is a subrelation of G_2, that is, if $x \; G_1 \; y$ then $x \; G_2 \; y$. To see this, we observe that if $x \; G_1 \; y$ then

(1) $\quad x_i \geqslant y_i \quad$ for all i
$\quad x_i > y_i \quad$ for some i,

whence

$$\sum x_i > \sum y_i .$$

Moreover, if each individual is given a strictly positive weight, that is, $a_i > 0$ for all i, then G_1 is a subrelation of G_3, and hence compatible with it. The reasoning is obvious. From (1) and the hypothesis that $a_i > 0$ we have:

$$a_i x_i \geqslant a_i y_i \quad \text{for all } i$$
$$a_i x_i > a_i y_i \quad \text{for some } i,$$

whence by addition of inequalities

$$\sum a_i x_i > \sum a_i y_i .$$

On the other hand, when C has any abundance of different distribution vectors, G_2 and G_3 are incompatible. For instance, let $n = 3$ and

$$x = \langle 1, 2, 4 \rangle$$
$$y = \langle 4, 1, 1 \rangle$$
$$a = \langle 2, \tfrac{1}{2}, \tfrac{1}{4} \rangle .$$

Then

$$x\ G_2\ y$$

since

$$\sum x_i = 7 \quad \text{and} \quad \sum y_i = 6,$$

but

$$y\ G_3\ x$$

since

$$\sum a_i x_i = 4 \quad \text{and} \quad \sum a_i y_i = 8\tfrac{3}{4}.$$

Examples which satisfy (ii) of Theorem 1 are easy to construct but will not be considered here. The intuitive idea of (ii) is the truism that grading principles concerned with entirely different spheres of activity are compatible.

The use of the word 'activity' in the last sentence underlines the difficulty of not speaking of grading principles as referring to acts or decisions rather than consequences. Before turning to social decision situations in the next section, something more needs to be said about the status here advocated for grading principles. One natural tendency is to formulate grading principles in the imperative mood so as to command the execution of certain acts. But Hare (1952, Part III) has cogently argued it is more appropriate to use the indicative mood and the auxiliary verb 'ought' to obtain the proper sort of universal formulation. The one further emendation required here is to add the infinitive 'to prefer' after 'ought'. Thus, we go from the imperative:

"Honor thy father"

to:

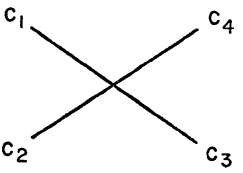
"Everyone ought to honor his father",

and on to:

(1) "Everyone ought to prefer to honor his father."

I maintain that ordinary usage addresses moral principles of grading directly to acts because the problem of acting without knowing the true state of nature is ignored. This point is important enough to be amplified by referring again to Example 1. Consider the moral imperative 'Punish the guilty and defend the innocent'. Suppose this is the only moral

imperative guiding my choice of action a_1 or a_2 in Example 1. It seems patently obvious that without knowing the true state of nature I can make no direct application of the imperative to choose between a_1 and a_2. If s_1 is the true state of nature I should choose a_1, but if s_2 is the true state, I should choose a_2. To be sure, I could first sum up the factual evidence for s_1 and s_2, decide which is more likely, assume the more likely state is nearly certain to be the true state, and then take the appropriate action. But this is surely a crude way to proceed and is wholly inadequate in more complicated situations; for instance, suppose there were three states of nature to each of which I assigned a subjective probability of $\frac{1}{3}$. On the other hand, the imperative may be applied directly to constrain my preference relation on the set $\{c_1, c_2, c_3, c_4\}$ of consequences. The Hasse diagram of the resulting strict partial ordering is obviously:



which may be compared with the diagram for Example 2. When applied directly to consequences, application of the imperative need not be confounded with the difficult and distinct problem of weighing factual evidence regarding the true state of nature.

The particular homily about honoring fathers illustrates another point: it and all principles of a similar form lead to a simple and crude partial ordering of consequences, namely, consequences are divided into two classes and all members of one are preferred to all members of the other. As Examples 3 and 4 emphasize, such principles are not of much help in making a rational decision in a complicated situation like that generated by a labor-management dispute or the problem of pricing policy in a semi-controlled economy.

IV. SOCIAL DECISION MODEL

Against the analysis of previous sections may be brought the charge that the indivual decision model unduly and unrealistically isolates the

behavior of one man from another. In the remainder of this paper social situations shall be considered. For reasons of technical simplicity the discussion shall be restricted to two persons, although most of the concepts introduced readily generalize to n persons.

The structure of the basic model is still relatively simple. We shall call an ordered sextuple $\mathscr{S} = \langle S, C_1, C_2, D_1, D_2, f \rangle$ a *two-person decision situation* when S, C_1, C_2, D_1, D_2 are sets and f is a function mapping the Cartesian product $S \times D_1 \times D_2$ into $C_1 \times C_2$. The intended interpretation is:

S = set of states of nature,
C_1 = set of consequences for person I,
C_2 = set of consequences for person II,
D_1 = set of decisions or acts available to I,
D_2 = set of decisions or acts available to II,
f = social decision function.

Some examples will be given in the next section in connection with the theory of two-person justice.

The definition of grading principles is an obvious generalization of the one already given for the individual case.

DEFINITION 3: *Let $\mathscr{S} = \langle S, C_1, C_2, D_1, D_2, f \rangle$ be a two-person decision situation. Then a grading principle with respect to $\mathscr{S}$ is a strict partial ordering of the Cartesian product $C_1 \times C_2$.*

This definition does not require that in applying a grading principle person I need consider consequences to person II, but does make possible such a consideration. We could in fact use Definition 3 as a basis for defining a wholly egocentric person, namely a person, say I, whose grading principles and preference relations in all two-person situations are orderings uniquely determined by elements of C_1 (the character of C_2 being never considered).

The same arguments given previously apply to the requirement that ordinary grading principles in two-person situations be partial orderings on consequences and not on acts. On the other hand, the arguments for a rigid adherence to the M.E.U. maxim are not so persuasive, since other rules of behavior like minimax or minimax regret can be strongly defended for two-person situations. But these matters will not be gone into here; for our purposes adoption of any of these alternative rules requires admission only that a utility or value function on $C_1 \times C_2$ is needed for

both persons I and II. We want to investigate how a formal principle of justice may be introduced which will put non-trivial constraints on the utility function. Moreover, it will be of interest to investigate the adequacy of a justice maxim compared to the M.E.U. or minimax kind of maxim, as an over-all rule of behavior.

In concentrating attention on justice no claim is intended that it is the most significant grading principle for social situations, nor even that the definitions given here provide more than the merest beginning of a formal theory of justice.

To begin with we need the notion of a preference relation on $C_1 \cup C_2$, that is, on the set of consequences to both I and II. The idea is that one consequence in $C_1 \times C_2$ will be deemed more just or fair than another relative to a preference ranking of all consequences together. How in fact would a person make such a ranking? Presumably by treating himself and the other person on an "equal" basis. A suggestion as to how this idea of equality or symmetry may be formalized will be given the following definition:

DEFINITION 4: *A system* $\mathscr{S} = \langle S, C_1, C_2, D_1, D_2, f, R_1, R_2 \rangle$ *is a two-person decision situation with preference rankings if, and only if,* $\langle S, C_1, C_2, D_1, D_2, f \rangle$ *is a two-person decision situation, and* R_1 *and* R_2 *are weak orderings of* $C_1 \cup C_2$. (A weak ordering is a relation which is transitive and strongly connected.)

The intended interpretation is that R_1 is the preference ranking of person I and R_2 that of II. Formally we might say that a person's preference ranking R of $C_1 \cup C_2$ is equitable or symmetric if it remains unchanged when the two persons change positions in the decision situation. Difficulties of making this suggestion precise will not be pursued here, but it would seem best to do it in terms of a specific game, or at least gamelike, structure, with the exchange being defined in terms of becoming a different player in the game, not, by all means, in terms of the personal attributes of the players somehow being exchanged. It is intended that in constructing R_1, say, person I will say to himself, it is better that II have x in C_2 than that I have y in C_1 whence $x \, R_1 \, y$ and not $y \, R_1 \, x$, etc. For example, a man should judge it better that his neighbor of equal economic status receive a thousand dollars than that he himself should receive fifty dollars. Unfortunately, I see no way of characterizing in an adequate formal manner the intuitive notion of *better than* used in this example.

But it would be a mistake to consider this situation peculiar to moral philosophy. The notion of preference or *better than* has a status in formal moral philosophy very similar to that of the notion of force in mechanics. It is not a problem of mechanics proper to classify forces according to their physical origin.[11]

We now define the notion of *more just than* (abbreviated by J) relative to each person's preference ranking.

DEFINITION 5: *If* $x_1, y_1 \in C_1$ *and* $x_2, y_2 \in C_2$ *and* $x = \langle x_1, x_2 \rangle$ *and* $y = \langle y_1, y_2 \rangle$ *then for* $i = 1, 2$, $x J_i y$ *if, and only if, either* (i) $x_1 R_i y_1$ *and* $x_2 R_i y_2$ *and not* ($y_1 R_i x_1$ *and* $y_2 R_i x_2$), *or* (ii) $x_1 R_i y_2$ *and* $x_2 R_i y_1$ *and not* ($y_2 R_i x_1$ *and* $y_1 R_i x_2$).

This definition is simpler than it may appear at first glance. It is framed so as to make J_i (for $i = 1, 2$) a strict partial ordering of the Cartesian product $C_1 \times C_2$, and yet permits the comparison of elements of C_1 with C_2. The two 'not' clauses in the definition guarantee that J_i is asymmetric.

Examples of J_i are at the beginning of the next section. We conclude this section with the theorem:

THEOREM 2: *Both* J_1 *and* J_2 *are grading principles with respect to* $\mathscr{S}$.

Proof: For $i = 1, 2$, to prove that J_i is asymmetric, suppose by way of contradiction that for some $x = \langle x_1, x_2 \rangle$ and $y = \langle y_1, y_2 \rangle$ in $C_1 \times C_2$ that

$$x J_i y \quad \text{and} \quad y J_i x .$$

From $x J_i y$ it follows from the definition that (dropping subscript i on R for brevity) $x_1 R y_1$ or $x_1 R y_2$, and similarly from $y J_i x$ it follows that $y_1 R x_1$ or $y_1 R x_2$. We thus have four cases to consider:

Case 1: $x_1 R y_1$ and $y_1 R x_1$. Case 3: $x_1 R y_2$ and $y_1 R x_1$.
Case 2: $x_1 R y_1$ and $y_1 R x_2$. Case 4: $x_1 R y_2$ and $y_1 R x_2$.

Since the proofs for all cases are similar, we shall look only at Case 2 in detail. From the hypothesis of this case, we have from (i) of the definition:

(1) $x_1 R y_1$,

(2) $x_2 R y_2$,

(3) not $y_1 R x_1$ or not $y_2 R x_2$,

and from (ii):

(4) $y_1 \, R \, x_2$

(5) $y_2 \, R \, x_1$

(6) not $x_2 \, R \, y_1$ or not $x_1 \, R \, y_2$.

From (1), (4) and the transitivity of R we infer:

(7) $x_1 \, R \, x_2$

and from (7) and (2):

(8) $x_1 \, R \, y_2$.

From (2) and (5) (and transitivity of R):

(9) $x_2 \, R \, x_1$,

and from (9) and (1):

(10) $x_2 \, R \, y_1$,

but (8) and (10) contradict (6).

To prove now that J_i is transitive, we assume

$$x \, J_i \, y \quad \text{and} \quad y \, J_i \, z,$$

which leads to four cases also. Again we shall consider only one typical case:

(11) $x_1 \, R \, y_2$ and $y_1 \, R \, z_2$.

From (11) and (ii) of the definition, we have:

(12) $x_2 \, R \, y_1$

(13) $y_2 \, R \, z_1$

(14) not $y_2 \, R \, x_1$ or not $y_1 \, R \, x_2$

(15) not $z_2 \, R \, y_1$ or not $z_1 \, R \, y_2$.

From (11), (13) and transitivity of R, we get:

(16) $x_1 \, R \, z_1$,

and similarly from (11) and (12):

(17) $x_2 \, R \, z_2$.

It remains to show that not $z_1 R x_1$ or not $z_2 R x_2$. Suppose by way of contradiction that

(18) $z_1 R x_1$ and $z_2 R x_2$.

From (18) and (11), we have:

(19) $z_1 R y_2$,

and from (18) and (12)

(20) $z_2 R y_1$,

but (19) and (20) contradict (15), which completes our proof, since for J_i to be a grading principle with respect to $\mathscr{S}$, it is required by definition that J_i be asymmetric and transitive in $C_1 \times C_2$.

V. POINTS OF JUSTICE AND THE PRISONER'S DILEMMA

It will be instructive to apply the ideas introduced in the last section to a simple but conceptually troublesome example of a two-person, non-zero-sum, non-cooperative game known as the prisoner's dilemma.[12] We quote the description from Chapter 5 of Luce and Raiffa (1957):

> Two suspects are taken into custody and separated. The district attorney is certain that they are guilty of a specific crime but he does not have adequate evidence to convict them at a trial. He points out to each prisoner that each has two alternatives: to confess to the crime the police are sure they have done, or not to confess. If they both do not confess, then the district attorney states he will book them on some very minor trumped-up charge such as vagrancy and they will both receive minor punishment; if they both confess they will be prosecuted, but he will recommend less than the most severe sentence; but if one confesses and the other does not, then the confessor will go free while the latter will get "the book" slapped at him.

Let n = no conviction on any charge,
v = vagrancy conviction,
r = reduced conviction (less than maximum),
m = maximum conviction.

Then the game may be represented by:

I \ II	confess	not confess
confess	$\langle r, r \rangle$	$\langle n, m \rangle$
not confess	$\langle m, n \rangle$	$\langle v, v \rangle$

where a pair like $\langle n, m\rangle$ is interpreted so that the first member n is the outcome to person I and the second member m the outcome to person II. We have not distinguished n_{I} and n_{II}, m_{I} and m_{II}, etc. These consequences are treated the same for each player. Keeping this in mind, the complete two-person decision situation with preference rankings may be identified, provided we introduce the one obvious preference ranking on the set of consequences:

S = one element set (trivial here),
$C_1 = C_2 = \{m, n, r, v\}$[13],
$D_1 = D_2 =$ {confess, not confess},
f = function defined by above game matrix,
$R_1 = R_2 =$ weak ordering arising from linear ordering n, v, r, m, with n most preferred.

Clearly here

$$J_1 = J_2,$$

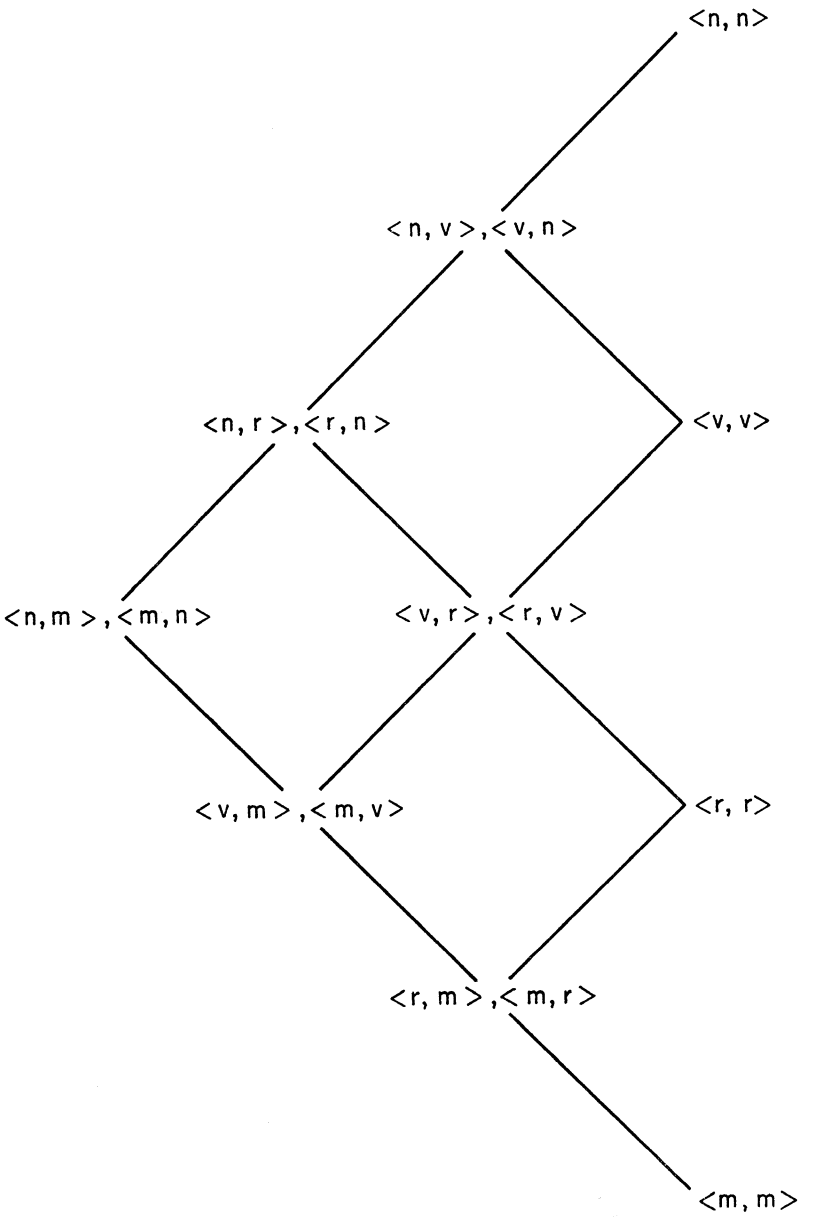
and the ordering *more just than* of $C_1 \times C_2$ may be represented by the following Hasse diagram (see next page), where two elements of $C_1 \times C_2$ standing at the same point in the diagram are not comparable under J_i.[14] Of course, only the four elements in the game matrix are of direct concern in discussing the prisoner's dilemma. The ordering induced by J_i on them may be represented by:

(1)

The important thing is that $\langle n, m\rangle$ is not related by J_i to any of the other three elements, nor is $\langle m, n\rangle$.

The weak relation expressed by (1) would not seem to be of much help in guiding the choice of an action or strategy by either prisoner. As a direct constraint on the utility function of either it scarcely imposes any structure. Before attempting to show that considerably more can be obtained from (1) by introducing the concept of a *point of justice*, it will be useful briefly to review the game-theoretic solution of the prisoner's dilemma.

Two concepts of optimality for two-person, non-zero-sum, non-


<n, n>
<n, v>, <v, n>
<n, r>, <r, n>
<v, v>
<n, m>, <m, n>
<v, r>, <r, v>
<v, m>, <m, v>
<r, r>
<r, m>, <m, r>
<m, m>

cooperative games yield the conclusion that both prisoners should choose the strategy of confessing, which leads to the outcome or consequence $\langle r, r \rangle$. One concept arises from the highly appealing *sure-thing* principle. A strategy or decision satisfies the sure-thing principle if no matter what your opponent does you are at least as well off, and possibly better off, with this strategy in comparison to any other available to you. Thus if person I adheres to the sure-thing principle he should confess, for if II confesses I gets r rather than m, and if II does not confess I gets n rather than v; whence for every choice of II, I is better off confessing. A similar situation obtains for II.

In many games no strategy satisfies the sure-thing principle. But every finite game of the class being discussed does have at least one *equilibrium point*, the second concept of optimality (introduced by Nash, 1950, 1951). Roughly speaking, an equilibrium point is a set of strategies, one for each player, with the property that these strategies provide a way of playing the game such that if all the players but one follow their given strategies, the remaining player cannot do better by following any strategy other than the one belonging to the equilibrium point. As is easily verified, the unique equilibrium point for the prisoner's dilemma is the pair of confession strategies, the same result obtained by application of the sure-thing principle.

In spite of the weight of these optimality principles there are several unsatisfactory aspects of the recommended solution. If both prisoners completely trust each other it seems more reasonable for both of them to adopt the strategy of not confessing. Moreover, the act of confessing might from a moral standpoint be distasteful. The various game-theoretical principles of behavior like the two just discussed are aimed at satisfying intuitive ideas of prudential rather than moral behavior – the notion of prudence being that of acting in one's own best interest without direct concern for others. The point of the remainder of this paper is to contrast moral and prudential behavior, with special reference to the prisoner's dilemma.

In Section III, I have argued that grading principles should be addressed to consequences rather than decisions or acts. I now want to suggest that a (first-order) grading principle concerned with consequences may lead to a (second-order) moral principle which is a direct guide to action. Such second-order moral principles may be termed *ethical rules behavior*, in *of*

contrast to game-theoretical *prudential* principles of behavior. I shall use the justice relation on consequences to formulate one such ethical rule of behavior. First we define a (J_i) *admissible element* as an element of $C_1 \times C_2$ which is not dominated under the relation J_i by any other element. In diagram (1) elements $\langle v, v \rangle$, $\langle n, m \rangle$ and $\langle m, n \rangle$ are (J_i) admissible. In the preceding diagram only $\langle n, n \rangle$ is. Next, in analogy with the definition of an equilibrium point, let us define a (J_i) *point of justice* as a set of strategies, one for each player, such that adoption of these strategies leads to an admissible element as outcome.

The simplest justice-oriented rule of behavior is then:

(I) *If $J_1 = J_2$ and there is a unique point of justice, the strategy belonging to this point ought to be chosen.*

Unfortunately, (I) is not applicable to the prisoner's dilemma, for the requirement that there be a unique point of justice is not satisfied.[15]

A more complicated, but still relatively simple ethical rule of behavior may be introduced in terms of the notion of a *justice-saturated* strategy. A strategy for player i is justice-saturated (with respect to J_i) if whatever strategies are picked by the other players the resulting set of strategies is a (J_i) point of justice. The rule of behavior is then:

(II) *If for any player this set of justice-saturated strategies is non-empty, he ought to choose one.*

In the prisoner's dilemma each player has a unique justice-saturated strategy, namely, the strategy of not confessing, joint use of which leads to the reasonable outcome $\langle v, v \rangle$.

To be sure, when a person's set of justice-saturated strategies contains more than one element, (II) does not lead to a unique action, and some supplementary ethical rule of behavior may be needed. A similar problem arises for prudential game-theoretical rules of behavior and should surprise only those who believe that satisfactory categorical rules of action are easily come by.

If neither (I) nor (II) is applicable (and simple two-person decision situations exist for which this is the case), the theory of justice outlined here is of no use in determining what action to take, except insofar as the relation J_i is a constraint on the person's utility function.[16]

But this last problem of applicability is one of the least difficulties

that face an adequate formal theory of justice. For example, even when (II) is applicable, an "ethical" man using it may be at a definite competitive disadvantage against a "prudential" man. In the prisoner's dilemma if prisoner I adopts his justice-saturated strategy and prisoner II his equilibrium-point strategy, then prisoner I will receive the maximum conviction. It is not easy, for me at least, to decide if this is an intuitive argument against the formal theory of justice or fair play set forth here, or if it is an intuitively reasonable instance of a just or fair man getting the worst of a situation. If the latter is the case, I think it may be claimed that a man who in all situations acts according to ethical rules of behavior may fare as well in the long run as the purely prudential man, provided knowledge of his standards of actions are known to his fellow man.

Another difficulty with the present theory is its structural weakness. It is a priori certain that no very elaborate theory of action can be built on the simple notion of a strict partial ordering. A major step in the development of rational theories of behavior has been the quantification of value (i.e., utility) and of subjective probability (i.e., reasonable degree of belief). Plausible assumptions which will lead to quantification of the theory of justice seem hard to find.

Making the theory of justice depend on the individual preference rankings is very much in the spirit of modern welfare economics, but may seem highly unsatisfactory to many philosophers. And I think it may be rightly objected that the intuitive success of the theory depends upon these individual preference rankings themselves satisfying certain criteria of justice. To admit this objection is not to accede to a charge of circularity, for moral principles of justice, logically independent of the theory developed here, can be consistently introduced as constraints on individual preference rankings of $C_1 \cup C_2$. I simply do not have at the present any such interesting formal principles to suggest.

However, it may be appropriate to mention an alternative way of treating the theory developed here. The one detailed application has been to a non-cooperative game. In a cooperative game, for instance, an arbitration situation, it might be reasonable for the two participants who are in conflict, but who are upholders of ethical rules of behavior, to appoint an arbitrator they both trust. The arbitrator is then asked to make what he considers the fairest preference ranking of $C_1 \cup C_2$ in terms of his knowledge of the participants' needs and wants. Rules (I) and (II), if

applicable, might then determine the outcome of arbitration. The immediate objection to this seems to be that if the arbitrator is going to do the ranking, why not simply let him rank the outcomes, and then agree on the one he considers fairest as the negotiated outcome. There is a simple answer to such an objection. It may be easy to rank $C_1 \cup C_2$, but very difficult to rank $C_1 \times C_2$. For example, let

$C_1 = \{\text{trip to Hawaii, trip to N.Y.}\}$,
$C_2 = \{\text{trip to Florida, trip to Chicago}\}$,

and the arbitrator, knowing persons I and II, may find it easy to rank $C_1 \cup C_2$:

trip to Hawaii,
trip to Florida,
trip to N.Y.,
trip to Chicago,

but he finds it very difficult to compare elements of $C_1 \times C_2$ like ⟨trip to Hawaii, trip to Chicago⟩ and ⟨trip to N.Y., trip to Florida⟩.

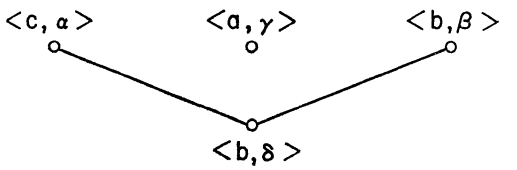
In conclusion an example may be constructed for which equilibrium-point analysis seems to lead to a more equitable and just solution of a non-cooperative game than the theory of justice outlined here. Let

$C_1 = \{a, b, c\}$
$C_2 = \{\alpha, \beta, \gamma, \delta\}$,

let $R_1 = R_2 =$ the ranking: a, α, b, β, c, δ, γ, and let the game matrix be:

I \ II	1	2
1	$\langle a, \gamma \rangle$	$\langle b, \beta \rangle$
2	$\langle c, \alpha \rangle$	$\langle b, \delta \rangle$

Then $J_1 = J_2$, and we have as the Hasse diagram of the justice partial ordering of $C_1 \times C_2$:



It is easily checked that decision 1 is the unique justice-saturated strategy for each player, yielding the outcome $\langle a, \gamma \rangle$, whereas the unique equilibrium point strategies yield $\langle b, \beta \rangle$ as the outcome. In terms of the ordinal properties of the consequences at least, outcome $\langle b, \beta \rangle$ seems fairer than $\langle a, \gamma \rangle$. I conclude that the theory of justice developed here satisfactorily solves only a certain perhaps small proper subset of two-person, non-cooperative games.

The difficulties of formulating a theory of justice for even a very restricted set of situations suggests there may be something seriously wrong with this kind of effort, at least in terms of any principles we seem able to formulate at present. What seems needed as a prolegomena is the painstaking working out of some less sweeping, more concrete grading principles of the sort needed to take a position on particular issues of economic, political or social significance. Example 4 is a sketch of one sort in this direction.

NOTES

[1] I am indebted to Richard Brandt, Donald Davidson and F. Studnicki for a number of useful and penetrating criticisms of a much earlier draft of this paper written in 1957 and circulated as a technical report in that year under the title, 'Two formal models for moral principles'.

[2] Kant's views are typical: "...in matters which concern all men without distinction, nature cannot be accused of any partial distribution of her gifts; and that with regard to essential interests of human nature, the highest philosophy can achieve no more than that guidance which nature has vouchsafed even to the meanest understanding" (1949a, p. 666).

[3] I emphasize that consequences are to be construed broadly here. Causal as well as logical relationships are relevant, but an exact discussion of the significance of causal concepts in the present context would require too lengthy a digression to be appropriate.

[4†] Article 6 in this volume.

[5] In this respect it seems unfortunate that in his inaugural lecture *Theory of Games as a Tool for the Moral Philosopher* (1955) Professor Braithwaite picked for detailed analysis an example which would not in practice be subject to elaborate calculations. His painstakingly careful presentation would apply equally well to more realistic labor-management bargaining situations.

[6] However, second-order moral principles as ethical rules of behavior directly governing acts are introduced in the final section.

[7] The intuitive idea behind a Hasse diagram is simple: if point x may be reached from point y by a continually ascending, not necessarily straight line, then xGy.

[8] The proof is immediate that the principle of unanimity yields a strict partial ordering of C.

[9] The symbol $\cup$ denotes the union of two sets. A binary relation is a set of ordered couples, whence we may speak of the union of two relations.

[10] In the literature of socialist economics, Administrator A is often the Central Planning Board, but a bureaucratic assignment of weights is not essential to the economic theory of the welfare state (cf. Lange and Taylor, 1938).

[11] These remarks are admittedly Kantian in flavor. Cf., "And just as nothing follows from the primary formal principles of our judgments of truth except when primary material grounds are given, so also no particular definite obligation follows from these ... rules except when indemonstrable material principles of practical knowledge are connected with them" [Kant, 1949b, pp. 283–284].

[12] A game is non-cooperative when no precommunication or bargaining between the players is permitted. The prisoner's dilemma is attributed to A. W. Tucker.

[13] The identification of C_1 and C_2 merely simplifies the presentation and is not essential.

[14] Under the equivalence relation which "identifies" elements like $\langle r, m \rangle$ and $\langle m, r \rangle$, $C_1 \times C_2$ is a lattice with respect to J_i, but this fact is of no significance here.

[15] It is perhaps useful to mention that in general a game of the type being considered here does not have a unique equilibrium point; the prisoner's dilemma is a happy exception.

[16] In general, finite games only have equilibrium points when mixed strategies (i.e., probability mixtures of pure strategies) are admitted. A discussion of Rules (I) and (II) with respect to mixed strategies would take us too far afield.

11. PROBABILISTIC INFERENCE AND THE CONCEPT OF TOTAL EVIDENCE*

I. INTRODUCTION

My purpose is to examine a cluster of issues centering around the so-called statistical syllogism and the concept of total evidence. The kind of paradox that is alleged to arise from uninhibited use of the statistical syllogism is of the following sort.

(1) The probability that Jones will live at least fifteen years given that he is now between fifty and sixty years of age is r. Jones is now between fifty and sixty years of age. Therefore, the probability that Jones will live at least fifteen years is r.

On the other hand, we also have:

(2) The probability that Jones will live at least fifteen years given that he is now between fifty-five and sixty-five years of age is s. Jones is now between fifty-five and sixty-five years of age. Therefore, the probability that Jones will live at least fifteen years is s.

The paradox arises from the additional reasonable assertion that $r \neq s$, or more particularly that $r > s$. The standard resolution of this paradox by Carnap (1950, p. 211), Barker (1957, pp. 76–77), Hempel (1965, p. 399) and others is to appeal to the concept of total evidence. The inferences in question are illegitimate, because the total available evidence has not been used in making the inferences. Taking the premises of the two inferences together, we know more about Jones than either inference alleges, namely, that he is between fifty-five and sixty years of age. (Parenthetically I note that if Jones happens to be a personal acquaintance what else we know about him may be beyond imagining, and if we were asked to estimate the probability of his living at least fifteen years we might find

* Reprinted from *Aspects of Inductive Logic* (ed. by J. Hintikka and P. Suppes), North-Holland Publ. Co., Amsterdam, 1966, pp. 49–65.

it impossible to lay out the total evidence that we should use according to Carnap *et al.*, in making our estimation.)

There are at least two good reasons for being suspicious of the appeal to the concept of total evidence. In the first place, we seem in ordinary practice continually to make practical estimates of probabilities, as in forecasting the weather, without explicitly listing the evidence on which the forecast is based. At a deeper often unconscious level the estimations of probabilities involved in most psychomotor tasks – from walking up a flight of stairs to catching a ball – do not seem to satisfy Carnap's injunction that any application of inductive logic must be based on the total evidence available. Or, at the other end of the scale, many actually used procedures for estimating parameters in stochastic processes do not use the total experimental evidence available, just because it is too unwieldy a task (see, e.g., the discussion of pseudo-maximum-likelihood estimates in Suppes and Atkinson (1960, Chap. 2). It might be argued that these differing sorts of practical examples have as a common feature just their deviation from the ideal of total evidence, but their robustness of range, if nothing else, suggests there is something wrong with the idealized applications of inductive logic with an *explicit* listing of the total evidence as envisioned by Carnap.

Secondly, the requirement of total evidence is totally missing in deductive logic. If it is taken seriously, it means that a wholly new principle of a very general sort must be introduced as we pass from deductive to inductive logic. In view of the lack of a sharp distinction between deductive and inductive reasoning in ordinary talk, the introduction of such a wholly new principle should be greeted with considerable suspicion.

I begin my critique of the role of the concept of total evidence with a discussion of probabilistic inference.

II. PROBABILISTIC INFERENCE

As a point of departure, consider the following inference form:

$$\frac{\begin{array}{l} P(A \mid B) = r \\ P(B) = \rho \end{array}}{P(A) \geqq r\rho\,.} \tag{3}$$

In my own judgment (3) expresses the most natural and general rule of detachment in probabilistic inference. (As we shall see shortly, it is often useful to generalize (3) slightly and to express the premises also as inequalities,

$$\text{(3a)} \qquad \frac{\begin{array}{r} P(A \mid B) \geqq r \\ P(B) \geqq \rho \end{array}}{P(A) \geqq r\rho .}$$

The application of (3a) considered below is to take $r=\rho=1-\varepsilon$.) It is easy to show two things about (3); first, that this rule of probabilistic inference is derivable from elementary probability theory (and Carnap's theory of confirmation as well, because a confirmation function $c(h, e)$ satisfies all the elementary properties of conditional probability), and secondly, no contradiction can be derived from two instances of (3) for distinct given events B and C, but they may, as in the case of deductive inference, be combined to yield a complex inference.

The derivation of (3) is simple. By the theorem on total probability, or by an elementary direct argument

$$\text{(4)} \qquad P(A) = P(A \mid B)\, P(B) + P(A \mid \bar{B})\, P(\bar{B}),$$

whence because probabilities are always non-negative, we have at once from the premises that $P(A \mid B)=r$ and $P(B)=\rho$, $P(A)\geqq r\rho$. Secondly, from the four premises

$$\begin{array}{r} P(A \mid B) = r \\ P(B) = \rho \\ P(A \mid C) = s \\ P(C) = \sigma , \end{array}$$

we conclude at once that $P(A)\geqq \max(r\rho, s\sigma)$, and no contradiction results. Moreover, by considering the special case of $P(B)=P(C)=1$, we move close to (1) and (2) and may prove that $r=s$. First we obtain, again by an application of the theorem on total probability and observation of the fact that $P(\bar{B})=0$ if $P(B)=1$, the following inference form as a special case of (3)

$$\text{(5)} \qquad \frac{\begin{array}{r} P(A \mid B) = r \\ P(B) = 1 \end{array}}{P(A) = r.}$$

The proof that $r=s$ when $P(B)=P(C)=1$ is then obvious:

$$(6)\qquad \begin{cases} (1)\ P(A \mid B) = r & \text{Premise} \\ (2)\ \quad P(B) = 1 & \text{Premise} \\ (3)\ P(A \mid C) = s & \text{Premise} \\ (4)\ \quad P(C) = 1 & \text{Premise} \\ (5)\ \quad P(A) = r & 1, 2 \\ (6)\ \quad P(A) = s & 3, 4 \\ (7)\ \quad\ \ r = s & 5, 6. \end{cases}$$

The proof that $r=s$ seems to fly in the face of statistical syllogisms (1) and (2) as differing predictions about Jones. This matter I want to leave aside for the moment and look more carefully at the rule of detachment (3), as well as the more general case of probabilistic inference.

For a given probability measure P the validity of (3) is unimpeachable. In view of the completely elementary – indeed, obvious – character of the argument establishing (3) as a rule of detachment, it is in many ways hard to understand why there has been so much controversy over whether a rule of detachment holds in inductive logic. Undoubtedly the source of the controversy lies in the acceptance or rejection of the probability measure P. Without explicit relative frequency data, objectivists with respect to the theory of probability may deny the existence of P, and in similar fashion confirmation theorists may also if the language for describing evidence is not explicitly characterized. On the other hand, for Bayesians like myself, the existence of the measure P is beyond doubt. The measure P is a measure of partial belief, and it is a condition of coherence or rationality on my simultaneously held beliefs that P satisfy the axioms of probability theory (forceful arguments that coherence implies satisfaction of the axioms of probability are to be found in the literature, starting at least with de Finetti, 1937). It is not my aim here to make a general defense of the Bayesian viewpoint, but rather to show how it leads to a sensible and natural approach to the concept of total evidence.

On the other hand, I emphasize that much of what I have to say can be accepted by those who are not full-fledged Bayesians. For example, what I have to say about probabilistic inference will be acceptable to anyone who is able to impose a common probability measure on the events or premises in question.

For the context of the present paper the most important thing to

emphasize about the rule of detachment (3) is that its application in an argument requires no query as to whether or not the total evidence has been considered. In this respect it has exactly the same status as the rule of detachment in deductive logic. On the other hand it is natural from a logical standpoint to push for a still closer analogue to ordinary deductive logic by considering Boolean operations on events.

It is possible to assign probabilities to at least three kinds of entities: sentences, propositions and events. To avoid going back and forth between the sentence-approach of confirmation theory and the event-approach of standard probability theory, I shall use event-language but standard sentential connectives to form terms denoting complex events. For those who do not like the event-language, the events may be thought of as propositions or elements of an abstract Boolean algebra. In any case, I shall use the language of logical inference to talk about one event implying the other, and so forth.

First of all, we define $A \rightarrow B$, as $\bar{A} \vee B$ in terms of Boolean operations on the events A and B. And analogous to (3), we then have, as a second rule of detachment:

$$(7) \qquad \begin{array}{l} P(B \rightarrow A) \geqq r \\ \quad P(B) \geqq \rho \\ \hline \therefore P(A) \geqq r + \rho - 1 . \end{array}$$

The proof of (7) uses the general addition law rather than the theorem on total probability:

$$\begin{aligned} P(B \rightarrow A) &= P(\bar{B} \vee A) \\ &= P(\bar{B}) + P(A) - P(\bar{B} \,\&\, A) \\ &\geqq r , \end{aligned}$$

whence, solving for $P(A)$,

$$\begin{aligned} P(A) &\geqq r - P(\bar{B}) + P(\bar{B} \,\&\, A) \\ &\geqq r - (1 - \rho) \\ &\geqq r + \rho - 1 , \end{aligned}$$

as desired. The general form of (7) does not seem very enlightening, and we may get a better feeling for it if we take the special but important case that we want to claim both premises are known with near certainty, in

particular, with probability equal to or greater than $1-\varepsilon$. We then have

$$\begin{array}{l} P(B \rightarrow A) \geqq 1-\varepsilon \\ \quad P(B) \geqq 1-\varepsilon \\ \hline \therefore P(A) \geqq 1-2\varepsilon. \end{array} \tag{8}$$

It is worth noting that the form of the rule of detachment in terms of conditional probabilities does not lead to as much degradation from certainty as does (8), for

$$\begin{array}{l} P(A \mid B) \geqq 1-\varepsilon \\ \quad P(B) \geqq 1-\varepsilon \\ \hline \therefore P(A) \geqq (1-\varepsilon)^2, \end{array} \tag{9}$$

and for $\varepsilon > 0$, $(1-\varepsilon)^2 > 1-2\varepsilon$. It is useful to have this well-defined difference between the two forms of detachment, for it is easy, on casual inspection, to think that ordinary-language conditionals can be translated equivalently in terms of conditional probability or in terms of the Boolean operation corresponding to material implication. Which *is* the better choice I shall not pursue here, for application of either rule of inference does not require an auxiliary appeal to a court of total evidence.

Consideration of probabilistic rules of inference is not restricted to detachment. What is of interest is that classical sentential rules of inference naturally fall into two classes, those for which the probability of the conclusion is less than that of the individual premises, and those for which this degradation in degree of certainty does not occur. Tollendo ponens, tollendo tollens, the rule of adjunction (forming the conjunction), and the hypothetical syllogism all lead to a lower bound of $1-2\varepsilon$ for the probability of the conclusion given that each of the two premises is assigned a probability of at least $1-\varepsilon$. The rules that use only one premise, e.g., the rule of addition (from A infer $A \vee B$), the rule of simplification, the commutative laws and De Morgan's laws assign a lower probability bound of $1-\varepsilon$ to the conclusion given that the premise has probability of at least $1-\varepsilon$.

We may generalize this last sort of example to the following theorem.

THEOREM 1: *If* $P(A) \geqq 1-\varepsilon$ *and* A *logically implies* B *then* $P(B) \geqq 1-\varepsilon$.

Proof: We observe at once that if A logically implies B then $\bar{A} \cup B = X$, the whole sample space, and therefore $A \subseteq B$, but if $A \subseteq B$, then $P(A) \leqq P(B)$, whence by hypothesis $P(B) \geqq 1-\varepsilon$.

It is also clear that Theorem 1 can be immediately generalized to any finite set of premises.

THEOREM 2: *If each of the premises* $A_1, \ldots, A_n$ *has probability of at least* $1-\varepsilon$ *and these premises logically imply* B *then* $P(B) \geqq 1-n\varepsilon$.

Moreover, in general the lower bound of $1-n\varepsilon$ *cannot be improved on,* i.e., *equality holds in some cases whenever* $1-n\varepsilon \geqq 0$.

Proof: By hypothesis for $i=1, \ldots, n$, $P(A_i) \geqq 1-\varepsilon$. We prove by induction that under this hypothesis $P(A_1 \,\&\cdots\&\, A_n) \geqq 1-n\varepsilon$. The argument for $n=1$ is immediate from the hypothesis. Suppose it holds for n. Then by an elementary computation

$$\begin{aligned} P(A_1 \,\&\cdots\&\, A_n \,\&\, A_{n+1}) &= 1 - (1 - P(A_1 \,\&\cdots\&\, A_n)) \\ &\quad - (1 - P(A_{n+1})) \\ &\quad + P((A_1 \,\&\cdots\&\, A_n) \,\&\, \bar{A}_{n+1}) \\ &\geqq 1 - (1 - P(A_1 \,\&\cdots\&\, A_n)) \\ &\quad - (1 - P(A_{n+1})) \\ &\geqq 1 - n\varepsilon - \varepsilon \\ &\geqq 1 - (n+1)\,\varepsilon, \end{aligned}$$

as desired. (Details of how to handle quantifiers, which are not explicitly treated in the standard probability discussions of the algebra of events, may be found in Gaifman, 1964, or Krauss and Scott, 1966. The basic idea is to take as the obvious generalization of the finite case

$$P((\exists x)\, Ax) = \sup \{P(Aa_1 \vee Aa_2 \vee \cdots \vee Aa_n)\},$$

where the sup is taken over all finite sets of objects in the domain. Replacing sup by inf we obtain a corresponding expression for $P((\forall x)Ax)$. Apart from details it is evident that however quantifiers are handled, the assignment of probabilities must be such that Theorem 1 is satisfied, i.e., that if A logically implies B then the probability assigned to B must be at least as great as the probability assigned to A, and this is all that is required for the proof of Theorem 2.)

The proof that the lower bound $1-n\varepsilon$ cannot in general be improved upon reduces to constructing a case for which each of the n premises has probability $1-\varepsilon$, but the conjunction, as a logical consequence of the premises taken jointly has probability $1-n\varepsilon$, when $1-n\varepsilon \geqq 0$. The example I use is most naturally thought of as a temporal sequence of events

$A_1, \ldots, A_n$. Initially we assign

$$P(A_1) = 1 - \varepsilon$$
$$P(\bar{A}_1) = \varepsilon.$$

Then

$$P(A_2 \mid A_1) = \frac{1 - 2\varepsilon}{1 - \varepsilon}$$
$$P(A_2 \mid \bar{A}_1) = 1,$$

and more generally

$$P(A_n \mid A_{n-1}A_{n-2} \ldots A_1) = \frac{1 - n\varepsilon}{1 - (n-1)\varepsilon}$$
$$P(A_n \mid A_{n-1}A_{n-2} \ldots \bar{A}_1) = 1$$
$$\vdots$$
$$P(A_n \mid \bar{A}_{n-1}\bar{A}_{n-2} \ldots \bar{A}_1) = 1,$$

in other words for any combination of preceding events on trials 1 to $n-1$ the conditional probability of A_n is 1, except for the case $A_{n-1}A_{n-2} \ldots A_1$. The proof by induction that $P(A_n) = 1 - \varepsilon$ and $P(A_nA_{n-1} \ldots A_1) = 1 - n\varepsilon$ is straightforward. The case for $n = 1$ is trivial. Suppose now the assertion holds for n. Then by inductive hypothesis

$$\begin{aligned} P(A_{n+1}A_n \ldots A_1) &= P(A_{n+1} \mid A_n \ldots A_1)\, P(A_n \ldots A_1) \\ &= \frac{1 - (n+1)\varepsilon}{1 - n\varepsilon}(1 - n\varepsilon) \\ &= 1 - (n+1)\varepsilon, \end{aligned}$$

and by the theorem on total probability

$$\begin{aligned} P(A_{n+1}) = {} & P(A_{n+1} \mid A_n \ldots A_1)\, P(A_n \ldots A_1) \\ & + [P(A_{n+1} \mid A_n \ldots \bar{A}_1)\, P(A_n \ldots \bar{A}_1) + \cdots \\ & + P(A_{n+1} \mid \bar{A}_n \ldots \bar{A}_1)\, P(\bar{A}_n \ldots \bar{A}_1)]. \end{aligned}$$

By construction all the conditional probabilities referred to in the bracketed expression are 1, and the unconditional probabilities in this expression by inductive hypothesis simply sum to $n\varepsilon$, i.e., $1 - (1 - n\varepsilon)$, whence

$$P(A_{n+1}) = \frac{1 - (n+1)\varepsilon}{1 - n\varepsilon} \cdot (1 - n\varepsilon) + n\varepsilon = 1 - \varepsilon,$$

which completes the proof.

It is worth noting that in interesting special cases the lower bound of $1-n\varepsilon$ can be very much improved. For example, if the premises $A_1, \ldots, A_n$ are statistically independent, then the bound is at least $(1-\varepsilon)^n$.

The intuitive content of Theorem 2 reflects a common-sense suspicion of arguments that are complex and depend on many premises, even when the logic seems impeccable. Overly elaborate arguments about politics, personal motives or circumstantial evidence are dubious just because of the uncertainty of the premises taken jointly rather than individually.

A natural question to ask about Theorem 2 is whether any nondeductive principles of inference that go beyond Theorem 2 arise from the imposition of the probability measure P on the algebra of events. Bayes' theorem provides an immediate example. To illustrate it with a simple artificial example, suppose we know that the composition of an urn of black (B) and white (W) balls may be exactly described by one of two hypotheses. According to hypothesis H_r, the proportion of white balls is r, and according to H_s, the proportion is s. Moreover, suppose we assign a priori probability ρ to H_r and $1-\rho$ to H_s. Our four premises may then be expressed so:

$$\begin{aligned} P(W \mid H_r) &= r \\ P(W \mid H_s) &= s \\ P(H_r) &= \rho \\ P(H_s) &= 1-\rho. \end{aligned}$$

Given that we now draw with replacement, let us say, two white balls, we have as the likelihood of this event as a consequence of the first two premises

$$\begin{aligned} P(WW \mid H_r) &= r^2 \\ P(WW \mid H_s) &= s^2, \end{aligned}$$

and thus by Bayes' theorem, we may infer

$$P(H_r \mid WW) = \frac{r^2\rho}{r^2\rho + s^2(1-\rho)}, \tag{10}$$

and this is clearly not a logical inference from the four premises. Logical purists may object to the designation of Bayes' theorem as a principle of inference, but there is little doubt that ordinary talk about inferring is very close to Bayesian ideas, as when we talk about predicting the weather or Jones' health, and such talk also has widespread currency among

statisticians and the many kinds of people who use statistical methods to draw probabilistic inferences.

The present context is not an appropriate one in which to engage upon a full-scale analysis of the relation between logical and statistical inference. I have only been concerned here to establish two main points about inference. First, in terms of standard probability theory there is a natural form of probabilistic inference, and inference from probabilistically given premises involves no appeal to the concept of total evidence. Second, all forms of such probabilistic inference are not subsumed within the forms of logical inference, and two examples have been given to substantiate this claim, one being the rule of detachment as formulated for conditional probability and the other being Bayes' theorem.

III. THE STATISTICAL SYLLOGISM RE-EXAMINED

There is, however, a difficulty about the example of applying Bayes' theorem that is very similar to the earlier difficulty with the statistical syllogism. I have not stated as explicit premises the evidence *WW* that two white balls were drawn, and the reason I have not provides the key for re-analyzing the statistical syllogism and removing all air of paradox from it.

The evidence *WW* has not been included in the statement of the premises of the Bayesian example, because the probability measure *P* referred to in the premises is the measure that holds *before* any taking of evidence (by drawing a ball) occurs. The measure *P* does provide a means of expressing the a posteriori probability *after* the evidence is taken as a conditional probability, but the hypothetical or conditional nature of this assertion has been too little appreciated. Using just the measure *P* there is no way to express that *in fact* two white balls were drawn, rather than, say, a white ball and then a black ball. Using conditional probabilities we can express the a posteriori probabilities of the two hypotheses under any *possible* outcomes of one, two or more drawings. What we cannot express in these terms is the *actual* evidence, and it is a mistake to try. (It should be apparent that these same remarks apply to Carnapian confirmation functions.) Commission of this mistake vitiates what appears to be the most natural symbolic formulation of the statistical syllogism – the inference form (5) as a special case of (3).

We can symbolize statistical syllogism (1) as follows, where $e(x)$ is the life expectancy of person x and $a(x)$ is the age of person x, and let j=Jones:

$$(11)\qquad \frac{\begin{array}{l}P(e(j) \geqq 15 \mid 50 < a(j) < 60) = r \\ 50 < a(j) < 60\end{array}}{\therefore P(e(j) \geqq 15) = r.}$$

Now let us schematize this inference in terms of *hypothesis* and *evidence* as these notions occur in Bayes' theorem

$$(12)\qquad \frac{\begin{array}{l}P(\text{hypothesis} \mid \text{evidence}) = r \\ \text{evidence}\end{array}}{\therefore P(\text{hypothesis}) = r,}$$

and the incorrect character of this inference is clear. From the standpoint of Bayes' theorem it asserts that once we know the evidence, the a posteriori probability $P(H \mid E)$ is equal to the a priori probability $P(H)$, and this is patently false. The difficulty is that the measure P cannot be used to assert that $P(50<a(j)<60)=1$, which is, I take it, a direct consequence of the assertion that $50<a(j)<60$. (I shall expand on this point later.) The measure P is the measure used to express the conditional probabilities about Jones' life expectancy generated by any possible evidence. The inference-form expressed by (11) is illegitimate because the same probability measure does not apply to the two premises and the conclusion, as the scheme (12) makes clear when compared to Bayes' theorem.

Because there seems to be something genuine even if misleading about the statistical syllogism, it is natural to ask what are nonparadoxical ways of symbolizing it. One way is simply to adopt the symbolism used in Bayes' theorem, and then the conclusion is just the same as the first premise, the assertion of the a posteriori probability P(hypothesis | evidence). A related approach that makes the inference seem less trivial is the following. First, we symbolize the major premise in universal form, rather than with particular reference to Jones, for example:

The probability that a male resident of California will live at least 15 years given that he is now between 50 and 60 years is r, or symbolically, where $m(x)$ is male resident of California,

$$P(e(x) \geqq 15 \mid m(x) \;\&\; 50 < a(x) < 60) = r,$$

and secondly, given as second premise the event A that

$$m(j) \;\&\; 50 < a(j) < 60,$$

we may write the conclusion in terms of a new probability measure P_A conditionalized on A: $P_A(e(j) \geqq 15) = r$. Moreover, it is clear that no paradox arises from (2) because the evidence expressed in the second premise of (2) represents an event B distinct from A, and the conclusion $P_B(e(j) \geqq 15) = s$, is consistent with the conclusion $P_A(e(j) \geqq 15) = r$ of (1).

There is still another way of putting the matter which may provide additional insight into the inferential kernel inside the dubious statistical syllogism. We may think of the premises as all the a posteriori probabilities given all the different possible kinds of evidence. As an additional final premise, some evidence A is asserted. On this basis a new measure P_A is generated and the probability of the hypothesis is then asserted in terms of this new measure P_A, as the conclusion of the inference.

At this point it might seem easy to insist that delicate questions of consistency or coherence about the probability measure P do indeed differentiate deductive and inductive logic, but this is not at all the case. The problem of temporal order of knowledge is as characteristic of deductive as of inductive logic. In discussing deductive canons of inference we tacitly assume the statements whose inferential relations are being considered are all asserted or denied at a given time or are timeless in character. It is not a paradox of deductive logic that the joint assertion of two statements true at different times leads to a paradox – for example, *it rained yesterday*, and *it did not rain yesterday*. The same thing, I have argued, is to be said about the statistical syllogism. The same probability measure does not apply to the first and second premise; the measure referred to in the first premise is temporally earlier than the one implicit in the second premise and the conclusion.

In the next section I turn to these temporal problems and their relation to the complex task of defining rationality, but before doing this I want briefly to pull several strands together and summarize in slightly different fashion the place given to the concept of total evidence by the view of probability advocated here.

According to this view it is automatic that if a person is asked for the probability of an event at a given time it will follow from the conditions of

coherence on all his beliefs at that time that the probability he assigns to the event automatically takes into account the total evidence that he believes has relevance to the occurrence of the event. The way in which the total evidence is brought in is straightforward and simple through the theorem on total probability. To be quite clear how this theorem operates it may be useful to take a somewhat detailed look at the gradual expansion of the probability of an event A in terms of given evidence B and C. For purposes of generality we may assume that the probability of B and C is not precisely 1 and therefore deal with the general case. (In the interest of compactness of notation, I here let juxtaposition denote intersection or conjunction.) First we have

$$P(A) = P(A \mid B)\,P(B) + P(A \mid \bar{B})\,P(\bar{B}). \tag{13}$$

We also have

$$P(A) = P(A \mid C)\,P(C) + P(A \mid \bar{C})\,P(\bar{C}). \tag{14}$$

And in terms of both B and C we have the more complex version:

$$\begin{aligned} P(A) = {} & P(A \mid BC)\,P(BC) + P(A \mid B\bar{C})\,P(B\bar{C}) \\ & + P(A \mid \bar{B}C)\,P(\bar{B}C) + P(A \mid \bar{B}\bar{C})\,P(\bar{B}\bar{C}). \end{aligned} \tag{15}$$

In the special case that $P(B)=P(C)=1$, we then have

$$P(A) = P(A \mid B) = P(A \mid C) = P(A \mid BC).$$

We have in the general case, as indications of the relations between "partial" and "total" evidence,

$$P(A \mid B)\,P(B) = P(A \mid BC)\,P(BC) + P(A \mid B\bar{C})\,P(B\bar{C})$$

and

$$P(A \mid \bar{B})\,P(\bar{B}) = P(A \mid \bar{B}C)\,P(\bar{B}C) + P(A \mid \bar{B}\bar{C})\,P(\bar{B}\bar{C}).$$

The point of exhibiting these identities is to show that no separate concept of total evidence need be added to the concept of a probability measure on an individual's beliefs. It also may seem that these identities show that the notion of conditional probability is not even needed. The important point however is that the serious use contemplated here for the notion of conditional probability is in terms of passing from the probability measure expressing partial beliefs at one time to a later time. It is just by

conditionalizing in terms of the events that actually occurred that this passage is made a good deal of the time by most information-processing organisms.

I conclude this paper with a more detailed look at the processes by which beliefs are changed.

IV. RATIONAL CHANGES IN BELIEF

It seems important to recognize that the partial beliefs, or probability beliefs as we may term them, that an individual holds as a mature adult are not in any realistic way, even for an ideally rational individual, to be obtained simply by conditionalization, that is, in terms of conditional probability, from an overall probability measure which the individual was born with or acquired early in life. The patent absurdity of this idea does not seem to have been adequately reflected upon in Bayesian discussions of these matters. The static adherence to a single probability measure independent of time is characteristic of de Finetti (1937) and Savage (1954), but even a superficial appraisal of the development of a child's information-processing capacities makes it evident that other processes than conditionalization are required to explain the beliefs held by a mature adult. Moreover, even an adult who does not live in a terribly static and simple world will need other processes than conditionalization to explain the course of development of his beliefs during his years as an adult.

Acceptance of the view that a person's beliefs at time t are to be expressed by a probability measure special for that time raises certain problems that go beyond ordinary talk about probabilities. Under this view what probability are we to assign to events that a person knows have occurred? I can see no other course but to assert that such events have probability 1. Thus if I know that Jones is between 50 and 60 years of age then this event has probability 1 for me. We shall take as an axiom linking knowing and probabilistic beliefs as well as an axiom linking beliefs and probabilistic beliefs exactly this assertion, that is, if a person knows or believes that an event occurred then the probability of this event for that person is 1. For example, since I now believe with certainty in my own mind that Julius Caesar stayed several weeks in Gaul the probability of this event for me is 1.

The kind of problem that is solved by this axiom, and that is troublesome in a more detailed look at the standard Bayesian viewpoint, is the following. It is customary to remark, as has already been indicated, that the probability of an event should be conditionalized on that which has occurred. But it is also natural to ask what currently is the probability to be attached to an event that has occurred. Thus if the event B has occurred what probability is to be assigned to it (here I engage in the standard simplification of talking about the event occurring and not using the more careful locution of talking about knowing or believing that the event has occurred). It is not sufficient to say that we may refer to the probability of B given B because in this vein we can talk about the conditional probability of any event given that event. It is necessary to assign probability 1 to all such conditional probabilities independent of whether or not the event B has actually occurred.

To put in another way an argument already given, once the continual adjustment of probability to the current state of affairs is supposed, and there is much in ordinary thinking and language that supports this idea, the problem of seeming to need to introduce a separate assumption about using all the evidence available simply disappears. As has been shown, it is an immediate consequence of the theorem on total probability that when we discuss the probability of an event A the relevance of any information we have about the event is immediately absorbed in a calculation of this probability, as a direct consequence of the theorem on total probability.

Unfortunately, there is another problem of total evidence, related to but not the same as, the one we started with. This new problem would seem to occupy a central position in any analysis of how we change our beliefs from one time to another. The problem is that of characterizing what part of the welter of potential information impinging on an organism is to be accepted as evidence and how this evidence is to be used to change the organism's basic probability measure. From the standpoint of psychology and physiology, a satisfactory empirical answer seems a great distance away. The fact of our empirical ignorance about matters conceptually so close to central problems of epistemology is philosophically important, but of still greater philosophical importance is the fact that our general concept of rationality seems intimately involved in any answer, empirical or not, that we give. Until we can say how an organism of given capacities

and powers should process the information impinging on it we cannot have a very deep running concept of rationality.

The point that I want to make here needs some delineation. The problem is not one of giving many instances in which there is great agreement on what is the rational way to process information. If a man is in a house that is burning, he is irrational calmly to continue to listen to music, or if a man is driving down a highway at high speed, he is irrational if he becomes absorbed in the beauty of the landscape and looks for forty or fifty seconds at an angle nearly perpendicular to the road itself. These simple instances can be multiplied to any extent desired, but what is not easy to multiply is the formal characterization of the principles that should be applied and that govern the variety of cases in which we all have a clear intuitive judgment.

To a large extent work in inductive logic and statistical inference tends to obscure the fundamental character of this problem of giving principles by which information is to be judged important and to be responded to. The reason for this neglect in inductive logic or theoretical statistics is that once the formal language or the random variables are selected, then the problem of information-processing is reduced to relatively simple proportions. The selection of the language or the selection of the random variables, as the case may be, is the largest single decision determining how information will be processed, and to a very large extent the simple rule of conditionalizing, so that the measure held at a later time arises from an earlier probability measure as a conditional probability measure, then furnishes the appropriate way to proceed. It might almost be said that the rule of *pure prudence* is always to derive beliefs held at time t from beliefs held at time t' by conditionalization on the probability measure characterizing partial beliefs at the earlier time. Although such a principle of pure prudence may seem attractive at first glance, in my own judgment it is a piece of pure fantasy. A myriad of events are occurring at all times and are noticeable by a person's perceptual apparatus. What is not the least bit clear is what sort of filter should be imposed by the individual on this myriad of events in order to have a workable simplifying scheme of decision and action. The highly selective principles of attention that must necessarily be at work do not seem to be characterizable in any direct way from the concept of conditional probability.

The two interrelated processes that any adequate theory of rationality

must characterize are the process of information selection and the process of changing beliefs according to the significance of the information that is selected. In other words, to what should the rational man pay attention in his ever-changing environment and how should he use the evidence of that to which he does attend in order to change his beliefs and thereby modify his decisions and actions.

What seems particularly difficult to describe are rational mechanisms of selection. In a first approximation the classical theory of maximizing expected utility as developed by Ramsey, de Finetti, Savage and others uses the mechanism of conditional probabilities to change beliefs once the information to be attended to has been selected. This classical mechanism is certainly inadequate for any process of concept formation, and thus for any very deep running change in belief, and as Jeffrey (1965) points out, it is not even adequate to the many cases in which changes in belief may be expressed simply as changes in probability but not explicitly in terms of changes in conditional probability, because the changes in probability are not completely analyzable in terms of the explicitly noticed occurrence of events. For example, the probability that I will assign to the event of rain tomorrow will change from morning to afternoon, even though I am not able to express in explicit form the evidence I have used in coming to this change. In Suppes (1966) I have attempted to show how inadequate the Bayesian approach of conditional probability is in terms of even fairly simple processes of concept formation, and I do not want to go over that ground again here, except to remark that it is clear on the most casual inspection that all information processing that an organism engages in cannot be conceived of in terms of conditional probabilities.

It is even possible to question whether *any* changes can be so expressed. The measure P effectively expressing my beliefs at time t cannot be used to express what I actually observe immediately after time t, for P is already "used up" so to speak in expressing the a priori probability of each possible event that might occur, and cannot be used to express the unconditional occurrence of that which in fact did happen at time t. Pragmatically, the situation is clear. If an event A occurs and is noticed, the individual then changes his belief pattern from the measure P to the conditional measure P_A, and our reformulated version of the statistical syllogism is exemplified. What has not been adequately commented upon in discussions of these matters by Bayesians is that the probability

measure P held at time t cannot be used to express what actually happened immediately after t, but only to express, at the most, how P would change *if* so and so did happen.

In any case, genuine changes in conceptual apparatus cannot in any obvious way be brought within this framework of conditional probability. As far as I can see, the introduction of no genuinely new concept in the history of scientific or ordinary experience can be adequately accounted for in terms of conditional probabilities. A fundamental change in the algebra of events itself, not just their probabilities, is required to account for such conceptual changes. Again, it is a problem and a responsibility of an adequate theory of normative information processing to give an account of how such changes should take place.

What I would like to emphasize in conclusion is the difficulty of expressing in systematic form the mechanisms of attention a rationally-operating organism should use. It is also worth noting that any interesting concept of rationality for the behavior of men must be keyed to a realistic appraisal of the powers and limitation of the perceptual and mental apparatus of humans. These problems of attention and selection do not exist for an omniscient God or an information-processing machine whose inputs are already schematized into something extraordinarily simple and regular. The difficulty and subtlety of characterizing the mechanisms of information selection and at the same time a recognition of their importance in determining the actual behavior of men make me doubt that the rather simple Carnapian conception of inductive logic can be of much help in developing an adequate theory of rational behavior. Even the more powerful Bayesian approach provides a very thin and inadequate characterization of rationality, because only one simple method for changing beliefs is admitted. It is my own view that there is little chance of defining an adequate concept of rationality until analytical tools are available to forge a sturdy link between the selection and use of evidence and processes of concept formation.

PART III

FOUNDATIONS OF PHYSICS

My earliest interest in philosophy was in the philosophy of physics, and as I remarked in the preface, I have not included any work on the foundations of classical mechanics, because that work, undertaken jointly with J. C. C. McKinsey, is covered in some detail in my *Introduction to Logic*. Over the past decade, my interests have shifted more to mathematical psychology and to the foundations of psychology, but I continue to retain my original interest in the foundations of physics, and hope that I shall be able to make contributions to the subject in the years to come. If time and energy permitted, I would like best to write a kind of Bourbaki of physics showing how set-theoretical methods can be used to organize all parts of theoretical physics and bring to all branches of theoretical physics a uniform language and conceptual approach. I rather suspect, however, that only a very small circle of scholars would be interested in such work. Physicists consider the subject of no real interest from the standpoint of physics, the present generation of philosophers with some mathematical training are little interested in physics, and mathematicians are interested only if new results of mathematical interest are obtained. Perhaps the message is that system building, even of this austere kind, is not currently in fashion as a way of doing the philosophy of physics.

The four articles in this part do not represent any sort of system building, but concentrate on conceptually important aspects of relativity theory and quantum mechanics. The first of the four articles is concerned mainly with the derivation of the Lorentz transformations of the special theory of relativity from an explicit, but minimal set of assumptions. Since the initial publication in 1959 of Article 12 on the derivation of the Lorentz transformations, important new results have been published by Walter Noll (1964) and also by E. C. Zeeman (1964).

The elegant aspect of Noll's paper is that he axiomatizes Minkowskian chronometry using coordinate-free methods. The representation and derivations in Article 12 depend upon coordinate methods. Zeeman shows that it is not necessary to assume invariance of timelike intervals as

in Article 12, but that it is sufficient to assume the preservation of order, that is, the relativistic partial ordering of beforeness between points is sufficient to guarantee derivation of the Lorentz transformations. Like many simple and beautiful ideas, it is surprising that this did not occur to someone sooner. The key to the results was already present in the early work of Robb (1936), which shows that the binary relation of beforeness is a sufficient conceptual basis for the kinematical theory of special relativity. (Reference to Robb's very original book is to be found at the end of Article 12.)

The phrase 'with or without parity' in the title of Article 12 refers to whether it is possible to derive the direction of time from the axioms of relativistic kinematics. It is clearly not possible for the axioms given in that paper. In order to do this, I discuss in the final section the possibility of introducing a relation of signaling. It is obvious that this can be done very directly in an *ad hoc* fashion. What is needed, however, is some natural approach that is fully satisfying from an intuitive and a conceptual standpoint. In his article, Noll makes some remarks about this, and he raises the question of whether his approach solves the problem I raised. Essentially, Noll introduces a directed signal relation that is asymmetric, and of course if we postulate that the numerical representation must preserve the signal direction in terms of signals passing from earlier to later events, the direction of time is guaranteed. I find this approach unsatisfactory since this is an arbitrary stipulation in the definition of isomorphism, and we get just as good an isomorphism from a structural standpoint if the direction in time is reversed. What is needed are substantive postulates about the nature of signaling, probably in terms of the spread of information, but any such postulates will necessarily take us far beyond the ordinary kinematics of the special theory of relativity. I suppose my present view is that there is no hope of deriving the direction of time within a framework of ideas natural to the special theory of relativity, but deeper investigation of this question is certainly desirable.

The last three articles of Part III, Articles 13 to 15, are concerned with probability concepts in quantum mechanics and the relation of these probability concepts to a nonclassical logic of quantum mechanics. The number of philosophically interesting problems that remain open in this domain is large, and I certainly recognize the modest nature of my own

contribution to the problems. From a purely logical standpoint, a thorough investigation of the axiomatization as a sentential logic of the nonclassical logic of quantum mechanics expressed in algebraic form in Article 15 would be a matter of interest to some logicians and some philosophers of science. It is too remote from physics itself to be of major interest to those wrapped up in the conceptual problems of quantum mechanics. Much more pressing from the standpoint of the foundations of physics is the working out of a fully satisfactory axiomatization of classical quantum mechanics. Probably the best presentation as yet available is that found in Mackey (1963). Mackey, however, does not reach the level of a specific representation of physical problems; and it is not clear how to extend Mackey's axioms in a natural way to give particular experimental situations a categorical representation. Such extensions in the case of classical mechanics are obvious. In other words, once we specify the state of the system at a given instant in terms of the positions and velocities of all particles, and express a force function, the motions of the particles are determined uniquely. But general assumptions that match observables to operators in a constructive way are not a part of Mackey's axiomatization.

The other essential topic in the philosophy of quantum mechanics not touched upon here is the collection of problems generally lumped together under the heading of problems of measurement in quantum mechanics. What I have said in the three articles included here about probability in quantum mechanics is, I believe, important for working out a correct theory of measurement in quantum mechanics, but that is about all that can be claimed. For a look at some of the problems central to the quantum mechanical theory of measurement, I would recommend an article by my younger colleague, Joseph Sneed (1966).

Finally, for a spirited defense of classical logic in quantum mechanics, I would recommend the recent article by Arthur Fine (1968). In spite of Fine's clearly stated arguments, I continue to think the case is strong for defending the nonclassical character of both probability and logic in quantum mechanics. If the view I defend in Articles 13–15 is correct, this deviation from classical probability theory and classical logic is the most philosophically significant aspect of quantum mechanics.

12. AXIOMS FOR RELATIVISTIC KINEMATICS WITH OR WITHOUT PARITY*

I. INTRODUCTION

The primary aim of this paper is to give an elementary derivation of the Lorentz transformations, without any assumptions of continuity or linearity, from a single axiom concerning invariance of the relativistic distance between any two space-time points connected by an inertial path. The concluding section considers extensions of the theory of relativistic kinematics which will destroy conservation of temporal parity, that is, extensions which are not invariant under time reversals.

It is philosophically and empirically interesting that the Lorentz transformations can be derived without any extraneous assumptions of continuity or differentiability. In a word, the single assumption needed for relativistic kinematics is that all observers at rest in inertial frames get identical measurements of relativistic distances along inertial paths when their measuring instruments have identical calibrations. Note that it is a consequence and *not* an assumption that these observers are moving with a uniform velocity with respect to each other. Granted the possibility of perfect measurements everywhere of relativistic intervals, this single axiom isolates in a precise way the narrow operational basis needed for the special theory of relativity.

Prior to any search of the literature it would seem that this result would be well known, but I have not succeeded in finding the proof anywhere. Every physics textbook on relativity makes a linearity assumption at the minimum. In geometrical discussions of indefinite quadratic forms it is often remarked that the relativistic interval is invariant under the Lorentz group, but it is not proved that it is invariant under no wider group, which is the main fact established here. Some further remarks in this connection are made at the end of Section II.

* Reprinted from *The Axiomatic Method with Special Reference to Geometry and Physics* (ed. by L. Henkin, P. Suppes, and A. Tarski), North-Holland Publ. Co., Amsterdam, 1959, pp. 291–307.

II. PRIMITIVE NOTIONS AND SINGLE AXIOM

Our single initial axiom for relativistic kinematics is based on three primitive notions, each of which has a simple physical interpretation. The first notion is an arbitrary set X interpreted as the set of *physical space-time points.* The second notion is a nonempty family $\mathfrak{F}$ of one-one functions mapping X onto R_4, the set of all ordered quadruples of real numbers. (Thus X must have the power of the continuum.) Intuitively each function in $\mathfrak{F}$ represents an *inertial space-time frame of reference*, or, more explicitly, a space-time measuring apparatus at rest in an inertial frame. If $x \in X, f \in \mathfrak{F}$, and $f(x) = \langle x_1, x_2, x_3, t \rangle$ then x_1, x_2, and x_3 are the three orthogonal spatial coordinates of the point x, and t the time coordinate, with respect to the frame f. For a more explicit formal notation, $f_i(x)$ is the ith coordinate of the space-time point x with respect to the frame f, for $i = 1, \ldots, 4$. The third primitive notion is a positive number c, which is to be interpreted as the *speed of light.*

It is convenient to have a notation for the *relativistic distance* with respect to a frame f between any two space-time points x and y.

DEFINITION 1: *If $x, y \in X$ and $f \in \mathfrak{F}$ then*

$$I_f(xy) = \sqrt{\sum_{i=1}^{3} [f_i(x) - f_i(y)]^2 - c^2 [f_4(x) - f_4(y)]^2}.$$

(We always take the square root with positive sign.) If f is an inertial frame, then (i) $I_f(xy) = 0$ if x and y are connected by a light line; (ii) $I_f^2(xy) < 0$ if x and y lie on an inertial path (the square is negative since $I_f(xy)$ is imaginary); (iii) $I(xy) > 0$ if x and y are separated by a "space-like" interval. We use (ii) for a formal definition.

DEFINITION 2: *If $x, y \in X$ and $f \in \mathfrak{F}$ then* x AND y LIE ON AN INERTIAL PATH WITH RESPECT TO f *if and only if* $I_f^2(xy) < 0$.

It will also occasionally be useful to characterize inertial paths in terms of their speed. We may do this informally as follows. By the *slope* of a line α in R_4, whose projection on the 4th coordinate (the time coordinate) is a non-degenerate segment, we mean the three-dimensional vector W such that for any two distinct points $\langle Z_1, t_1 \rangle$ and $\langle Z_2, t_2 \rangle$ of α

$$W = \frac{Z_1 - Z_2}{t_1 - t_2}.$$

By the *speed* of α we mean the nonnegative number $|W|$. An *inertial path* is a line in R_4 whose speed is less than c; and a *light line* is of course a line whose speed is c.

The single axiom we require is embodied in the following definition.

DEFINITION 3: *A system* $\mathfrak{X}=\langle X, \mathfrak{F}, c\rangle$ *is a* COLLECTION OF RELATIVISTIC FRAMES *if and only if for every* x, y *in* X, *whenever* x *and* y *lie on an inertial path with respect to some frame in* $\mathfrak{F}$, *then for all* f, f' *in* $\mathfrak{F}$

$$(1) \qquad I_f(xy) = I_{f'}(xy).$$

I originally formulated this invariance axiom so as to require that Equation (1) hold for *all* space-time points x and y, that is, without restricting them to lie on an inertial path (with respect to some frame in $\mathfrak{F}$). Walter Noll pointed out to me that with this stronger axiom no physically motivated arguments of the kind given below are required to prove that any two frames in $\mathfrak{F}$ are related by a linear transformation; a relatively simple algebraic argument may be given to show this.

On the other hand, when the invariance assumption is restricted, as it is here, to distances between points on inertial paths, the line of argument formalized in the theorems of the next section seems necessary. This restriction to pairs of points on inertial paths is physically natural because their distances $I_f(xy)$ are more susceptible to direct measurements than are the distances of points separated by a space-like interval (i.e., $I_f(xy)>0$).

III. THEOREMS

In proving the main result that any two frames in $\mathfrak{F}$ are related by a Lorentz transformation, some preliminary definitions, theorems and lemmas will be useful. We shall use freely the geometrical language appropriate to Euclidean four-dimensional space with the ordinary positive definite quadratic form.

THEOREM 1: *If* $k\geqslant 0$ *and* $f(x)-f(y)=k[f(u)-f(v)]$ *then* $I_f(xy)=kI_f(uv)$.

Proof: If $k=0$, the theorem is immediate. So we need to consider the case for which $k>0$. It follows from the hypothesis of the theorem that

$$(1) \qquad x_i - y_i = k(u_i - v_i) \quad \text{for} \quad i = 1, \ldots, 4,$$

where, for brevity here and subsequently, when we are considering a fixed element f of $\mathfrak{F}$, $f_i(x)=x_i$, etc. Using (1) and Definition 1 we then have:

$$\begin{aligned} I_f(xy) &= \sqrt{\sum_{i=1}^{3}(x_i-y_i)^2 - c^2(x_4-y_4)^2} \\ &= \sqrt{\sum_{i=1}^{3}k^2(u_i-v_i)^2 - c^2k^2(u_4-v_4)^2} \\ &= kI_f(uv). \quad \text{Q.E.D.} \end{aligned}$$

In the next theorem we use the notion of *betweenness* in a way which is meant not to exclude identity with one of the end points.

THEOREM 2: *If the points $f(x)$, $f(y)$ and $f(z)$ are collinear and $f(y)$ is between $f(x)$ and $f(z)$ then*

$$I_f(xy) + I_f(yz) = I_f(xz).$$

Proof: Extending our subscript notation, let $f(x)=\boldsymbol{x}$, etc. Since the three points $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are collinear, and $\boldsymbol{y}$ is between $\boldsymbol{x}$ and $\boldsymbol{z}$, there is a number k such that $0\leqslant k\leqslant 1$ and

(1) $$\boldsymbol{y} = k\boldsymbol{x} + (1-k)\,\boldsymbol{z},$$

whence

$$\boldsymbol{y} - \boldsymbol{z} = k(\boldsymbol{x} - \boldsymbol{z}),$$

and thus by Theorem 1

(2) $$I_f(yz) = kI_f(xz).$$

By adding and subtracting $\boldsymbol{x}$ from the right-hand side of (1), we get:

$$\boldsymbol{y} = k\boldsymbol{x} + (1-k)\,\boldsymbol{z} + \boldsymbol{x} - \boldsymbol{x},$$

whence

$$\boldsymbol{x} - \boldsymbol{y} = (1-k)\,(\boldsymbol{x} - \boldsymbol{z}),$$

and thus by virtue of Theorem 1 again,

(3) $$I_f(xy) = (1-k)\,I_f(xz).$$

Adding (2) and (3) we obtain the desired result:

$$I_f(xy) + I_f(yz) = I_f(xz). \quad \text{Q.E.D.}$$

Our next objective is to prove a partial converse of Theorem 2. Since the notion of Lorentz transformation is needed in the proof, we introduce the appropriate formal definitions at this point. $\mathscr{I}$ is the identity matrix of the necessary order.

DEFINITION 4: *A matrix* $\mathscr{A}$ *(of order* 4*) is a* LORENTZ MATRIX *if and only if there exist real numbers* β, δ, *a three-dimensional vector* U, *and an orthogonal matrix* $\mathscr{E}$ *of order* 3 *such that*

$$\beta^2\left(1-\frac{U^2}{c^2}\right)=1$$

$$\delta^2=1$$

$$\mathscr{A}=\begin{pmatrix}\mathscr{E} & 0\\ 0 & \delta\end{pmatrix}\begin{pmatrix}\mathscr{I}+\dfrac{\beta-1}{U^2}U^*U & -\dfrac{\beta U^*}{c^2}\\ -\beta U & \beta\end{pmatrix}.$$

(In this definition and elsewhere, if A is a matrix, A^* is its transpose, and vectors like U are one-rowed matrices – thus U^* is a one-column matrix.) The physical interpretation of the various quantities in Definition 1 should be obvious. The number β is the *Lorentz contraction factor*. When $\delta=-1$, we have a reversal of the direction of time. The matrix $\mathscr{E}$ represents a *rotation* of the spatial coordinates, or a rotation followed by a reflection. The vector U is the *relative velocity* of the two frames of reference. For future reference it may be noted that every Lorentz matrix is nonsingular.

DEFINITION 5: *A Lorentz transformation is a one-one function* φ *mapping* R_4 *onto itself such that there is a Lorentz matrix* $\mathscr{A}$ *and a* 4*-dimensional vector* B *so that for all* Z *in* R_4

$$\varphi(Z)=Z\mathscr{A}+B.$$

The physical interpretation of the vector B is clear. Its first three coordinates represent a translation of the origin of the spatial coordinates, and its last coordinate a translation of the time origin. Definition 5 makes it clear that every Lorentz transformation is a nonsingular affine transformation of R_4, a fact which we shall use in several contexts. The important consideration for the proof of Theorem 3 is that affine transformations preserve the collinearity of points.

THEOREM 3: *If any two of the three points* x, y, z *are distinct and lie on an*

inertial path with respect to f and if $I_f(xy)+I_f(yz)=I_f(xz)$, then the points $f(x), f(y)$ and $f(z)$ are collinear, and $f(y)$ is between $f(x)$ and $f(z)$.

Proof: Three cases naturally arise.

Case 1: $I^2(xy)<0$. In this case the line segment $f(x)-f(y)$ is an inertial path segment from x to y, and there exists a Lorentz transformation φ which will transform the segment $f(x)-f(y)$ to "rest", that is, more precisely, φ may be chosen so as to transform f to a frame f', which need not be a member of $\mathfrak{F}$, such that the spatial coordinates of x and y are at the origin, the time coordinate of x is zero, and z has but one spatial coordinate, by appropriate spatial rotation. That is, we have:

$$f'(x) = \langle 0, 0, 0, 0\rangle,$$
$$f'(y) = \langle 0, 0, 0, y'_4\rangle,$$
$$f'(z) = \langle z'_1, 0, 0, z'_4\rangle.$$

We shall prove that $f'(x), f'(y)$ and $f'(z)$ are collinear. Since φ is non-singular and affine, its inverse φ^{-1} exists and is affine, whence collinearity is preserved in transforming from f' back to f.

It is a familiar fact that the relativistic intervals $I_f(xy)$, $I_f(yz)$ and $I_f(xz)$ are Lorentz invariant and thus have the same value with respect to f' as f. Consequently, from the additive hypothesis of the theorem, we have:

$$(1) \qquad \sqrt{-c^2y'^2_4} + \sqrt{z'^2_1 - c^2(y'_4 - z'_4)^2} = \sqrt{{}'^2_1 - c^2z'^2_4}.$$

Squaring both sides of (1), then cancelling and rearranging terms, we obtain:

$$(2) \qquad \sqrt{-y'^2_4}\cdot\sqrt{z'^2_1 - c^2(y'_4 - z'_4)^2} = cy'_4(y'_4 - z'_4).$$

If $y'_4=0$, then x and y are identical, contrary to the hypothesis that $I^2(xy)<0$. Taking then $y'_4\neq 0$, dividing it out in (2), squaring both sides and cancelling, we infer:

$$-z'^2_1 = 0,$$

whence

$$z'_1 = 0,$$

which establishes the collinearity in f' of the three points, since their spatial coordinates coincide, and obviously $f'(y)$ is between $f'(x)$ and $f'(z)$.

Case 2: $I_f^2(yz)<0$. Proof similar to Case 1.

Case 3: $I_f^2(xz)<0$. By an argument similar to that given for Case 1, we may go from f to a frame f' by a Lorentz transformation which will transform the inertial segment $f(x)-f(z)$ to "rest". That is, we obtain:

$$f'(x)=\langle 0, 0, 0, 0\rangle,$$
$$f'(y)=\langle y'_1, 0, 0, y'_4\rangle,$$
$$f'(z)=\langle 0, 0, 0, z'_4\rangle.$$

Then by the additive hypothesis of the theorem:

$$(3)\qquad \sqrt{y_1'^2-c^2y_4'^2}+\sqrt{y_1'^2-c^2(y'_4-z'_4)^2}=\sqrt{-c^2z_4'^2}.$$

Proceeding as before, by squaring and cancelling, we obtain from (3):

$$(4)\qquad \sqrt{-c^2z_4'^2}\cdot\sqrt{y_1'^2-c^2y_4'^2}=-c^2y'_4z'_4.$$

Squaring again and cancelling yields:

$$(5)\qquad y_1'^2z_4'^2=0.$$

There are now two possibilities to consider: either $y'_1=0$ or $z'_4=0$. If the former is the case, then the three points are collinear in R_4, for they are all three placed at the origin of the spatial coordinates. On the other hand, if $z'_4=0$, then x and z are identical points, contrary to hypothesis. Again it is obvious that $f'(y)$ is between $f'(x)$ and $f'(z)$. Q.E.D.

That a full converse of Theorem 2 cannot be proved, in other words that the additive hypothesis

$$I_f(xy)+I_f(yz)=I_f(xz)$$

does not imply collinearity, is shown by the following counterexample:

$$f(x)=\langle 0, 0, 0, 0\rangle,$$
$$f(y)=\langle 1, 1, 0, 0\rangle$$
$$f(z)=\langle\sqrt{2c}, 0, 0, 1\rangle.$$

Clearly, $f(x)$, $f(y)$ and $f(z)$ are not collinear in R_4, but $I_f(xy)+I_f(yz)=I_f(xz)$, that is,

$$(1)\qquad \sqrt{2}+\sqrt{(1-\sqrt{2}c)^2+1-c^2}=\sqrt{2c^2-c^2}.$$

For, simplifying and rearranging (1), we see it is equivalent to:

$$\sqrt{2-2\sqrt{2}c+c^2}=c-\sqrt{2} \tag{2}$$

and the left-hand of (2) is simply

$$\sqrt{(c-\sqrt{2})^2}=c-\sqrt{2}.$$

(It may be mentioned that the full converse of Theorem 2 does hold for R_2, that is, when there is a restriction to one spatial dimension.)

We now want to prove some theorems about properties which are invariant in $\mathfrak{F}$. Formally, a property is *invariant* in $\mathfrak{F}$ if and only if it holds or does not hold uniformly for every member f of $\mathfrak{F}$. Thus to say that the property of a line being an inertial path is invariant in $\mathfrak{F}$ means that a line with respect to f in $\mathfrak{F}$, is an inertial path with respect to f if and only if it is an inertial path with respect to every f' in $\mathfrak{F}$. All geometric objects referred to here are with respect to the frames in $\mathfrak{F}$.

THEOREM 4: *The property of being the midpoint of a finite segment of an inertial path is invariant in* $\mathfrak{F}$.

Proof: Suppose x, y and z lie on an inertial path with respect to f and

$$f(y)=\tfrac{1}{2}f(x)+\tfrac{1}{2}f(z), \tag{1}$$

and thus

$$f(y)-f(x)=\tfrac{1}{2}[f(z)-f(x)].$$

Consequently by virtue of Theorem 1

$$I_f(xy)=\tfrac{1}{2}I_f(xz) \tag{2}$$

and similarly

$$I_f(yz)=\tfrac{1}{2}I_f(xz), \tag{3}$$

whence

$$I_f(xy)+I_f(yz)=I_f(xz). \tag{4}$$

Now by the invariance axiom of Definition 3, for any f' in $\mathfrak{F}$

$$I_{f'}(xy)=I_f(xy)$$
$$I_{f'}(yz)=I_f(yz)$$
$$I_{f'}(xz)=I_f(xz).$$

Substituting these identities in (4) we obtain:

$$I_{f'}(xy) + I_{f'}(yz) = I_{f'}(xz).$$

Thus by virtue of Theorem 3, $f'(x)$, $f'(y)$ and $f'(z)$ are collinear with $f'(y)$ between $f'(x)$ and $f'(z)$. Moreover, since by the invariance axiom (2) and (3) hold for f', we conclude $f'(y)$ is actually the midpoint. Q.E.D.

This proof is easily extended to show that the property of being an inertial path is invariant in $\mathfrak{F}$, but we do not directly need this fact. We next want to show that this midpoint property is invariant for arbitrary segments. In view of the counterexample following Theorem 3 it is evident that a direct proof in terms of the relativistic intervals cannot be given. The method we shall use consists essentially of constructing a parallelogram whose sides are segments of inertial paths. A similar but somewhat more complicated proof is given in Rubin and Suppes (1954).

THEOREM 5: *The property of being the midpoint of an arbitrary finite segment is invariant in* $\mathfrak{F}$.

Proof: Let $A = \langle Z_1, t_1 \rangle$ and $B = \langle Z_2, t_2 \rangle$ where A is an arbitrary segment in R_4. (The points A to G defined here are with respect to f in $\mathfrak{F}$.) For definiteness assume $t_1 \geqslant t_2$. We set

$$Z_0 = \frac{Z_1 + Z_2}{2}$$

and we choose t_0 and t_3 so that

$$t_0 < t_2 - \frac{|Z_1 - Z_2|}{2c},$$

$$t_3 > t_1 + \frac{|Z_1 - Z_2|}{2c},$$

$$|A - \langle Z_0, t_3 \rangle| = |\langle Z_0, t_0 \rangle - B|,$$

$$|A - \langle Z_0, t_0 \rangle| = |\langle Z_0, t_3 \rangle - B|.$$

We now let (see Figure 1)

$$C = \langle Z_0, t_0 \rangle, \quad D = \langle Z_0, t_3 \rangle,$$

$$E = \frac{A + B}{2}, \qquad F = \frac{B + D}{2},$$

$$G = \frac{A + C}{2}.$$

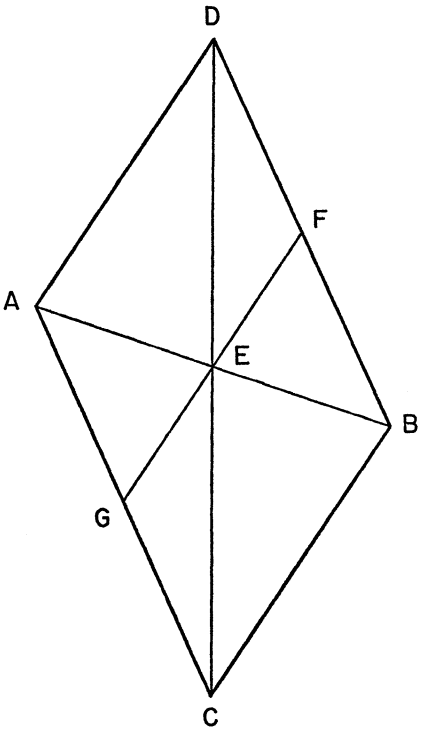


Fig. 1.

Denoting now the same points with respect to f' in $\mathfrak{F}$ by primes, we have by virtue of this construction in f and the invariance property of Theorem 4,

(1) $E' = \frac{1}{2}(C' + D')$,

(2) $F' = \frac{1}{2}(B' + D')$,

(3) $G' = \frac{1}{2}(A' + C')$,

(4) $E' = \frac{1}{2}(F' + G')$.

Substituting (2) and (3) into (4) we have:

$$E' = \tfrac{1}{2}[\tfrac{1}{2}(B' + D') + \tfrac{1}{2}(A' + C')]$$
$$= \tfrac{1}{2}[\tfrac{1}{2}(A' + B') + \tfrac{1}{2}(C' + D')].$$

Now substituting (1) into the right-hand side of the last equation and

simplifying, we infer the desired result:

$$E' = \tfrac{1}{2}(A' + B'),$$

since by construction $E=\frac{1}{2}(A+B)$. Thus the midpoint of an arbitrary segment is preserved. Q.E.D.

THEOREM 6: *The property of two finite segments of inertial paths being parallel and in a fixed ratio is invariant in* $\mathfrak{F}$.

Proof: Let $f(x)-f(y)=k[f(u)-f(v)]$, with $f(x)-f(y)$ and $f(u)-f(v)$ segments of inertial paths. Without loss of generality we may assume $k \geqslant 1$. Let z be the point such that $f(x)-f(y)=k[f(x)-f(z)]$. We now construct a parallelogram with $f(u)-f(v)$ and $f(x)-f(z)$ as two parallel sides. By the previous theorem any parallelogram in f is carried into a parallelogram in f' since the midpoint of the diagonals is preserved. Thus

$$f'(u) - f'(v) = f'(x) - f'(z), \tag{1}$$

but by Theorems 2 and 3

$$f'(x) - f'(y) = k[f'(x) - f'(z)], \tag{2}$$

(for details see proof of Theorem 4), whence from (1) and (2)

$$f'(x) - f'(y) = k[f'(u) - f'(v)]. \quad \text{Q.E.D.}$$

As the final theorem about properties invariant in $\mathfrak{F}$, we want to generalize the preceding theorem to arbitrary finite segments.

THEOREM 7: *The property of two arbitrary finite segments being parallel and in a fixed ratio is invariant in* $\mathfrak{F}$.

Proof: In view of preceding theorems, the crucial thing to show is that if

$$f(x) - f(y) = k[f(x) - f(z)]$$

then

$$f'(x) - f'(y) = k[f'(x) - f'(z)].$$

Our approach is to use an "inertial" parallelogram similar to the one used in the proof of Theorem 5. In fact an exactly similar construction will be used; points A to E are constructed identically, where $A=f(x)$ and $B=f(y)$. Without loss of generality we may assume $k>2$, that is, that $f(z)=F$ is between A and E. We then have that

$$A - E = (k/2)[A - F]. \tag{1}$$

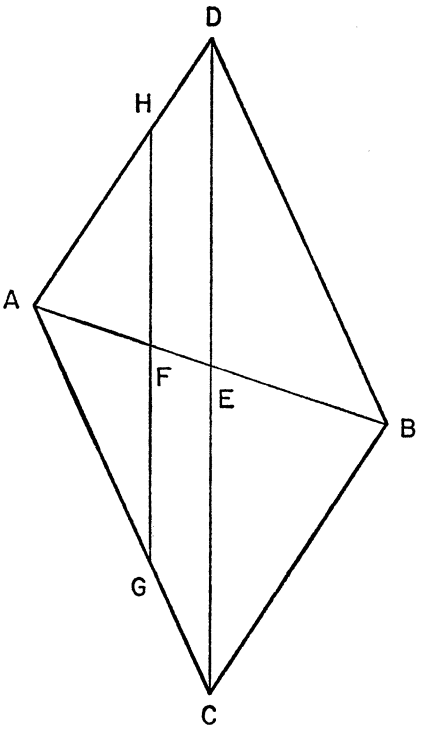


Fig. 2.

We draw through F a line parallel to CD, which cuts AC at G and AD at H. (See Figure 2.)
Now (1) is equivalent to:

$$F = (1 - 2/k)\, A + (2/k)\, E. \tag{2}$$

Moreover, by construction

$$F = \tfrac{1}{2}(G + H) \tag{3}$$

$$E = \tfrac{1}{2}(C + D) \tag{4}$$

$$G = (1 - 2/k)\, A + (2/k)\, C \tag{5}$$

$$H = (1 - 2/k)\, A + (2/k)\, D. \tag{6}$$

Since GFH, AGC, AHD and CED are by construction segments of in-

ertial paths, by virtue of Theorem 7, we have from (3)–(6):

$$F' = \tfrac{1}{2}(G' + H') \tag{7}$$

$$E' = \tfrac{1}{2}(C' + D') \tag{8}$$

$$G' = (1 - 2/k)\, A' + (2/k)\, C' \tag{9}$$

$$H' = (1 - 2/k)\, A' + (2/k)\, D'. \tag{10}$$

Substituting (9) and (10) in (7), we get:

$$F' = (1 - 2/k)\, A' + (1/k)(C' + D'). \tag{11}$$

And now substituting (8) in (11), we obtain the desired result:

$$F' = (1 - 2/k)\, A' + (2/k)\, E'. \tag{12}$$

But now by virtue of Theorem 5

$$E' = \tfrac{1}{2}(A' + B'),$$

which together with (12) yields:

$$F' = (1 - 1/k)\, A' + (1/k)\, B',$$

which is equivalent to:

$$f'(x) - f'(y) = k[f'(x) - f'(z)]. \tag{13}$$

The remainder of the proof, based upon considering $f(x)-f(y)=k[f(u)-f(v)]$, is exactly like that of Theorem 6 and may be omitted. (In place of Theorems 2 and 3 in that proof we use the result just established.) Q.E.D.

We now state the theorem toward which the preceding seven have been directed.

THEOREM 8: *Any two frames in* $\mathfrak{F}$ *are related by a nonsingular affine transformation.*

Proof: A familiar necessary and sufficient condition that a transformation of a vector space be affine is that parallel finite segments with a fixed ratio be carried into parallel segments with the same fixed ratio. (See, e.g. Birkhoff and MacLane, 1941, p. 263.) Hence by virtue of Theorem 7 any two frames are related by an affine transformation. Nonsingularity of the transformation follows from the fact that each frame in $\mathfrak{F}$ is a one-one mapping of X onto R_4. Q.E.D.

Once we have any two frames in $\mathfrak{F}$ related by an affine transformation, it is not difficult to proceed to show that they are related by a Lorentz transformation. In the proof of this latter fact, it is convenient to use a Lemma about Lorentz matrices, which is proved in Rubin and Suppes (1954) and is simply a matter of direct computation.

Lemma 1: *A matrix* $\mathscr{A}$ *(of order* 4*) is a Lorentz matrix if and only if*

$$\mathscr{A}\begin{pmatrix}\mathscr{I} & 0\\ 0 & -c^2\end{pmatrix}\mathscr{A}^* = \begin{pmatrix}\mathscr{I} & 0\\ 0 & -c^2\end{pmatrix}.$$

We now prove the basic result:

THEOREM 9: *Any two frames in* $\mathfrak{F}$ *are related by a Lorentz transformation.*

Proof: Let f, f' be two frames in $\mathfrak{F}$. As before, for x in X, $f(x)=\boldsymbol{x}$, $f_1(x)=x_1$, $f'(x)=\boldsymbol{x}'$, etc. We consider the transformation φ such that for every x in X, $\varphi(\boldsymbol{x})=\boldsymbol{x}'$. By virtue of Theorem 8 there is a nonsingular matrix (of order 4) and a four-dimensional vector B such that for every x in X

$$\varphi(\boldsymbol{x}) = \boldsymbol{x}\mathscr{A} + B.$$

The proof reduces to showing that $\mathscr{A}$ is a Lorentz matrix.

Let

$$\mathscr{A} = \begin{pmatrix}\mathscr{D} & E^*\\ F & g\end{pmatrix}. \tag{1}$$

And let α be a light line (in f) such that for any two distinct points $\boldsymbol{x}$ and $\boldsymbol{y}$ of α if $\boldsymbol{x}=\langle Z_1, t_1\rangle$ and $\boldsymbol{y}=\langle Z_2, t_2\rangle$, then

$$\frac{Z_1 - Z_2}{t_1 - t_2} = W. \tag{2}$$

Clearly $|W|=c$. Now let

$$W' = \frac{Z_1' - Z_2'}{t_1' - t_2'}. \tag{3}$$

From (1), (2) and (3) we have:

$$W' = \frac{(Z_1 - Z_2)\,\mathscr{D} + (t_1 - t_2)\,F}{(Z_1 - Z_2)\,E^* + (t_1 - t_2)\,g}. \tag{4}$$

Dividing all terms on the right of (4) by $t_1 - t_2$, and using (2), we obtain:

$$W' = \frac{W\mathscr{D} + F}{WE^* + g}. \tag{5}$$

At this point in the argument we need to know that $|W'| = c$, that is to say, we need to know that if $I_f(xy) = 0$, then $I_{f'}(xy) = 0$. The proof of this fact is not difficult. From our fundamental invariance axiom we have that $I_{f'}(xy) \geqslant 0$, that is,

$$| W' | \geq c. \tag{6}$$

Consider now a sequence of inertial lines $\alpha_1, \alpha_2, \ldots$ whose slopes $W_1, W_2, \ldots$ are such that

$$\lim_{n \to \infty} W_n = W. \tag{7}$$

Now corresponding to (5) we have:

$$|W_n'| = \left| \frac{W_n\mathscr{D} + F}{W_nE^* + g} \right| < c. \tag{8}$$

Whence, from (8) we conclude that if $WE^* + g \neq 0$, then

$$|W'| = |\lim_{n \to \infty} W_n'| \leqslant c. \tag{9}$$

Thus from (6) and (9) we infer

$$|W'| = c, \tag{10}$$

if $WE^* + g \neq 0$, but that this is so is easily seen. For, suppose not. Then

$$\lim_{n \to \infty} (W_nE^* + g) = 0,$$

and thus

$$\lim_{n \to \infty} (W_n\mathscr{D} + F) = 0.$$

Consequently $W\mathscr{D} + F = 0$, and $\langle W, 1 \rangle_{\mathscr{A}} = 0$, which is absurd in view of the nonsingularity of $\mathscr{A}$.

Since $|W'| = c$, we have by squaring (5):

$$\frac{W\mathscr{D}\mathscr{D}^*W^* + 2W\mathscr{D}F^* + |F|^2}{(WE^* + g)^2} = c^2, \tag{11}$$

and consequently

$$W(\mathscr{D}\mathscr{D}^* - c^2E^*E)\,W^* + 2W(\mathscr{D}F^* - c^2E^*g) + |F|^2 - c^2g = 0. \tag{12}$$

Since (12) holds for an arbitrary light line, we may replace W by $-W$, and obtain (12) again. We thus infer:

$$W(\mathscr{D}F^* - c^2E^*g) = 0,$$

but the direction of W is arbitrary, whence

$$\mathscr{D}F^* - c^2E^*g = 0. \tag{13}$$

Now let $x = \langle 0, 0, 0, 0\rangle$ and $y = \langle 0, 0, 0, 1\rangle$. Then

$$I_f^2(xy) = -c^2.$$

But it is easily seen from (1) that

$$I_{f'}^2(xy) = |F|^2 - c^2g^2,$$

and thus by our fundamental invariance axiom

$$c^2g^2 - |F|^2 = c^2. \tag{14}$$

From (12), (13), (14) and the fact that $|W|^2 = c^2$, we infer:

$$W(\mathscr{D}\mathscr{D}^* - c^2E^*E)\,W^* = |W|^2,$$

and because the direction of W is arbitrary we conclude:

$$\mathscr{D}\mathscr{D}^* - c^2E^*E = \mathscr{I}, \tag{15}$$

where $\mathscr{I}$ is the identity matrix.

Now by direct computation on the basis of (1),

$$\mathscr{A}\begin{pmatrix}\mathscr{I} & 0\\ 0 & -c^2\end{pmatrix}\mathscr{A}^* = \begin{pmatrix}\mathscr{D}\mathscr{D}^* - c^2E^*E & \mathscr{D}F^* - c^2E^*g\\ (\mathscr{D}F^* - c^2E^*g)^* & FF^* - c^2g^2\end{pmatrix}. \tag{16}$$

From (13), (14), (15) and (16) we arrive finally at the result:

$$\mathscr{A}\begin{pmatrix}\mathscr{I} & 0\\ 0 & -c^2\end{pmatrix}\mathscr{A}^* = \begin{pmatrix}\mathscr{I} & 0\\ 0 & -c^2\end{pmatrix},$$

and thus by virtue of Lemma 1, $\mathscr{A}$ is a Lorentz matrix. Q.E.D.

IV. TEMPORAL PARITY

Turning now to problems of parity, we may for simplicity restrict the discussion to time reversals. Similar considerations apply to spatial reflections.

A simple axiom, which will prevent time reversal between frames in $\mathfrak{F}$, is:

(T1) *There are elements x and y in X such that for all f in $\mathfrak{F}$*

$$f_4(x) < f_4(y).$$

There is, however, a simple objection to this axiom. It is unsatisfactory to have time reversal depend on the existence of special space-time points, which could possibly occur only in some remote region or epoch. This objection is met by T2.

(T2) *If $I_f^2(xy) < 0$ then either for all f in $\mathfrak{F}$*

$$f_4(x) < f_4(y).$$

or for all f in $\mathfrak{F}$

$$f_4(y) < f_4(x).$$

T2 replaces the postulation of special points by a general property: given any segment of an inertial path, all frames in $\mathfrak{F}$ must orient the direction of time for this segment in the same way.

Nevertheless, there is another objection to T1 which holds also for T2: the appropriate axiom should be formulated so that a given observer in a frame f may verify it without observing any other frames, that is, he may decide if he is a qualified candidate for membership in $\mathfrak{F}$ without observing other members of $\mathfrak{F}$. (This issue is relevant to the single axiom of Definition 3 but cannot be entered into here.) From a logical standpoint this means eliminating quantification over elements of $\mathfrak{F}$, which may be done by introducing a fourth primitive notion, a binary relation σ of *signaling* on X. To block time reversal we need postulate but two properties of σ:

(T3.1) *For every x in X there is a y in X such that $x\sigma y$.*

(T3.2) *If $x\sigma y$ then $f_4(x) < f_4(y)$.*

However, a third objection to (T1) also applies to (T2) and (T3). Namely, we are essentially postulating what we want to prove. The axioms stated here correspond to postulating artifically in a theory of measurement of mass that a certain object must be assigned the mass of one. I pose the question: *Is it possible to find "natural" axioms which fix a direction of time*? It may be mentioned that Robb's meticulous axiomatization (1936) in terms of the notion of *after* provides no answer.

13. PROBABILITY CONCEPTS IN QUANTUM MECHANICS*

I. INTRODUCTION

The fundamental problem considered is that of the existence of a joint probability distribution for momentum and position at a given instant. The philosophical interest of this problem is that for the potential energy functions (or Hamiltonians) corresponding to many simple experimental situations, the joint "distribution" derived by the methods of Wigner (1932) and Moyal (1949) is not a genuine probability distribution at all.

If this "distribution" is accepted as the most reasonable one possible, then we may infer a stronger result than the Heisenberg uncertainty principle, namely, not only are position and momentum not precisely measurable simultaneously, they are not simultaneously measurable at all.

To make the discussion as accessible and elementary as possible, the next section is devoted to the most relevant probability concepts. In Section III the general expression for the joint distribution of momentum and position is derived, and then computed, as an illustration, for the ground state and first excited state of a one-dimensional harmonic oscillator. The implications for the Heisenberg uncertainty principle of the results obtained are analyzed in Section IV. The final section (Section V) consists of some observations concerning the axiomatic foundations of quantum mechanics.

II. SOME PROBABILITY CONCEPTS

For subsequent discussion it will be desirable to have at hand several notions from general probability theory. For some of the definitions given below questions of continuity and differentiability can arise. To avoid these issues, which have no bearing on the central theme, adequate smoothness properties will be assumed in general arguments, and will be exhibited in all particular cases.

* Reprinted from *Philosophy of Science* **28** (1961), 378–389.

To begin with, we assume as the starting point for probability considerations the underlying sample space. Points in this space represent possible experimental outcomes. For example, the set of all sequences of 1's and 0's is an appropriate sample space to represent the possible outcomes of flipping a coin an infinite number of times. Or, to be more finitistic, the set of 16 finite sequences of four terms, each term being 1 or 0 represents the set of possible outcomes of flipping a coin four times. In addition, we need a countably additive probability measure on an appropriately specified (Borel) field of sets. (In the finitistic cases, the probability measure may be defined for all subsets of the sample space.)

A random variable is then simply a (measurable) function defined on the sample space. Later on, we shall be considering the position and momentum of a particle as random variables. For the experiment of flipping a coin four times, a typical random variable would be the function h whose values are the number of 1's in any experimental outcome, i.e., any point of the sample space. Thus, $h(\langle 1, 1, 1, 1\rangle)=4$, $h(\langle 0, 1, 1, 1\rangle)=h(\langle 1, 0, 1, 1\rangle)=h(\langle 1, 1, 0, 1\rangle)=h(\langle 1, 1, 1, 0\rangle)=3$, etc. The practical justification of random variables should be apparent; they permit simplification of the structure of experimental outcomes in order to concentrate on the aspects of the experiment considered most significant.

Let Ξ be the sample space, let P be the probability measure on $B(\Xi)$, the given Borel field of subsets of Ξ, and let $\boldsymbol{X}$ be a (measurable) real-valued function defined on Ξ, i.e., let $\boldsymbol{X}$ be a random variable on Ξ. Then the *distribution function* F of $\boldsymbol{X}$ is defined for every real number x as follows:

$$F(x) = P\{\xi: \xi \in \Xi \ \& \ \boldsymbol{X}(\xi) \leqslant x\}.$$

It is easily shown that F is a monotonically increasing function such that

$$0 \leqslant F(x) \leqslant 1$$
$$F(-\infty) = 0$$
$$F(\infty) = 1.$$

The derivative of F, which we assume exists, is the *density* f. Put the other way round,

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

The expected value or mean $E(\boldsymbol{X})$ is defined as:

$$E(\boldsymbol{X}) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

Let $\bar{x}$ be the mean of $\boldsymbol{X}$. Then the variance of $\boldsymbol{X}$ is defined as:

$$\sigma_{\boldsymbol{X}}^2 = \int_{-\infty}^{\infty} (x-\bar{x})^2 f(x)\,dx.$$

We now turn to the characteristic function of F. Let F be the distribution function of a single random variable. Then the complex-valued function φ of the real variable t such that

$$(1) \qquad \varphi(t) = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx}\,dF(x)$$

is the *characteristic function* of F.

Note that

(i) $\varphi(O)=1$

(ii) $|\varphi(t)| \leqslant \int_{-\infty}^{\infty} dF(x)=1$

(iii) $\varphi(-t)=\overline{\varphi(t)}$.

A distribution is uniquely determined by its characteristic function and conversely. Moreover, various properties of a distribution can be inferred from properties of its characteristic function. Granted that F has a derivative f which is the density function, we obtain at once from (1)

$$(2) \qquad \varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx.$$

From (2) we conclude that f is the Fourier transform of φ, and thus we obtain at once by the standard result for Fourier transforms the important inversion:

$$(3) \qquad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\,dt.$$

The analogue of (3) for two random variables will give us the joint density function of position and momentum.

In fact, developments corresponding to the above for the joint distribution of two (or more) random variables proceed in the expected fashion. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two real-valued random variables defined on the sample space Ξ. Their joint distribution $F(x, y)$ is defined as follows:

$$F(x, y) = P\{\xi : \xi \in \Xi \;\&\; \boldsymbol{X}(\xi) \leqslant x \;\&\; \boldsymbol{Y}(\xi) \leqslant y\}.$$

The joint density function $f(x, y)$ is defined by:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \, \partial y}.$$

The marginal densities $f_1(x)$ and $f_2(y)$ are defined by:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

The two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are (*statistically*) *independent* if for all x and y

$$f(x, y) = f_1(x)\, f_2(y).$$

The means and variances of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined as above, using now $f_1(x)$ for $\boldsymbol{X}$ and $f_2(y)$ for $\boldsymbol{Y}$. The covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})(y - \bar{y})\, f(x, y)\, dx\, dy$$

and the correlation coefficient $\rho_{\boldsymbol{XY}}$ as:

$$\rho_{\boldsymbol{XY}} = \frac{\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})}{\sigma_{\boldsymbol{X}} \sigma_{\boldsymbol{Y}}}.$$

Note that if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent then the covariance and correlation coefficient equal zero, but the converse is not necessarily true.

The characteristic function of the joint distribution function $F(x, y)$ is defined by

$$\varphi(t, u) = E(e^{itX+iuY}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{itx+iuy}\, dF(x, y). \tag{4}$$

Analogously to (2), we have:

$$\varphi(t, u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{itx+iuy} f(x, y)\, dx\, dy. \tag{5}$$

Analogously to (3), we then have by the Fourier inversion theorem

$$f(x, y) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-itx-iuy}\, \varphi(t, u)\, dt\, du. \tag{6}$$

III. JOINT DISTRIBUTION OF MOMENTUM AND POSITION

Consider now the momentum and position random variables $\boldsymbol{P}$ and $\boldsymbol{Q}$. Following (4) the characteristic function $\varphi(t, u)$ is defined by

$$\varphi(u, v) = E(e^{iu\boldsymbol{P}+iv\boldsymbol{Q}}). \tag{7}$$

Throughout this paper we shall for simplicity consider only time-independent phenomena. Using the Hilbert space formulation, let (ψ, ψ) be the inner product of a state with itself. Following the usual formalism, the expectation $E(\boldsymbol{R})$ of an operator $\boldsymbol{R}$ when the quantum mechanical system is in state ψ is simply $(\psi, \boldsymbol{R}\psi)$. In view of (7) the characteristic function $\varphi(u, v)$ for the joint distribution of $\boldsymbol{P}$ and $\boldsymbol{Q}$ is given by:

$$\varphi(u, v) = (\psi, e^{i(up+vq)}\, \psi). \tag{8}$$

Corresponding to (6) we then have from (7) and (8) by Fourier inversion

$$f(p, q) = 1/4\pi^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(up+vq)}(\psi, e^{i(up+vq)}\, \psi)\, du\, dv. \tag{9}$$

For canonically conjugate operators $\boldsymbol{P}$ and $\boldsymbol{Q}$, i.e., $\boldsymbol{PQ}-\boldsymbol{QP}=\hbar/i$, it may

be shown that (8) simplifies to[1]

$$\varphi(u, v) = \int \psi^*(q - \tfrac{1}{2}\hbar u)\, e^{ivq}\, \psi(q + \tfrac{1}{2}\hbar u)\, dq$$

and whence by Fourier inversion

$$(10) \qquad f(p, q) = 1/2\pi \int \psi^*(q - \tfrac{1}{2}\hbar u)\, e^{-iup}\, \psi(q + \tfrac{1}{2}\hbar u)\, du\,.$$

As is well known in probability theory not every characteristic function determines a proper probability distribution, and this is indeed the difficulty with (10).

The expression given by (10) for the joint density was first proposed by Wigner (1932). The derivations just sketched follow Moyal (1949). It is natural to ask how "inevitable" is (10) in the framework of classical quantum mechanics. To the writer it seems to be by far the most obvious approach compatible with standard probability theory and the formalism of quantum mechanics. Different approaches are to be found in Dirac (1945) and Feynman (1948). Dirac's theory leads to distributions which are complex-valued and thus cannot be interpreted as probabilities; formidable mathematical problems beset the Feynman pathintegral approach.

Before considering the Heisenberg uncertainty principle in the light of (10), it will be useful to consider two simple one-dimensional cases, for one of which – the harmonic oscillator in the ground state – $f(p, q)$ is a genuine density, and for the other – the oscillator in the first excited state – it is not.

A. *Ground State*[2]

The potential energy is given by

$$V(x) = \tfrac{1}{2}Kx^2\,,$$

and the time-independent wave equation is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + \frac{1}{2}Kx^2\psi(x) = E\psi(x)\,.$$

The solution of this equation in terms of Hermite polynomials is familiar

from the literature. In the lowest energy state H_0

$$\psi(x) = \left(\frac{\alpha}{\pi^{1/2}}\right)^{1/2} \exp(-\tfrac{1}{2}\alpha^2 x^2), \tag{11}$$

where

$$\alpha^2 = \sqrt{Km/\hbar^2}.$$

Thus

$$|\psi(x)|^2 = (\alpha/\pi^{1/2})\, e^{-\alpha^2 x^2} \tag{12}$$

which is a normal density with mean zero and variance $\sigma^2 = 1/2\alpha^2 = \hbar/2\sqrt{Km}$.

We now apply (10) and (11) to obtain the joint distribution of momentum and position. For convenience of calculation, we replace p by the propagation vector $k = p/\hbar$. We have at once:

$$\begin{aligned} f(k, x) &= \frac{1}{2\pi}\int \psi^*(x - u/2)\, e^{-iku}\, \psi(x + u/2)\, du \\ &= 1/2\pi\left(\frac{\alpha}{\pi^{1/2}}\right)\int \exp[\alpha^2(x^2 + (u/2)^2)]\, e^{-iku}\, du \\ &= \frac{1}{2\pi}\left(\frac{\alpha}{\pi^{1/2}}\right) e^{-\alpha^2 x^2}\, \frac{\pi^{1/2}}{\alpha/2} \exp\left[-\frac{k^2}{4(\alpha/2)^2}\right] \\ &= (1/\pi)\exp(-\alpha^2 x^2 - k^2/\alpha^2). \end{aligned}$$

Or, in terms of the momentum p, the joint density is

$$f(p, x) = \frac{1}{\hbar\pi}\exp\left(-\alpha^2 x^2 - \frac{p^2}{\hbar^2\alpha^2}\right). \tag{13}$$

Integrating out x in (13) we obtain for the marginal distribution of momentum

$$f_1(p) = \frac{1}{\alpha\hbar\pi^{1/2}}\exp\left(-\frac{p^2}{\hbar^2\alpha^2}\right), \tag{14}$$

which is a normal density with mean zero and variance $\sigma_p^2 = (\hbar^2\alpha^2)/2 = \hbar/2\sqrt{Km}$. And integrating out p in (13) we obtain precisely (12), i.e.,

$$f_2(x) = (\alpha/\pi^{1/2})\, e^{-\alpha^2 x^2}. \tag{15}$$

We note at once that the Heisenberg uncertainty relation is satisfied by the product of the standard deviations σ_p and σ_x of the two marginal distributions (14) and (15), for

$$\sigma_p \sigma_x = \frac{\hbar\alpha}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}\,\alpha} = \hbar/2 . \tag{16}$$

It is interesting to note that from the full marginal distributions we get the equality (16) as a stronger form than the usual inequality – for this special case of the harmonic oscillator in the lowest energy level. It may also be noted that the distributions of momentum and position are statistically independent, because $f_1(p) f_1(x) = f(p, x)$. *A fortiori*, the covariance and correlation coefficient are zero. This prediction of statistical independence is one which is not discussed in the usual treatments of the harmonic oscillator – in fact, I have not seen it anywhere. Consideration of this prediction illustrates one of the difficulties in analyzing the foundations of quantum mechanics. The difficulty is that it is so hard to come by an exact account of the experimental data. For example, for what classes of experiments have standard χ^2 goodness-of-fit tests been applied to test the null hypothesis that the observed data fit the quantum mechanical marginal distributions for momentum and position? Even more to the point, for the experimental setups corresponding to a potential energy $V(x)$ from which we derive a genuine *joint* distribution is it possible to collect data on this joint distribution? Admittedly my own ignorance of the experimental literature is partly responsible for these questions, but it would seem that not only the philosophical discussions of quantum mechanics, but the standard treatises as well provide inadequate answers.

B. *First Excited State*

In the first excited state, we have from the literature

$$\psi(x) = \left(\frac{4\alpha^3}{\pi^{1/2}}\right)^{1/2} x \exp(-\tfrac{1}{2}\alpha^2 x^2),$$

whence

$$|\psi(x)|^2 = \frac{4\alpha^3}{\sqrt{\pi}}\, x^2\, e^{-\alpha^2 x^2} .$$

Applying now (10) and (11), and again replacing p by the propagation vector $k=p/\hbar$, we have:

$$(17) \qquad f(k, x) = (1/2\pi)(4\alpha^3/\sqrt{\pi}) \times \int (x^2 - (u/2)^2) \exp[-\alpha^2(x^2 + (u/2)^2)]\, e^{-iku}\, du .$$

Integrating (17) we obtain

$$(18) \qquad f(k, x) = (4/\pi)[\exp(-\alpha^2x^2 - k^2/\alpha^2)](\alpha^2x^2 + k^2/\alpha^2 - \tfrac{1}{2}),$$

and the function $f(k, x)$ is negative for those values of k and x such that

$$\alpha^2x^2 + k^2/\alpha^2 < \tfrac{1}{2},$$

which means that $f(k, x)$ is not a proper joint density.

A quite different example for which $f(p, x)$ is not a proper density is given in Rubin (1959).

IV. HEISENBERG UNCERTAINTY PRINCIPLE

For the momentum and position random variables the Heisenberg uncertainty relation is the inequality

$$\sigma_x\sigma_p \geqslant \hbar/2 .$$

On the basis of the general results in the preceding section I want to urge that many of the interpretations given this inequality are mistaken.

The first thing to note is that this inequality for the product of the standard deviations of two random variables in itself tells us nothing about the process of measuring the values of these random variables. Suppose, for example, we can measure the height and the weight of any human being (at a given time) with absolute accuracy. It will still be the case that for any reasonably sized sub-population of humans the product of the standard deviations of height and weight will exceed some positive number. The Heisenberg relation is disturbing because we think of 'identically prepared' particles having the same momentum and position, as we do not in the case of the height and weight of humans. But the disturbance of classical ideas is not only this psychological one.

It is commonly said that the Heisenberg principle shows we cannot measure both momentum and position simultaneously with arbitrary

precision. As we have already remarked, there is in fact a logical gap between the relationship itself and this measurement statement. This logical gap is not, however, the main problem. The real point is that the uncertainty relation does not represent a genuine statistical relationship at all, for there does not in general exist a joint probability distribution of the momentum and position random variables. The real claim to be made is that when a proper joint distribution of momentum and position does not exist, then these two properties are not simultaneously measurable at all.

The conclusion that momentum and position are not simultaneously measurable at all does not follow from the Heisenberg relation but from the more fundamental results about the absence of a genuine joint distribution. There is no underlying sample space which may be used to represent the simultaneous measurements, exact or inexact. It is in this sense that the Heisenberg inequality is not a genuine statistical relation. On the other hand, it may be admitted that in a weaker sense it is a real statistical relation. For particles prepared in a state described by a given Hamiltonian or potential energy function, it may be possible to measure the position of some and the momentum of others. And we would expect the product of the standard deviations of these two sets of measurements to be equal to or greater than $\hbar/2$.

The physically important fact, which is not equivalent to the uncertainty relation, is that the results of the successive measurement of two noncommuting observables like momentum and position depend on the order in which the measurements are performed. Let the pair of numbers (p, x) represent in a one-dimensional case the result of measuring the momentum p and then the position x, and let (x', p') represent for an 'identically prepared' particle the results of measuring the position x' and then the momentum p'. The important thing is that in general $p \neq p'$ and $x \neq x'$. Formally this relation for noncommuting canonically conjugate observables is expressed by the fundamental commutation rule

$$PQ - QP = \hbar/i,$$

which, under the interpretation urged here, is more physically significant than the uncertainty relation $\sigma_p \sigma_x \geqslant \hbar/2$. Weyl (1931) clearly shows the importance of the commutation rules in the basic structure of quantum mechanics.

Koopman (1957) is concerned to argue that quantum mechanics provides no real evidence for changing the foundations of probability. And with this I am in complete agreement, for the simple reason that there is, so far as I know, no substantial argument for making any such changes. The use of functions like the improper joint densities of the last section is for purposes of calculation. In no sense do they help make a case for changing the basic laws of probability. Far from it, rather it may be argued that we can use their very improperness as a key to inferring what is not possible to measure experimentally. Koopman goes on to claim that the inability to measure momentum and position simultaneously is not an unusual one. He says the following:

> The standard example of *incompatibility* is in quantum mechanics, where α and β are statements concerning the same components of position and momentum. But it is possible to give a much more obvious example. Suppose that, as a result of a gamma-ray mutation, an altogether different and unique creature were produced from rat parents, and suppose that its resistance to lethal doses of certain poisons A and B is of interest. If α is a statement concerning the number of days before it dies when injected solely with A, while β is a similar statement regarding B, then α and β are clearly incompatible, for the simple reason that you can kill your creature with A alone or else with B alone, but not "with both alone." Therefore no logical combinations such as $\alpha\beta$, $\alpha+\beta$ have any meaning: the situation assumed in the definition of α is inconsistent with that needed to make β meaningful. Nor is the situation improved by restating α and β as implications (α being interpreted as that the A experiment *implies* that the creature survives so many days, etc.), simply because, in order to become experimental propositions, a single experiment ("trial") must be capable of telling whether α is true ("succeeds") or false; and similarly for β – and we are back where we were [p. 100].

His rat example is well put, but it goes too far in reducing the incompatibility of two quantum mechanical statements to what is essentially ordinary incompatibility. In his example the conjunction $\alpha\&\beta$ would ordinarily simply be regarded as false. The quantum mechanical situation is more complicated. We have two functions, the momentum and position random variables, describing two quantitative properties of some physical phenomenon – in realistic language, of some physical particle. We consider the functions to be defined for every instant of time. We then feel uneasy when we cannot talk about the values of these functions at the same time t. In a given experimental setup that we can appropriately describe by a Hamiltonian or potential energy function, it is possible to derive the marginal probability distributions of momentum and position.

Moreover, it seems possible with some slight modifications to talk about the joint distribution of the tendency of the two poisons A and B to kill

the rats in a and b days, respectively. Suppose we find, for example, that according to the amount C_i of a certain organic substance in the kidneys of the rats the number of days until death with injection of poison A is $a_i \pm 1$ and with injection of B is $b_i \pm 2$, with probability one. On the basis of observation of the amount C_i in the kidneys of a rat we could then assert that the probability of the rat having the property of dying in a_i days from poison A and the property of dying in b_i days from poison B is such and such. And by studying the distribution of the given organic substance in the kidneys of rats of a particular colony, we could go on to make unconditional probability statements about the rats in the colony having the property of dying in a days from poison A and b days from poison B. Because the effects of only one poison can be observed, the inference to the joint distribution depended upon discovering controlling factors like the organic substance in the kidney. In fact, the aim of the scientist is to investigate more and more thoroughly the chemistry of poisons in order to predict more and more accurately the effect of giving any particular one. It is, so far as I can see, completely meaningful to summarize the results of such investigations by joint probability statements. Note that given C_i, the probabilities of a_i and b_i are statistically independent, but without this information they are not.

Koopman asserts in the passage quoted above that the conjunction $\alpha \& \beta$ is meaningless. What I am claiming is that by rephrasing α and β in terms of properties or dispositions we obtain new statements α' and β' whose conjunction is perfectly meaningful. The defense of this change is the same as the standard defense against a too narrow criterion of empirical or operational meaningfulness in evaluating scientific theories. It is, I take it, now a truism that not all scientific concepts or terms are directly observable, and *a fortiori* are not simultaneously observable.

The qualitative difference between the rat experiment and a quantum mechanical one is that in the quantum mechanical case we seem to be prevented even in principle from talking about the joint distribution of momentum and position.

V. AXIOMATIC FOUNDATIONS

I conclude with some general remarks about the axiomatic foundations of classical quantum mechanics. It is widely but mistakenly thought that von

Neumann provided an exact axiomatization in his well-known book on quantum mechanics. He gives there exact axioms for Hilbert space, but he does not adjoin in exact fashion an axiomatic characterization of quantum mechanics. (By 'exact' I do not mean in logical symbolism but according to the standards of axiomatization in geometry or abstract algebra.) Perhaps the clearest axiomatization yet given is that of Mackey (1957).

Briefly speaking, Mackey proceeds in the following fashion for the time-independent case. Let θ be the set of observables and let S be the set of states; any structure on the sets θ and S is explicitly stated in the axioms. The function $p(A, \alpha, E)$ is defined whenever $A \in \theta$, $\alpha \in S$ and E is a Borel set of real numbers. Intuitively $p(A, \alpha, E)$ is the probability of measuring observable A in set E when the state of the system is α. The first axiom states in fact that for every A in θ and α in S, $p(A, \alpha, E)$ is a probability measure in the argument E on the set of all real numbers. The second axiom guarantees uniqueness of observables with a given probability distribution, and similarly for states. It is a kind of extensionality axiom for observables and states.

If $p(A, \alpha, E) = p(A', \alpha, E)$ *for all* α *in* S *and Borel sets* E *then* $A = A'$, *and if* $p(A, \alpha, E) = p(A, \alpha', E)$ *for all* A in θ *and all Borel sets* E *then* $\alpha = \alpha'$. The remaining axioms are more technical and will not be given here. Properties as two-valued observables are defined, and a certain partial ordering in terms of probability distributions on properties is defined. The final and most powerful axiom is then the assertion that the set of all properties under the given ordering is isomorphic to the partially ordered set of all closed subspaces of a separable infinite dimensional complex Hilbert space.

It is important to point out that Mackey's axiomatization is not complete, for to get a system that is adequate to analyze detailed physical examples, further axioms are needed to set up the proper correspondence between observables and operators. To solve particular problems it is by no means sufficient to know only that such a one-one correspondence exists; the correspondence must be given constructively by additional axioms.

From the standpoint of probability theory, the present formalism of quantum mechanics in terms of operators on a complex Hilbert space is not a natural one. It is apparent that no one approaching the subject from

the standpoint of stochastic processes would be in any direct way led to this formulation. Because most of the observable predictions of the theory are in fact probabilistic predictions, the question naturally arises of giving an axiomatic formulation in terms more congenial to the theory of stochastic processes. It should be clear from Section III that such a formulation is possible in terms of quasi-probability distributions, a term commonly used for the joint distribution of conjugate observables because of the possible negativity of the distribution. On the basis of Moyal (1949) and Baker (1958) such a formulation is not too difficult to give, although it is beyond the scope of the present paper to enter into the details, particularly since the discussion here has been restricted to quantum statics.

Moyal's paper demonstrates the conceptual advantages of this statistical approach in making detailed comparisons between classical and quantum statistical mechanics. From a more general philosophical standpoint it seems to me there are at least two reasons for interest in the quasi-probability distribution formulation. The first is the negative but therapeutic one of reducing the interminable discussion of the wave vs. particle interpretation of quantum mechanics. A statistical approach to the observable quantities is neutral as between these two interpretations. On another occasion I hope to recast in terms of the present statistical approach the many distinctions introduced by Reichenbach (1944) in his philosophical analysis of quantum mechanics.

The second reason for the philosophical interest of the statistical approach to quantum mechanics has already been mentioned. It provides a more natural place for probability concepts, thereby leading to a formulation in which it is easier to discuss general philosophical concepts like those of determinism and causality and to compare in a relatively direct way the role of these concepts in probabilistic theories of phenomena outside the quantum domain. I have tried to illustrate in Section IV how such a formulation can clarify the meaning of Heisenberg's uncertainty principle. Consideration of quantum dynamics in terms of quasi-probability distributions is also helpful in clarifying the status of causality in quantum mechanics, but this is a task beyond the scope of the present paper.

NOTES

[1] Henceforth the range of integration is understood to be $(-\infty, \infty)$ and notation for it is omitted.

[2] I am indebted to Sidney Drell and Raymond Willey for some helpful suggestions concerning these two examples.

14. THE ROLE OF PROBABILITY IN QUANTUM MECHANICS*

The view I attempt to support in this paper by various sorts of arguments is that if we examine the structure of quantum mechanics for paradoxical or surprising implications, it is not the interpretation of the uncertainty principle as showing the untenability of determinism that is most important. I take it that many would now agree that strict determinism was a fragile flower nurtured for a few decades in the special climate of classical mechanics, particularly as formulated by Laplace. The once broad claims for determinism have in many ways been replaced by universal claims for the theory of probability and the methodology of statistics, although I do not wish to suggest for a moment that determinism as a thesis and probability theory as a methodological cornerstone of science are on the same conceptual footing. Rather, my point is to try to show that the really radical intellectual thing about quantum mechanics is not its challenge to determinism as a philosophical thesis, but its challenge to probability theory and modern mathematical statistics as the universal methodology of all empirical science.

Let us begin by considering some of the standard interpretations of the Heisenberg uncertainty relation. Landau and Lifshitz have this to say:

> whereas in classical mechanics a particle has definite coordinates and velocity at any given instant, in quantum mechanics the situation is entirely different. If, as a result of measurement, the electron is found to have definite coordinates, then it has no definite velocity whatever. Conversely, if the electron has a definite velocity, it cannot have a definite position in space. For the simultaneous existence of the coordinates and velocity would mean the existence of a definite path, which the electron has not. Thus, in quantum mechanics, the coordinates and velocity of an electron are quantities which cannot be simultaneously measured exactly, i.e., they cannot simultaneously have definite values. We may say that the coordinates and velocity of the electron are quantities which do not exist simultaneously [1958, p. 4].

Similar views are to be found in Reichenbach. He says that the uncertainty

* Reprinted from *Philosophy of Science, The Delaware Seminar*, Vol. 2: *1962–63* (ed. by B. Baumrin), Wiley, New York, 1963, pp. 319–337.

relation ... can be interpreted in the form: When the position of the particle is well determined, the momentum is not sharply determined and vice versa [1944, p. 13].

As can be seen from these two quotations the orthodox viewpoint is to emphasize that the uncertainty relation means that when position is determined exactly the momentum is not determined at all, or when momentum is sharply determined, position is badly determined, if at all. It is particularly interesting to note that Landau and Lifshitz, and Reichenbach as well, do not give a quantitative interpretation of the uncertainty relation.

A somewhat more sophisticated characterization is given in Schiff's well-known textbook:

The relation ... means that the component of the momentum of a particle cannot be precisely specified without our loss of all knowledge of the corresponding component of its position at that time, that a particle cannot be precisely localized in a particular direction without our loss of all knowledge of its momentum component in that direction, and that in intermediate cases the product of the uncertainties of the simultaneously measurable values of corresponding position and momentum components is at least of the order of magnitude of h [1949, p. 7].

Schiff does mention the "intermediate cases" when neither position nor momentum is sharply determined, but none of the authors cited give anything like a thorough probability interpretation of the inequality. I have, for instance, seen no book on quantum mechanics in which an "intermediate case" is described in quantitative or experimental detail. What single experiment can be cited on behalf of the presumed intermediate cases, that is, in what experiment can we determine approximately, apart from other sources of experimental error, that

$$\sigma_x > 0$$
$$\sigma_p > 0$$
$$\sigma_x \sigma_p \geqslant \hbar/2$$

and more particularly, what numerical values of the standard deviations σ_x and σ_p are found? The third of the three inequalities is a standard formulation of the uncertainty principle. From its form one would naturally infer that it is possible simultaneously and jointly to determine the standard deviations σ_x and σ_p, but, as we shall see, this is not in general possible.

If we consider this inequality from the standpoint of probability theory, there seem to be no persuasive arguments whatsoever for accepting the

orthodox interpretation. To say this is not to say that the orthodox interpretation is necessarily wrong, but to emphasize that there is a very large logical gap between the claims that are ordinarily made for the meaning of the uncertainty principle and the literal interpretation of the principle itself as a statement about the product of two standard deviations. At this point it may be instructive to consider other interpretations that are possible within the framework of standard probability theory.

Perhaps the first point that needs clarification is the distinction between measurements that are exact when taken at a given instant and the probability distribution of quantitative properties of objects, when those objects are "identically prepared" for experimental manipulation. Consider the following simple example. Let H be the random variable defined on the set of all human beings such that for any person x, $H(x)$ is the height of x, and let W be the random variable measuring the weights of human beings. Now we do not need to collect any new experimental data in order to know that both of these random variables will have considerable variation in the population of human beings. This variation in itself has nothing to do with the measuring process. For purposes of any investigation in which we would be interested in the height and weight of human beings, we can assume that both these variables may be measured with absolute accuracy. It will still be the case that the product of the standard deviations of the distributions of the two random variables will exceed some fairly large positive number, in spite of the fact that for any individual x, $H(x)$ and $W(x)$ are perfectly definite. On the basis of the positive variances of both height and weight, we could paraphrase Landau and Lifshitz and say that if the height of a person has a definite value, then his weight has no definite value whatsoever. The patent absurdity of this is obvious, which is just to say that Landau and Lifshitz have explained the meaning of the uncertainty relation rather badly.

The immediate rejoinder in considering this example in relation to the Heisenberg uncertainty inequality is that of course there is a drastic difference. For we know that the members of the set of human beings are distinct individuals with quite different properties of height and weight. On the other hand, when we prepare an electron or some other "elementary" particle for experimental observation, we think of all electrons brought into the experimental situation as being in some sense identical or equivalent, at least being identical in a sense that we do not expect the

various human beings to be identical. Yet there does not seem any reason to accept this principle of identity or equivalence for elementary particles other than as a kind of prejudice derived from two hundred years of classical particle mechanics. It seems apparent that if this principle of identity is not adopted, then it is perfectly consistent to hold that in any given instant, position and momentum of an elementary particle may be measured exactly, but if the measurements are repeated on other elementary particles of the same kind under what the experimenter thinks are identical experimental conditions, then a genuine probability distribution of both kinds of measurements will be obtained, and in fact the product of the standard deviations of these two distributions will satisfy the Heisenberg uncertainty relation.

This is the kind of situation that obtains in a wide range of psychological experiments. It may be useful to describe in detail one kind of such experiment that has a number of affinities from a methodological standpoint with the issues we are discussing in connection with the uncertainty principle. I have in mind a simple paired-associate experiment. This simple example from mathematical behavior theory conjoins four properties that collectively challenge overly simple analyses of the uncertainty relation: (1) homogeneity or 'identity' of items is assumed; (2) a detailed theory of the learning phenomena is given; (3) the measurements of the trial of last error are absolutely exact; (4) a positive variance for the distribution of last errors is predicted. The orthodox interpretation of quantum mechanics needs to make clearer why a similar conjunction of properties could not be postulated in quantum mechanics. On this point I must be absolutely clear. I am not saying that an interpretation of quantum mechanics in terms of exact joint measurements of position and momentum is correct. We need to be told in more detail why it is wrong. (It is the burden of the latter part of this paper to show it must in fact be rejected for deeper reasons than those provided by the uncertainty relation.)

In our simple paired-associate experiment, the task for the subject is to learn to associate each one of a list of nonsense syllables with an appropriate response. In a typical setup the list might consist of 20 nonsense syllables of the form cvc (consonant, vowel, consonant). The responses are given by pressing one of two keys. On a random basis, ten of the syllables are assigned to key 1 and ten to key 2. The subject is

shown each nonsense syllable in turn, is asked to make a response, and is then shown the correct response by one of several devices, for example, by the illumination of a small light above the correct key. After the subject has proceeded through the list once, he is taken through the list a second time, but the order or presentation of the 20 items is randomized. A criterion of learning is set; for example, four times through the list without a mistake. The subject is asked to continue to respond until he satisfies this criterion. The criterion is selected so as to give substantial evidence that the subject has indeed learned the correct association between each stimulus item and its appropriate response. There are a number of statistics to be derived from a simple and very adequate model for these phenomena. For the present purpose it will be sufficient to concentrate on the distribution of the trial of last error.

The simple model to be applied to the phenomena is the following. The subject begins the experiment by not knowing the arbitrary association established by the experimenter between individual stimuli and the response keys. He is thus in the unconditioned state U. On each trial there is a probability c that he will pass from the unconditioned state to the conditioned state C. It is postulated that this probability c is constant over trials and independent of responses on preceding trials. Once the subject passes into the conditioned state it is also postulated he remains there for the balance of the experiment. A simple transition matrix for the model, which is a first-order Markov chain in the two states U and C, is the following:

$$\begin{array}{c|cc} & C & U \\ \hline C & 1 & 0 \\ U & c & 1-c. \end{array}$$

To complete the model for the analysis of experimental data it is also necessary to state what the probabilities of response are in the two states U and C. When the subject is in the unconditioned state, it is postulated that there is a guessing probability p of making a correct response, and this guessing probability is independent of the trial number and the preceding pattern of responses. When the subject is in the conditioned state, the probability of making a correct response is postulated to be 1. With these assumptions it may be shown that the distribution of the trials on which state C is entered is geometric, where c is the parameter of

the geometric distribution. We recall that the mean of the geometric distribution with parameter c is $1/c$ and its variance is the quantity $(1-c)/c^2$. Unfortunately, it is not possible to observe when the subject enters state C, for he may or may not have guessed the correct responses on immediately preceding trials. What is behaviorally observable is the trial number of the last error on each item. This distribution is, as would be expected, approximately geometric. The exact distribution is as follows. Let n be any trial. Then the probability $f(n)$ that the last error with respect to responses to a given stimulus occurs on trial n is:

$$f(n) = \begin{cases} bp & \text{for } n = 0 \\ b(1-p)(1-c)^{n-1} & \text{for } n > 0, \end{cases}$$

where $b = c/[1-p(1-c)]$. The mean, μ, and variance, σ^2, of this distribution are:

$$\mu = b(1-p)/c^2$$
$$\sigma^2 = \mu[2/c - 1 - \mu].$$

(Note that if $p=0$, we just obtain the geometric distribution.) For our purposes the important thing about this example is that the variance of the trial of last error is certainly not zero, in any except degenerate cases. It is important to realize that this positive variance is not due to any hypothesis of individual differences. The hypothesis of individual differences among subjects or among stimulus items would be brought into the theory by postulating variation in the parameter c and possibly the guessing parameter p. The assumption of a constant c and a constant p either for all stimulus items or a group of subjects is equivalent to assuming homogeneity of items or of subjects. In spite of this assumption of homogeneity, the predicted behavior has a positive variance. The experimenter attempts to select items for a given subject as carefully as he can in terms of empirical criteria of homogeneity. No matter how carefully such items are selected, a wide variety of studies shows that a distribution of last errors with a positive variance is always obtained in such paired-associate experiments.

In connection with the earlier quotations about the interpretation of the uncertainty relation, it should be remarked that in the case of the empirical observation of the distribution of trials of last error in a given experiment there is no problem whatsoever concerning errors of measure-

ment. The measurements are absolutely exact. They are discrete observations, for which it is necessary only to record which of two keys was pressed by the subject on a given trial. In other words, the positive variance and thus the uncertainty in the trial of last error are in no sense due to the measurement procedure.

Let us try transferring these results on exact measurements with a positive variance to a quantum-mechanical context. As a simple example we may consider a one-dimensional harmonic oscillator in the ground state. It may be shown that the theoretical distribution $f(x)$ of position for this physical system is given by the following expression:

$$f(x) = (\alpha/\pi^{1/2})\, e^{-\alpha^2 x^2}.$$

This expression is that for a normal density with mean zero and variance $\sigma^2 = \alpha^2/2$, where α is a physical constant defined in terms of $\hbar$ and mass m. More precisely,

$$\sigma^2 = \hbar/2\sqrt{Km},$$

where K is the constant arising in the expression for the potential energy: $V(x) = \frac{1}{2}Kx^2$. This distribution with its positive variance for position is derived from the theory in quite straightforward terms. The important point at the moment is that this derivation in no way mentions any processes of measurement. An obvious empirical interpretation of the result is that if measurements of position are made for a harmonic oscillator in the ground state, then the obtained measurements will fit a normal distribution with mean zero and variance as indicated. There is nothing as such in the theory that says the procedures of measurement themselves give rise to the distribution.

A related point of interpretation of the uncertainty relation is the following. The standard verbal formulations or interpretations imply, or tend to imply, that the experimenter may at will decide to measure a particular physical variable with arbitrary precision. Of course, if he does so he must sacrifice corresponding precision in conjugate physical variables. But the example of the linear oscillator in the ground state suggests that this interpretation is wrong. For a given energy state of a system that is a linear oscillator it is not possible to make a precise measurement of position or momentum. In the case of the oscillator in the ground state, all that can be obtained, no matter how exact the

experimental procedure, is a normal distribution with mean zero and positive variance as indicated for the position, and similar results obtain for the momentum.

The tenor of these remarks suggests that I am proposing it may be consistent with the fundamental theory of quantum mechanics to measure position and momentum exactly on any given occasion. As the measurements are repeated on identically prepared particles, distributions with positive variance will of course be obtained, and these will satisfy the uncertainty relation. This state of affairs would fit in very well with ordinary probability concepts. Unfortunately, it is not a possible interpretation of the uncertainty relation. Such an interpretation requires that there be a *joint* probability distribution of position and momentum at any given instant. However, if the usual formalism is applied to the computation of the joint distribution of two conjugate physical variables like position and momentum, then the distribution obtained by standard arguments is in general not a *proper* probability distribution at all. This result raises serious difficulties for any attempt to interpret quantum mechanics within a standard probability framework. It is extremely hard to think of any other empirical examples of physical variables having this property of not possessing a genuine joint probability distribution.

In order to bring out the full significance of having a joint distribution of position and momentum, it will be useful first to consider a much simpler case that will require no detailed computations of any sort. Suppose that we have two fair coins and we are flipping them in the following two ways. In the first case, we flip the first coin and then whatever comes up, either heads or tails, we automatically turn the second coin so that the same side is facing up. In the second case, we flip the two coins at the same time and assume that the physical handling of the coins is such that the outcomes are statistically independent.

First of all, then, let us write down the joint distribution for these two different cases. We shall use a subscript 1 for the first case and a subscript 2 for the second case, and we shall use a subscript a for the first coin and a subscript b for the second coin. In the first case the four possible outcomes have the following probabilities

$$P_1(H_a, H_b) = \tfrac{1}{2} \quad P_1(T_a, H_b) = 0$$
$$P_1(H_a, T_b) = 0 \quad P_1(T_a, T_b) = \tfrac{1}{2}.$$

Because of the physical conditions of this case, it is clear that the probability must be zero of having H_a and T_b, or also T_a and H_b.

In the second case, the joint distribution of heads and tails of the two coins assumes quite a different form. On the assumption of statistical independence, we have at once that the probability of any two outcomes, for example, H_a and T_b, is simply one-fourth:

$$P_2(H_a, H_b) = \tfrac{1}{4} \quad P_2(T_a, H_b) = \tfrac{1}{4}$$
$$P_2(H_a, T_b) = \tfrac{1}{4} \quad P_2(T_a, T_b) = \tfrac{1}{4}.$$

In spite of the clear difference between these two joint distributions, that is, P_1 and P_2, there are a number of probability statements that are the same for both of them. In particular, the two marginal distributions for coins a and b are precisely the same. By marginal distributions we mean just the probability statements about the single coins. Thus, we have

$$P_1(H_a) = P_1(H_b) = P_2(H_a) = P_2(H_b) = \tfrac{1}{2}$$
$$P_1(T_a) = P_1(T_b) = P_2(T_a) = P_2(T_b) = \tfrac{1}{2}.$$

Another way of putting it is that if we are able only to observe one coin at a time, that is, either coin a or coin b, we are not able to infer from such observations which of the two joint distributions is the true one. As this example clearly illustrates, the marginal distributions tell us a great deal less about the behavior of a "system" of two coins than does the joint distribution.

To obtain some numerical results similar in spirit, but of course much simpler than those we obtain in quantum mechanics in dealing with such physical variables as position and momentum, we may define for each coin a random variable. This random variable is a function defined on the two possible outcomes, heads or tails, and for each coin the random variable has the value 1 when the outcome is heads and the value zero when the outcome is tails. We shall use X_a for the random variable corresponding to coin a, and X_b for the random variable corresponding to coin b. As would be expected from what we have said about the identity of the marginal distributions for the two different joint distributions P_1 and P_2, the expected values or means $E(X_a)$ and $E(X_b)$ are the same for P_1 and P_2. Namely,

$$E_1(X_a) = E_1(X_b) = E_2(X_a) = E_2(X_b) = \tfrac{1}{2}.$$

Similar results obtain for any other quantities we want to compute for X_a and X_b separately, that is, the quantities are the same for the two distributions P_1 and P_2. For example, the variances of X_a and X_b have the following form:

$$\mathrm{Var}_i(X_a) = \mathrm{Var}_i(X_b) = \tfrac{1}{2}(1 - \tfrac{1}{2}) = \tfrac{1}{4}, \quad \text{for} \quad i = 1, 2.$$

But once we move to any considerations involving both variables at the same time, the results are quite different. For example, consider the covariance of X_a and X_b, defined as follows:

$$\mathrm{Cov}(X_a, X_b) = \sum_{x_a} \sum_{x_b} (x_a - \bar{x}_a)(x_b - \bar{x}_b) P(X_a = x_a, X_b = x_b),$$

where $\bar{x}_a = E(X_a)$ and $\bar{x}_b = E(X_b)$. The covariance has the following distinct values for P_1 and P_2.

$$\begin{aligned}\mathrm{Cov}_1(X_a, X_b) = {} & (1 - \tfrac{1}{2})(1 - \tfrac{1}{2}) \cdot \tfrac{1}{2} \\ & + (1 - \tfrac{1}{2})(0 - \tfrac{1}{2}) \cdot 0 \\ & + (0 - \tfrac{1}{2})(1 - \tfrac{1}{2}) \cdot 0 \\ & + (0 - \tfrac{1}{2})(0 - \tfrac{1}{2}) \cdot \tfrac{1}{2} = \tfrac{1}{4}.\end{aligned}$$

And

$$\begin{aligned}\mathrm{Cov}_2(X_a, X_b) = {} & (1 - \tfrac{1}{2})(1 - \tfrac{1}{2}) \cdot \tfrac{1}{4} \\ & + (1 - \tfrac{1}{2})(0 - \tfrac{1}{2}) \cdot \tfrac{1}{4} \\ & + (0 - \tfrac{1}{2})(1 - \tfrac{1}{2}) \cdot \tfrac{1}{4} \\ & + (0 - \tfrac{1}{2})(0 - \tfrac{1}{2}) \cdot \tfrac{1}{4} = 0.\end{aligned}$$

Similarly, the correlation coefficient ρ of X_a and X_b is defined in general as follows

$$\rho(X_a, X_b) = \frac{\mathrm{Cov}(X_a, X_b)}{\sigma(X_a)\,\sigma(X_b)}$$

and from the calculations already made, it is clear that for the joint distribution P_1, $\rho_1(X_a, X_b) = 1$ and for the joint distribution P_2, $\rho_2(X_a, X_b) = 0$. The covariance and correlation are measures of the dependency relationship holding between two events or random variables. When the events are independent, the covariance and correlation coefficient are both zero, as is the case for X_a and X_b under distribution P_2.

These computations for the simple random variables we have defined correspond to computations for expected positions and the like in

quantum mechanics. It will be useful now to look at some corresponding results for a simple physical case in quantum mechanics. I shall not here justify the results stated, but I have indicated in a previous paper (Suppes, 1961b)[1†] how they may be derived in a straightforward fashion from the usual formulations of classical quantum mechanics. For purposes of discussion, I selected one of the simplest cases possible, namely, the one-dimensional harmonic oscillator already mentioned. By applying the usual methods of calculation, we may show that in the ground state the harmonic oscillator does indeed have a genuine joint distribution of position and momentum, given by the expression

$$(1) \qquad f(p, x) = (1/\hbar\pi) \exp(-\alpha^2 x^2 - p^2/\hbar^2\alpha^2),$$

where α is the familiar physical coefficient defined earlier. Integrating out x, we obtain for the marginal distribution of momentum

$$f(p) = \frac{1}{\alpha\hbar\sqrt{\pi}} \exp\left(-\frac{p^2}{\hbar^2\alpha^2}\right),$$

which is a normal density with mean zero and variance $\sigma_p^2 = \hbar^2\alpha^2/2$. Integrating out p, we obtain for the marginal distribution of position

$$f(x) = (\alpha/\sqrt{\pi}) \exp(-\alpha^2 x^2),$$

which is the normal density with mean zero and variance $\sigma_x^2 = 1/2\alpha^2$, previously mentioned. Then we know at once that the Heisenberg uncertainty relation is satisfied by the product of the standard deviations σ_p and σ_x of the two marginal distributions:

$$\sigma_p\sigma_x = \frac{\hbar\alpha}{\sqrt{2}} \cdot \frac{1}{\sqrt{2\alpha}} = \frac{\hbar}{2}.$$

It is perhaps worth noting that in the present case we get an actual equality rather than the weak inequality of the general case. Also of particular interest is the observation that the joint distribution (1) shows that the distributions of momentum and position are statistically independent because $f(p)f(x) = f(p, x)$. And it of course follows from this independence result that the covariance and correlation coefficient of position and momentum are zero.

The situation is quite different when we turn to the first excited state of a harmonic oscillator. For simplicity, let us replace the momentum p by

the propagation vector $k = p\hbar$. We then have for the joint density of k and x, the following expression:

$$f(k, x) = (4/\pi)[\exp(-\alpha^2 x^2 - k^2/\alpha^2)][\alpha^2 x^2 + k^2/\alpha^2 - 1/2].$$

Unfortunately, the function f in this case is not a genuine probability density. In particular, the function is negative for those values of k and x such that

$$\alpha^2 x^2 + k^2/\alpha^2 < 1/2.$$

Similar results are obtained for other states of the harmonic oscillator; namely, the joint distributions obtained by the usual arguments are not genuine distributions at all. This situation is characteristic in quantum mechanics. It is only the unusual or special case for which a genuine joint distribution is obtained. From a statistical standpoint, the covariance or correlation of position and momentum would be at least as natural and important to consider as the product of their standard deviations, but when their joint distribution does not exist we cannot in a meaningful way talk of their covariance or correlation.

The strangeness of these results from a methodological standpoint is difficult to overemphasize. As we all know, the applications of probability theory range over all domains of science, and have become increasingly important in the past several decades. The mathematical techniques of probability theory, as well as the conceptual foundations, have received an enormous amount of attention from mathematicians, statisticians, and philosophers. In the several domains of application with which I am familiar or which I have at least cursorily inspected I have not been able to find a single example having the conceptual status of these results about the nonexistence of a joint distribution in quantum mechanics. Moreover, the idea that two physical variables that we are able to observe even separately must have a joint probability distribution is a very deep and natural one from the standpoint of probability theory. It may, of course, not be the case that we can make direct observations on this joint distribution, but what is unusual and dismaying is to find a case in which the postulate of the mere *existence* of a joint distribution is inconsistent with the fundamental theory of the phenomena.

Those of you among the audience who are not already familiar with these results will, roughly speaking, probably have one of two reactions.

Those whose backgrounds are in physics will tend to accept the nonexistence of the joint distribution of conjugate physical variables with scarcely a murmur, and say that here is simply another piece of evidence of the inadequacy of classical ideas. I shall have more to say in a moment about this rather casual attitude toward probability theory. On the other hand, those who are more thoroughly acquainted with probability theory than quantum mechanics will undoubtedly tend to think initially that it is perhaps a simple and not very deep matter to change the quantum-mechanical computations in such a way as to obtain genuine distributions. In the present lecture I have not attempted to examine this rather technical point, or, in fact, even to substantiate my own computations in any mathematical detail. I do, however, offer assurances that the framework within which I have performed these computations is perfectly standard and derives from classical and fundamental work by Wigner (1932) and Moyal (1949). It is true that other approaches to probability questions within classical quantum mechanics can be found in the literature. But generally speaking, these lead to even more strange results. A typical example is to be found in Dirac's paper (1945), in which he formulates a theory that leads to probability distributions that are complex valued. It need hardly be said that no direct empirical significance from a probability standpoint can be attached to such mathematical functions. It would, in fact, seem to be a complete misnomer to call them probability distributions in any sense.

The recent proposals of Margenau (1963) and Margenau and Hill (1961) to avoid these difficulties by a bold redefinition of the joint probability distribution are clearly inconsistent with the standard formulations of quantum mechanics. Reasons for skepticism about their proposals are too technical to enter into here. Suffice it to say that it is seldom possible to redefine a derived quantity like that of a joint distribution in the context of a complex theory and hope to end up with a complete theory that is both mathematically exact and consistent.

It is doubtful that any simple way can be found of avoiding the perplexing and paradoxical problems that arise in quantum mechanics, when probability notions are developed with anything even faintly approaching the thoroughness with which they are used in other disciplines. For those familiar with the applications of probability and mathematical statistics in mathematical psychology or mathematical

economics, it is surprising indeed to read the treatments of probability even in the most respected texts of quantum mechanics. I do not mean by this remark to belittle in any sense the magnificent and profound accomplishments embodied in quantum mechanics. These accomplishments, as well as the subtlety of the kind of results achieved, we surely all recognize. What is surprising is that the level of treatment both in terms of mathematical clarity and in terms of mathematical depth and detail is surprisingly low. Probability concepts even have a strange and awkward appearance in quantum mechanics, as if they had been brought within the framework of the theory only as an afterthought and with apology for their inclusion. I cannot, for instance, recall reading a single book or article on quantum mechanics by a physicist which uses the fundamental notion of a random variable in an explicit manner, although this notion is central to every modern treatise on probability theory. That many physicists still find it difficult to feel at home with probability notions is well attested to by the large number of people who prefer the more classical physical language of wave mechanics to talk about probability distributions and expected values. The intellectual tension that exists between the widely applicable methodology of modern probability and statistical theory on the one hand, and quantum mechanics as the boldest and most important scientific theory of the twentieth century, with its peculiar and strange use of probability – this tension will surely receive a different final resolution than is now in view.

If my diagnosis of the antinomy between probability theory and quantum mechanics, exemplified in the nonexistence of the joint probability distribution of conjugate physical variables, is correct, then it is very likely also correct to say that we are confronted with one of those fundamental conflicts that have in the past been so important to the history of mathematics and science. Perhaps the oldest clear example of such a conflict is to be found in the discovery of the irrational by the Pythagoreans. The Pythagorean philosophers and mathematicians found that the ratio of the diagonal to the side of a square could not be expressed by a "number," that is, by what mathematicians would now call a rational number – a number which is the ratio of two integers. We can, if we like, regard the Pythagorean derivation of the irrationality or incommensurability of the diagonal of the unit square as the derivation of a contradiction within the mathematical framework accepted at that time.

In the same way, we could by adding plausible postulates to the usual methodology of probability theory construe the nonexistence of a joint distribution of conjugate physical variables as the derivation of a contradiction from the joint assumptions of probability theory and quantum mechanics. The decision as to whether we are to regard the conflict as one issuing in a formal contradiction is not of critical importance. What is of fundamental importance is the recognition that two of the most powerfully entrenched ways of looking at the world lead to a fundamental conceptual conflict.

What happened in the Pythagorean case of the discovery of incommensurable line segments is well known. The fundamental concept of number was changed and extended to include irrational numbers, but the chronicle of this change and the acceptance of irrational numbers is a long and torturous one in the history of mathematics. The Greeks themselves did not really directly accept it, but in fact, rejected numerical algebra as an adequate instrument for the study of geometrical problems. The theory of proportion of Theatetus and Eudoxus, beautiful as it is from a mathematical standpoint, is clearly a geometrical escape. It was not really until many centuries later, due to the work of Weierstrass, Dedekind, and others in the nineteenth century, that the theory of real numbers, in particular of irrational numbers, was put on a sound mathematical basis as a theory in its own right.

It may be claimed that I have exaggerated the conflict between probability theory and quantum mechanics. Those who wish to support such a claim will emphasize that the problem of the nonexistence of the joint distribution of conjugate variables arises in a natural way out of the quantum-mechanical theory of measurement. I have not in this paper said a great deal about the difficult and numerous problems of measurement in quantum mechanics. For the purposes of the present point, I think it is sufficient to mention, however, that it is not the theory of measurement *per se* that can be made to assume the burden of the conflict. For the kind of peculiar probability results derivable from quantum-mechanical theory exist nowhere else in the theory of measurement. Secondly, these nonexistence results do not depend on any specific assumptions about the processes of measurement, but follow in a natural and simple way from the fundamental theory.

As I have already said, it is certainly possible to avoid any implication

of a formal contradiction from the nonexistence of a joint distribution of conjugate variables. What I think is much more difficult to avoid is the sense of tension and instability in the confrontation of quantum mechanics and probability theory. The existence of this conflict will surely lead to further fundamental conceptual changes in our basic scientific view of the world.

NOTE

[1†] Article 13 in this volume.

15. THE PROBABILISTIC ARGUMENT FOR A NONCLASSICAL LOGIC OF QUANTUM MECHANICS*

I. THE ARGUMENT

The aim of this paper is simple. I want to state as clearly as possible, without a long discursion into technical questions, what I consider to be the single most powerful argument for use of a nonclassical logic in quantum mechanics. There is a very large mathematical and philosophical literature on the logic of quantum mechanics, but almost without exception, this literature provides a very poor intuitive justification for considering a nonclassical logic in the first place. A classical example in the mathematical literature is the famous article by Birkhoff and von Neumann (1936). Although Birkhoff and von Neumann pursue in depth development of properties of lattices and projective geometries that are relevant to the logic of quantum mechanics, they devote less than a third of a page (p. 831) to the physical reasons for considering such lattices. Moreover, the few lines they do devote are far from clear. The philosophical literature is just as bad on this point. One of the better known philosophical discussions on these matters is that found in the last chapter of Reichenbach's book (1944) on the foundations of quantum mechanics. Reichenbach offers a three-valued truth-functional logic which seems to have little relevance to quantum-mechanical statements of either a theoretical or experimental nature. What Reichenbach particularly fails to show is how the three-valued logic he proposes has any functional role in the theoretical development of quantum mechanics. It is in fact fairly easy to show that the logic he proposes could not possibly be adequate for a systematic theoretical statement of the theory as it is ordinarily conceived. The reasons for this will become clear later on in the present paper.

The main premises of the argument I outline in this paper are few in number. I state them at this point without detailed justification in order to give the broad outline of the argument the simplest possible form.

* Reprinted from *Philosophy of Science* **33** (1966), 14–21.

PREMISE 1: *In physical or empirical contexts involving the application of probability theory as a mathematical discipline, the functional or working logic of importance is the logic of the events or propositions to which probability is assigned, not the logic of qualitative or intuitive statements to be made about the mathematically formulated theory.* (In the classical applications of probability theory, this logic of events is a Boolean algebra of sets; for technical reasons that are unimportant here this Boolean algebra is usually assumed to be countably additive, i.e., a σ-algebra.)

PREMISE 2: *The algebra of events should satisfy the requirement that a probability is assigned to every event or element of the algebra.*

PREMISE 3: *In the case of quantum mechanics probabilities may be assigned to events such as position in a certain region or momentum within given limits, but the probability of the conjunction of two such events does not necessarily exist.*

CONCLUSION: *The functional or working logic of quantum mechanics is not classical.*

From a scientific standpoint the conclusion from the premises is weak. All that is asserted is that the functional logic of quantum mechanics is not classical, which means that the algebra of events is not a Boolean algebra. Nothing is said about what the logic of quantum mechanics *is*. That question will be considered shortly. First I want to make certain that the support for the premises stated is clear, as well as the argument leading from the premises to the conclusion.

Concerning the first premise, the arguments in support of it are several. A source of considerable confusion in the discussion of the logic of quantum mechanics has been characterization of the class of statements whose logic is being discussed. On the one hand we are presented with the phenomenon that quantum mechanics is a branch of physics that uses highly developed mathematical tools, and on the other hand, discussions of logic deal with the foundations of mathematics itself. It is usually difficult to see the relation between characterization of the sentential connectives that seem appropriate for a new logic and the many mathematical concepts of an advanced character that must be available for actual work in quantum mechanics. The problem has often been posed as how can one consider changing the logic of quantum mechanics when the mathematics used in quantum mechanics depends in such a thorough fashion on classical logic. The point of this first premise is to narrow and

sharpen the focus of the discussion of the logic of an empirical science. As in the case of quantum mechanics, we shall take it for granted that probability theory is involved in the mathematical statement of the theory. In every such case a logic of events is required as an underpinning for the probability theory. The structure of the algebra of events expresses in an exact way the logical structure of the theory itself.

Concerning the second premise the arguments for insisting that a probability may be assigned to every event in the algebra is already a part of classical probability theory. It is only for this reason that one considers an algebra, or σ-algebra, of sets as the basis for classical probability theory. If it were permitted to have events to which probabilities could not be attached, then we could always take as the appropriate algebra the set of all subsets of the basic sample space. The doctrine that the algebra of events must have the property asserted in the second premise is too deeply embedded in classical probability theory to need additional argument here. One may say that the whole point of making explicit the algebra of events is just to make explicit those sets to which probabilities may indeed be assigned. It would make no sense to have an algebra of events that was not the entire family of subsets of the given sample space and yet not be able to assign a probability to each event in the algebra.

Concerning the third premise it is straightforward to show that the algebra of events in quantum mechanics cannot be closed under conjunction or intersection of events. The event of a particle's being in a certain region of space is well defined in all treatments of classical quantum mechanics. The same is true of the event of the particle's momentum's being in a certain region as well. If the algebra of events were a Boolean algebra we could then ask at once for the probability of the event consisting of the conjunction of the first two, that is, the event of the particle's being in a certain region at a given time t and also having its momentum lying in a certain interval at the same time t. What may be shown is that the probability of such a joint event does not exist in the classical theory. The argument goes back to Wigner (1932), and I have tried to make it in as simple and direct a fashion as possible in Suppes (1961b).[1†] The detailed argument shall not be repeated here. Its main line of development is completely straightforward. In the standard formalism, we may compute the expectation of an operator when the quantum-mechanical system is in a given state. In the present case the operator we

choose is the usual one for obtaining the characteristic function of a probability distribution of two variables. Having obtained the characteristic function we then invert it by the usual Fourier methods. Inversion should yield the density corresponding to the joint probability distribution of position and momentum. It turns out that for most states of any quantum-mechanical system the resulting density function is not the density function of any genuine joint probability distribution. We conclude that in general the joint distribution of two random variables like position and momentum does not exist in quantum mechanics and, consequently, we cannot talk about the conjunction of two events defined in terms of these two random variables. From the standpoint of the logic of science, the fundamental character of this result is at a much deeper level than the uncertainty principle itself, for there is nothing in the uncertainty principle as ordinarily formulated that runs counter to classical probability theory.

The inference from the three premises to the conclusion is straightforward enough hardly to need comment. From premise (1) we infer that the functional logic of events is the formal algebra of events on which a probability measure is defined. According to premise (2) every element, i.e., event, of the algebra must be assigned a probability. According to premise (3) the algebra of events in quantum mechanics cannot be closed under the conjunction of events and satisfy premise (2). Hence the algebra of events in quantum mechanics is not a Boolean algebra, because every Boolean algebra is closed under conjunction. Whence according to premise (1) the functional logic of quantum mechanics is not a Boolean algebra and thus is not classical.

II. THE LOGIC

Although the conclusion of the argument was just the negative statement that the logic of quantum mechanics is not classical, a great deal more can be said on the positive side about the sort of logic that does seem appropriate. To begin with it will be useful to record the familiar definition of an algebra, and σ-algebra, of sets.

DEFINITION 1: *Let X be a nonempty set. $\mathscr{F}$ is a classical algebra of sets on X if and only if $\mathscr{F}$ is a nonempty family of subsets of X and for every A and B in $\mathscr{F}$:*

1. $\sim A \in \mathscr{F}$.
2. $A \cup B \in \mathscr{F}$.

Moreover, if $\mathscr{F}$ is closed under countable unions, that is, if for $A_1, A_2, \ldots, A_n, \ldots \in \mathscr{F}$,

$$\bigcup_{i=1}^{\infty} A_i \in \mathscr{F},$$

then $\mathscr{F}$ is a classical σ-algebra on X.

It is then standard to use the concepts of Definition 1 in defining the concept of a classical probability space. In this definition we assume that the set-theoretical structure of X, $\mathscr{F}$ and P is familiar; in particular, that X is a nonempty set, $\mathscr{F}$ a family of subsets of X and P a real-valued function defined on $\mathscr{F}$.

DEFINITION 2: *A structure $\mathscr{X} = \langle X, \mathscr{F}, P \rangle$ is a finitely additive classical probability space if and only if for every A and B in $\mathscr{F}$:*

P1. *$\mathscr{F}$ is a classical algebra of sets on X;*
P2. *$P(A) \geqslant 0$;*
P3. *$P(X) = 1$;*
P4. *If $A \cap B = 0$, then $P(A \cup B) = P(A) + P(B)$.*

Moreover, $\mathscr{X}$ is a classical probability space (without restriction to finite additivity) if the following two axioms are also satisfied:

P5. *$\mathscr{F}$ is a σ-algebra of sets on $\mathscr{X}$;*
P6. *If $A_1, A_2, \ldots$, is a sequence of pairwise incompatible events in $\mathscr{F}$, i.e., $A_i \cap A_j = 0$ for $i \neq j$, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

In modifying the classical structures characterized in Definitions 1 and 2 to account for the truculent "facts" of quantum mechanics, there are a few relatively arbitrary choice points. One of them needs to be described in order to explain an aspect of the structures soon to be defined. I pointed out earlier that the joint probability of two events does not necessarily exist in quantum mechanics. A more particular question concerns the joint probability of two disjoint events. In this case there is

no possibility of observing both of them, since the very structure of the algebra of events rules this out. On the other hand, it is theoretically convenient to include the union of two such events in the algebra of sets, or a denumerable sequence of pairwise disjoint events, in the case of a σ-algebra. This liberal attitude toward the concept of event has been adopted here, but it should be noted that it would be possible to take a stricter attitude without affecting the concept of an observable in any important way. (This stricter attitude is taken by Kochen and Specker, 1965, but they also deliberately exclude all probability questions in their consideration of the logic of quantum mechanics.)

So the logic of quantum mechanics developed here permits the union of disjoint events apart from any question of noncommuting random variables' being involved in their definition. A more detailed discussion of this point may be found in Suppes (1965b). Roughly speaking, the definitions that follow express the idea that the probability distribution of a single quantum-mechanical random variable is classical, and the deviations arise only when several random variables or different kinds of events are considered.

The approach embodied in Definition 3 follows Varadarajan (1962); it differs in that Varadarajan does not consider an algebra of sets, but only the abstract algebra.

DEFINITION 3: *Let X be a nonempty set. $\mathscr{F}$ is a quantum-mechanical algebra of sets on X if and only if $\mathscr{F}$ is a nonempty family of subsets of X and for every A and B in $\mathscr{F}$:*

1. $\sim A \in \mathscr{F}$;
2. *If* $A \cap B = 0$ *then* $A \cup B \in \mathscr{F}$.

Moreover, if $\mathscr{F}$ is closed under countable unions of pairwise disjoint sets, that is, if $A_1, A_2, \ldots$ is a sequence of elements of $\mathscr{F}$ such that for $i \neq j$, $A_i \cap A_j = 0$

$$\bigcup_{i=1}^{\infty} A_i \in \mathscr{F},$$

then $\mathscr{F}$ is a quantum-mechanical σ-algebra of sets.

The following elementary theorem is trivial.

THEOREM 1: *If $\mathscr{F}$ is a classical algebra (or σ-algebra) of sets on X then $\mathscr{F}$ is also a quantum-mechanical algebra (or σ-algebra) of sets on X.*

The significance of Theorem 1 is apparent. It shows that the concept of a quantum-mechanical algebra of sets is a strictly weaker concept than that of a classical algebra of sets. This is not surprising in view of the breakdown of joint probability distributions in quantum mechanics. We cannot expect to say as much, and the underlying logical structure of our probability spaces reflects this restriction.

It is hardly necessary to repeat the definition of probability spaces, because the only thing that changes is the condition on the algebra $\mathscr{F}$, but in the interest of completeness and explicitness it shall be given.

DEFINITION 4: *A structure* $\mathscr{X}=\langle X, \mathscr{F}, P\rangle$ *is a finitely additive quantum-mechanical probability space if and only if for every* A *and* B *in* $\mathscr{F}$*:*

P1. $\mathscr{F}$ *is a quantum-mechanical algebra of sets on* X*;*

P2. $P(A)\geqslant 0$*;*

P3. $P(X)=1$*;*

P4. *If* $A\cap B=0$, then $P(A\cup B)=P(A)+P(B)$.

Moreover, X *is a quantum-mechanical probability space (without restriction to finite additivity) if the following two axioms are also satisfied:*

P5. $\mathscr{F}$ *is a quantum-mechanical* σ*-algebra of sets on* X*;*

P6. *If* $A_1, A_2, \ldots,$ *is a sequence of pairwise incompatible events in* $\mathscr{F}$, i.e., $A_i\cap A_j=0$ *for* $i=j$*, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right)=\sum_{i=1}^{\infty} P(A_i).$$

It is evident from the close similarity between Definitions 2 and 4 that we have as an immediate consequence of Theorem 1 the following result:

THEOREM 2: *Every classical probability space is also a quantum-mechanical probability space.*

It goes without saying that in the case of both of these theorems it is easy to give counterexamples to show that their converses do not hold.

Quantum-mechanical probability spaces can be used as the basis for an axiomatic development of classical quantum mechanics, but the restriction to algebras of sets in order to stress the analogy to classical probability spaces is too severe. The spaces defined are adequate for developing the theory of all observables that may be defined in terms of position and momentum, but not for the more general theory. The fundamental characteristic of the general theory is that not every quantum-

mechanical algebra may be embedded in a Boolean algebra, and thus is not isomorphic to a quantum-mechanical algebra of sets, because every such algebra of sets is obviously embeddable in the Boolean algebra of the set of all subsets of X.

It is thus natural to consider the abstract analogue of Definition 3 and define the general concept of a quantum-mechanical algebra. (The axioms given here simplify those in Suppes, 1965b, which are in turn based on Varadarajan, 1962.) Let A be a non-empty set, corresponding to the family $\mathscr{F}$ of Definition 2, let $\leqslant$ be a binary relation on A – the relation $\leqslant$ is the abstract analogue of set inclusion, let $'$ be a unary operation on A – the operation $'$ is the abstract analogue of set complementation, and let 1 be an element of A – the element 1 is the abstract analogue of the sample space X. We then have:

DEFINITION 5: *A structure* $\mathfrak{A}=\langle A, \leqslant, ', 1\rangle$ *is a quantum-mechanical algebra if and only if the following axioms are satisfied for every a, b and c in A:*

1. $a \leqslant a$;
2. *If* $a \leqslant b$ *and* $b \leqslant a$ *then* $a=b$;
3. *If* $a \leqslant b$ *and* $b \leqslant c$ *then* $a \leqslant c$;
4. *If* $a \leqslant b$ *then* $b' \leqslant a'$;
5. $(a')'=a$;
6. $a \leqslant 1$;
7. *If* $a \leqslant b$ *and* $a' \leqslant b$ *then* $b=1$;
8. *If* $a \leqslant b'$ *then there is a c in A such that* $a \leqslant c$, $b \leqslant c$, *and for all d in A if* $a \leqslant d$ *and* $b \leqslant d$ *then* $c \leqslant d$;
9. *If* $a \leqslant b$ *then there is a c in A such that* $c \leqslant a'$, $c \leqslant b$ *and for every d in A if* $a \leqslant d$ *and* $c \leqslant d$ *then* $b \leqslant d$.

The only axioms of any complexity are the last three. If the operation of addition for disjoint elements were given the three axioms would be formulated as follows:

7! $a+a'=1$;
8! $a+b$ *is in* A;
9! *If* $a \leqslant b$ *then there is a c in A such that* $a+c=b$.

The difficulty with the operation of addition is that we do not want it to be defined except for disjoint elements, i.e., elements a and b of A such that $a \leq b'$.

It should also be apparent that we obtain a σ-algebra by adding to the axioms of Definition 5 the condition that for any sequence of pairwise disjoint elements $a_1, a_2, \ldots, a_n, \ldots$ of A there is a c in A such that for all n, $a_n \leqslant c$ and for every d in A, if for every n, $a_n \leqslant d$, then $c \leqslant d$.

Although it may be apparent, in the interest of explicitness, it is desirable to prove the following theorem.

THEOREM 3: *Every quantum-mechanical algebra of sets is a quantum-mechanical algebra in the sense of Definition* 5.

Proof: Let $\mathscr{F}$ be a quantum-mechanical algebra of sets on X. The relation $\leqslant$ of Definition 5 is interpreted as set inclusion $\subseteq$, and Axioms 1–3 immediately hold. The complementation is interpreted as set complementation with respect to X, and Axioms 4 and 5 hold in this interpretation. The Unit I is interpreted as the set X, and Axiom 6 holds because for any A in $\mathscr{F}$, $A \subseteq X$. In the case of Axiom 7 it is evident from elementary set theory that if $A \subseteq B$ and $\sim A \subseteq B$, then $A U \sim A \subseteq B$, whence $X \subseteq B$, but $B \subseteq X$, and so $B = X$. Regarding Axiom 8, if $A \subseteq \sim B$ then $A \cap B = 0$, so $A \cup B \in \mathscr{F}$ by virtue of the second axiom for algebras of sets, and we may take $C = A \cup B$ to satisfy the existential requirement of the axiom, because $A \subseteq A \cup B$, $B \subseteq A \cup B$, and if $A \subseteq D$ and $B \subseteq D$ then $A \cup B \subseteq D$. Finally, as to Axiom 9, if $A \subseteq B$ then we first want to show that $B \sim A \in \mathscr{F}$. By hypothesis $A, B \in \mathscr{F}$, whence $\sim B \in \mathscr{F}$, and since $A \subseteq B$, $A \cap \sim B = 0$ and thus $A \cup \sim B \in \mathscr{F}$, but then because $\mathscr{F}$ is closed under complementation, $\sim(A \cup \sim B) = \sim A \cap B = B \sim A \in \mathscr{F}$, as desired. It is easily checked, in order to verify Axiom 9 that because $A \subseteq B$, we have $B \sim A \subseteq \sim A$, $B \sim A \subseteq B$ and for every set D in $\mathscr{F}$, if $A \subseteq D$ and $B \sim A \subseteq D$ then $B \subseteq D$, since $A \cup (B \sim A) \subseteq D$ and $A \cup (B \sim A) = B$. Thus $B \sim A$ is the desired C, which completes the proof.

To obtain a sentential calculus for quantum-mechanical algebras, we define the notion of validity in the standard way. More particularly, in the calculus implication $\rightarrow$ corresponds to the relation $\leqslant$ and negation $\neg$ to the complementation operation $'$. We say that a sentential formula is quantum-mechanically valid if it is satisfied in all quantum-mechanical algebras, i.e., if under the expected interpretation the formula designates the element 1 of the algebra. The set of such valid sentential formulas characterizes the sentential logic of quantum mechanics. The axiomatic structure of this logic will be investigated in a subsequent paper.

I conclude with a brief remark about Reichenbach's three-valued logic.

It is easy to show that the quantum-mechanical logic defined here is not truth-functional in his three values (for more details see Suppes, 1965b). It seems clear to me that his three-valued logic has little if anything to do with the underlying logic required for quantum-mechanical probability spaces, and I have tried to show why the logic of quantum-mechanical probability is *the* logic of quantum mechanics. What I have not been able to do within the confines of this paper is to make clear precisely why the algebras characterized in Definition 5 are exactly appropriate to express the logic of quantum-mechanical probability. The argument in support of this choice is necessarily rather long and technical. A fairly good case is made out in detail in Varadarajan (1962).

However, apart from giving a mathematically complete argument for Definition 5, it may be seen that quantum-mechanical algebras have many intuitive properties in common with Boolean or classical algebras. The relation of implication or inclusion has most of its ordinary properties, the algebras are closed under negation, and the classical law of double negation holds. What is lacking are just the properties of closure under union and intersection – or disjunction and conjunction – that would cause difficulties for nonexistent joint probability distributions.

NOTE

[1†] Article 13 in this volume.

PART IV

FOUNDATIONS OF PSYCHOLOGY

This part contains the most articles of the four parts and reflects properly the greater emphasis and stress of my own research over the past decade. The first three articles deal with general issues in the foundations of psychology. The first article gives an axiomatization of stimulus-sampling theory for a continuum of responses, which represents an extension of the theory to this experimentally and conceptually useful case. The second, more philosophical article discusses the nature and limitations of behaviorism. A main concern are the arguments of intentionalists like Chisholm about the limitations of behaviorism. The third article treats an interesting claim about the predictability of human behavior originally put forth by Michael Scriven. In these two philosophical articles, 18 and 19, I have used the apparatus laid out in more detail in the first article on stimulus-sampling theory.

The remaining articles in this part deal in one way or another with the foundations of mathematics or the foundations of psycholinguistics. The first article on this subject, Article 19, reports some of my earliest work in attempting to provide behavioral foundations for the learning of mathematical concepts by children. Article 20 extends this work to some very simple cases of mathematical proofs, and Article 21 continues this extension to a wider range of topics in the foundations of mathematics. Article 22 is a recently written general survey of the theory of cognitive processes and extends beyond the foundations of mathematics or psycholinguistics, but much of what is said bears on the psychological analysis of mathematics learning or thinking. Finally, the last article, in many ways the most substantial one in this part, deals with stimulus-response theory of finite automata. I try to link the work of linguists concerned with production or recognition automata for various types of languages to classical concepts and assumptions of stimulus-response theory. The work begun on this last article is not yet finished, but I think the direction of research, if it can be successfully continued, will be of some importance for psycholinguistics and also for the psychological foundations of mathematics itself.

Because of the conceptual richness and depth of mathematics, it is a particularly inviting topic for psychological analysis. It is surprising that so little scientific work of a complex or extended sort has yet been done on mathematics learning or thinking. The articles reprinted here certainly represent only a very modest beginning, and I know many mathematicians and philosophers are skeptical that this kind of beginning can lead to a more adequate framework for what goes on in the minds of students as they learn mathematics or in the minds of mathematicians as they create new concepts and domains of thought. In the opening pages of the last article, I try to give reasons for thinking that there is some ground for hope, and I shall not review the arguments given there.

During the period when these eight articles were written, I also wrote a number of articles concerned with more detailed questions in mathematical psychology, or in some cases, actual reports of experiments. Although I have a considerable personal involvement in this work, it did not seem sufficiently methodologically or philosophically oriented to include in the present volume, or to review in detail. Also, I have not attempted to provide additional leads into the literature, because the articles reprinted here were written recently and contain references to the relevant current literature.

It is clear that the foundations of psychology as discussed here is quite a different subject from what is currently called 'philosophical psychology'. The only one of the eight articles that makes serious contact with the current literature in philosophical psychology is the one on behaviorism (Article 18). This is not the proper occasion to defend my conceptual approach against that typical in philosophical psychology. I would say, however, that what is said here is much closer to the scientific spirit dominant in contemporary psychology, particularly in the U.S.A. Also, the main topic of these papers, the psychological foundations of mathematics, has scarcely been discussed in this recent literature of philosophical psychology.

Because the work in the final paper, the one on stimulus-response theory of finite automata, is perhaps the one of greatest philosophical interest, or at least because it is now of greatest interest to me, I would like to say something about the sort of additional results needed to convert what I think is a promising beginning into a substantial approach to language learning, of significance for both psycholinguistics and the philosophy of language.

At least six stages of additional results seem necessary to establish in any definitive way the viability of the approach in the article reprinted here.

(1) Detailed grammars of a probabilistic sort need to be worked out for the speech of young children. The developmental changes in these grammars need to be clearly identified, in order to provide a much sharper sense of what the child continues to acquire in terms of grammar or syntax as he matures from, say, 20 months to 5 years. (I leave aside all the important matters of phonology and the learning to recognize and produce the physical sounds of a language – primarily because I have nothing of significance to conjecture about them.)

(2) Such detailed grammars will provide a first estimate of the number of states required for language production or recognition on the part of a child. One of the apparently strongest lines of attack against a stimulus-response approach to language learning is centered on the claim that, if this approach were really correct, the child must learn an impossibly large number of conditioning connections. Perhaps everyone can agree that the number is too large if each sentence produced must somehow be learned as a discrete and indivisible unit. But a more sophisticated attack has scarcely begun on estimating the number of states in an automaton model of the child's behavior and determining by any serious quantitative argument whether the minimum number of states can possibly be acquired by conditioning processes.

There are two recurring but mistaken objections of linguists to stimulus-response ideas, which are appropriate to mention at this point. The first is the claim that an automaton with an unbounded number of states is needed to recognize or produce a natural language. There are many ways of showing how nonsensical this claim is when addressed to the actual performance of language users. Some of these ways are detailed in the first chapter of Crothers and Suppes (1967), and I shall not explore others here, except to remark that a very appropriate probabilistic approximation of any unbounded automaton generating a probabilistic grammar can be made by a fixed finite-state device, along lines familiar in the theory of ergodic stochastic processes, especially chains of infinite order. (I hope to publish some results on these matters in the near future.)

The second misplaced objection is to the asymptotic nature of the theorems proved in Article 23. Linguists unfamiliar with how such

asymptotic mathematical results are ordinarily related to finite experimental data claim that what is needed is *actual*, not asymptotic, modeling of a production automaton by a stimulus-response process. As can be seen by even the most casual perusal of the literature of modern mathematical learning theory, however, the behavior after a small finite number of trials in most straight learning experiments is already, for all practical purposes, at asymptote. When theoretical parameters of conditioning are then estimated from the experimental data, the difference in probabilities of responses on trial 500, say, and at asymptote (at trial infinity, so to speak), as predicted in terms of the estimated parameters, will be less than 10^{-4}.

The use of asymptotic results in learning theory is like the use of twice-differentiable paths in theoretical mechanics, a mathematical fiction for convenience of computation and analysis, and an approximation that remains close to the less tractable brute facts.

(3) Those who are not sympathetic to a stimulus-response approach rightly point out that it is not a simple matter to identify the reinforcing events required for conditioning to take place. No doubt the details are difficult to supply and the problems of analysis extraordinarily subtle, because of the intimate and rapid nonverbal means of communication and reinforcement between parent and child. But, some of the gross facts reside on the behavioristic side of the fence. Not even the most hidebound nativist suggests any child has ever learned any language except the one spoken around him. On the other hand, any behaviorist with the slightest modicum of scientific sense will admit that the genetic structure of the human child brings a high order of finely tuned equipment to the job of language learning. Identifying more sharply the nature and role of reinforcement will help immensely in locating with greater precision the line between innate and behavioral components of language learning.

(4) The article on finite automata has nothing at all to say about semantics, but in any successful theory of language learning, that omission can be only temporary. In many ways the associationistic nature of stimulus-response theory would seem to provide a natural home for semantics, but in simplest garb, this would be only a theory of reference and not a theory of meaning. It may well be that the currents and nuances of ordinary talk will be as difficult to analyze and predict with any

accuracy as are the currents and eddies of a gust of wind moving through a treetop, even though the fundamental dynamical theory of the motion of air is thought to be well understood. Our objective, therefore, for a theory must be more limited than that of providing a complete account. But, even though a complete semantical theory of nuance may be unattainable, a more modest semantical theory, comparable at least to what now exists for formal languages in the theory of models, will be required.

(5) Something not too far from the theory of models may provide a good first approximation to a semantical theory of ordinary language, but for any real account of language learning and use, something far more is required. A theory is required to explain why one utterance rather than another, or nothing at all, is spoken on a given occasion. A satisfactory and adequate account of a child's language-learning must enunciate principles that lead to at least probabilistic predictions about the contents of his utterances on given occasions. Such a theory might well be called the *semantical theory of utterances*, as opposed to the semantical theory of statements or sentences.

The first necessity of such a theory would seem to be the inclusion of a theory of perception, for the child above all responds to the stimuli immediately impinging on his peripheral receptors, not on his own brooding thoughts of yesteryear. A variety of theories of perception and concept formation centering on the notion of the organism's always working with and modifying an internal template of the environment are currently being developed and seem to hold some promise. The philosophically interesting point about these "activist" template models is that they are far removed from the conception of passive sense-data receivers often entertained by philosophers as models of perceiving.

(6) Perhaps the best test of any theory of language learning will be its ability to provide the plans for the construction of a computer that learns to talk. The line of behaviorism that derives from Wittgenstein does not offer any systematic or scientific account of the mechanisms by which behavior is learned, and at times, as in the case of Ryle, seems to suggest that such a deeper scientific account of mechanism would be otiose. But the man who wants to try to build, or more aptly, program, a computer to talk knows better after the first hour of the first day of attack on the problem. It is not unreasonable to forecast that in fifty years the

most controversial and important topic in the philosophy of language will be the conceptual attitude taken toward talking computers. And if this is so, it is very likely that the status of stimulus-response theories of language, albeit much better developed, will continue also to be a part of the controversy.

16. STIMULUS-SAMPLING THEORY FOR A CONTINUUM OF RESPONSES*

I. INTRODUCTION

The aim of the present investigation is to extend stimulus-sampling theory to situations involving a continuum of possible responses. The theory for a finite number of responses stems from the basic paper by Estes (1950); the present formulation will resemble most closely that given for the finite case in Suppes and Atkinson (1959). In a previous study (Suppes, 1959b) I was concerned with a corresponding extension of linear learning models, and several results of that study are, as we shall see, closely related to the present one.

The experimental situation consists of a sequence of trials. On each trial the subject (of the experiment) makes a response from a continuum of possible responses; his response is followed by a reinforcing event indicating the correct response for that trial. In situations of simple learning, which are characterized by a constant stimulating situation, responses and reinforcements constitute the only observable data, but stimulus-sampling theory postulates a considerably more complicated process which involves the conditioning and sampling of stimuli. In the finite case the usual assumption is that on any trial each stimulus is conditioned to exactly one response. Such a highly discontinuous assumption seems inappropriate for a continuum of responses, and I have replaced it with the postulate that the conditioning of each stimulus is *smeared* over a certain interval of responses, possibly the whole continuum. In these terms, the conditioning of any stimulus may be represented uniquely by a *smearing distribution*. These distributions, one for each stimulus, will play the same role as did the single smearing distribution introduced in my earlier paper on linear models (Suppes, 1959b).

* Reprinted from *Mathematical Methods in the Social Sciences, 1959* (ed. by K. J. Arrow, S. Karlin, and P. Suppes), Stanford University Press, Stanford, Calif., 1960, pp. 348–365.

The theoretically assumed sequence of events on any trial may then be described as follows:

trial begins with each stimulus in a certain state of conditioning	→ certain stimuli are sampled	→ response occurs	→ reinforcement occurs	→ possible change in conditioning occurs.

The sequence of events just described is, in broad terms, postulated to be the same for finite and infinite sets of possible responses. Differences of detail will become clear. The main point of the axioms in Section II is to state specific hypotheses about this sequence of events. As has already been more or less indicated, three kinds of axioms are needed: conditioning axioms, sampling axioms, and response axioms.

Section III contains some general theorems of the theory. Section IV considers in some detail the classical case of noncontingent reinforcement. Section V treats other cases more superficially.

Although no experimental data will be described in this paper, it will perhaps help to describe schematically one piece of apparatus which has been used to test the theory extensively. The subject is seated facing a large circular vertical disc. He is told that his task on each trial is to predict by means of a pointer where a spot of light will appear on the rim of the disc. The subject's pointer predictions are his responses in the sense of the theory. At the end of each trial the "correct" position of the spot is shown to the subject, which is the reinforcing event for that trial. The most important variable controlled by the experimenter is the choice of a particular probability distribution of reinforcement.

II. AXIOMS

The axioms are formulated verbally but with some effort to convey a sense of formal precision. It is not difficult, although not wholly routine, to convert them into a mathematically exact form. As already indicated, they fall naturally into three groups. In the statement of the axioms we use x for the response variable and z for the parameter of the smearing distribution $K_s(x; z)$ of any stimulus s. Moreover, z is the mode of the distribution; for the circular disc apparatus it is also assumed to be the mean, but not all apparatus to which the theory applies is so completely symmetric.

Conditioning axioms

C1. *For each stimulus s there is on every trial a unique smearing distribution* $K_s(x; z)$ *on the interval* $[a, b]$ *of possible responses such that* (a) *the distribution* $K_s(x; z)$ *is determined by its mode z and its variance;* (b) *the variance is constant over trials for a fixed stimulating situation;* (c) *the distribution* $K_s(x; z)$ *is continuous and piecewise differentiable in both variables.*

C2. *If a stimulus is sampled on a trial, the mode of its smearing distribution becomes, with probability* θ, *the point of the response (if any) which is reinforced on that trial; with probability* $1-\theta$ *the mode remains unchanged.*

C3. *If no reinforcement occurs on a trial, there is no change in the smearing distributions of sampled stimuli.*

C4. *Stimuli which are not sampled on a given trial do not change their smearing distributions on that trial.*

C5. *The probability* θ *that the mode of the smearing distribution of a sampled stimulus will become the point of the reinforced response is independent of the trial number and the preceding pattern of occurrence of events.*

Sampling axioms

S1. *Exactly one stimulus is sampled on each trial.*

S2. *Given the set of stimuli available for sampling on a given trial, the probability of sampling a given element is independent of the trial number and the preceding pattern of occurrence of events.*

Response axioms

R1. *If the sampled stimulus s and the mode z of its smearing distribution are given, then the probability of a response in the interval* $[a_1, a_2]$ *is* $K_s(a_2; z)-K_s(a_1; z)$.

R2. *This probability of response is independent of the trial number and the preceding pattern of occurrence of events.*

Because of the similarity of these axioms to those in Suppes and Atkinson (1959) I shall here mainly comment on those aspects peculiar to the continuum case. In the finite case the complicated form of Axiom C1 reduces simply to the assertion that on any trial each stimulus is conditioned to exactly one response. As already remarked, the assumption

[C1(a)] that the smearing distribution of any stimulus is determined by its mode and variance, rather than its mean and variance, is used in order to permit application of the theory to unsymmetrical apparatus. For instance, suppose the experimental set-up consists of a bar a meter or so in length on which the subject is to set a pointer to predict the occurrence of a spot of light. It seems unreasonable to suppose that the conditioning effect of a reinforcement near the end points of the bar will be smeared symmetrically to the left and to the right. For such a situation the mean of the smearing distribution (of a sampled stimulus) may not be at the point of reinforcement even though conditioning is effective. On the other hand, it seems psychologically sound to assume that the mode of the smearing distribution will be at the point of reinforcement – granted the effectiveness of conditioning. In the present formulation of the theory it is essential to have the one free parameter of the smearing distribution closely tied to the points of reinforcement, for when conditioning is effective, which occurs with probability θ, this parameter assumes the value of the point of reinforcement (Axiom C2). This corresponds to the assumption in the finite response case that with probability θ sampled stimuli become conditioned or connected to the reinforced response.

The remaining conditioning axioms (C3, C4, C5) have almost exactly the form which is also appropriate for the finite case. The same is true of the two sampling axioms. In contrast, the first response axiom, R1, has a much simpler form in the finite case: with probability 1 the response is made to which the sampled stimulus is conditioned. Axiom R1 generalizes this assumption in the obvious manner in terms of the smearing distribution of the sampled stimulus.

The three axioms C5, S2, and R2 are what have been termed in the literature *independence-of-path* assumptions. Only R2 is new here; the other two are also needed in the finite case. These three axioms are crucial in proving that for simple reinforcement schedules the sequence of random variables which take as values the modes of the smearing distributions of the stimuli constitutes a continuous-state Markov process.

We next introduce some notation. In particular, we need notation for five random variables, their values, and their distributions, as well as a notation for their joint distribution. Three of these random variables take values in the interval $[a, b]$, the continuum of possible responses and reinforcements fixed throughout the paper. Thus we have for trial n:

(i) the *response* random variable $\boldsymbol{X}_n$, with values x_n or simply x, distribution R_n, and density r_n;

(ii) the *reinforcement* random variable $\boldsymbol{Y}_n$, with values y_n or y, distribution F_n, and density f_n;

(iii) the *smearing-parameter* random variable $\boldsymbol{Z}_{s,n}$ of stimulus s, with values $z_{s,n}$ or z_s, distribution $G_{s,n}$, and density $g_{s,n}$. As indicated already, z_s is the mode of the smearing distribution of stimulus s. The random variable $\boldsymbol{Z}_n$, without the subscript s, shall take as values finite vectors $z = (z_{s_1}, \ldots, z_{s_N})$ relative to the ordering $(s_1, \ldots, s_N)$ of the set S of stimuli.

We also need for occasional use:

(iv) the *sampling* random variable $\boldsymbol{S}_n$, with values s_n or s for the sampled stimulus, and discrete density σ_n (it is always assumed that the set S of stimuli is finite);

(v) the *effectiveness-of-conditioning* random variable $\boldsymbol{D}_n$, with value 1 for effective and 0 for noneffective, and probability θ of value 1, following Axiom C2. I use $\delta_{i,n}$ for values of D_n. Thus $\delta_{i,n}$ is always either 1 or 0.

I use J_n for the joint distribution of any finite sequence of these random variables the last of which occurs on trial n, and j_n for the corresponding density. For occasional reference to points in the underlying sample space, ξ is used. Finally, the notation $K_s(x_n; z_n)$ for the smearing distribution of stimulus s was introduced earlier.

In terms of the five random variables introduced, the postulated sequence of events on any trial, which was described informally before, may be symbolized as follows:

$$\boldsymbol{Z}_n \to \boldsymbol{S}_n \to \boldsymbol{X}_n \to \boldsymbol{Y}_n \to \boldsymbol{D}_n \to \boldsymbol{Z}_{n+1}.$$

Note that the value of the random variable $\boldsymbol{Z}_n$ represents the conditioning of each stimulus at the beginning of trial n, for in the present continuous theory conditioning is in terms of a one-parameter family of smearing distributions.

It will also be useful to give a more precise formulation of the response axioms, R1 and R2, in terms of the notation just introduced. It is intended that R1 should simply make the following assertion:

$$P(a_1 \leqq \boldsymbol{X}_n \leqq a_2 \mid \boldsymbol{S}_n = s, \boldsymbol{Z}_{s,n} = z) = \int_{a_1}^{a_2} j_n(x \mid s, z)\, dx$$
$$= K_s(a_2; z) - K_s(a_1; z).$$

Axiom R2 states an independence-of-path assumption. Let w_{n-1} be any sequence of outcomes of the random variables defined up to trial $n-1$. Then R2 asserts:

$$\int_{a_1}^{a_2} j_n(x \mid s_n, z_{s,n}, w_{n-1})\, dx = \int_{a_1}^{a_2} j_n(x \mid s_n, z_{s,n})\, dx = K_s(a_2; z) - K_s(a_1; z).$$

The following obvious relations for the response density r_n will also be helpful later. First, we have that

$$r_n(x) = j_n(x),$$

i.e., r_n is just the marginal density obtained from the joint distribution j_n. Second, we have "expansions" like

$$r_n(x) = \int_a^b j_n(x, z_{s,n})\, dz_{s,n},$$

$$r_n(x) = \int_a^b \int_a^b \int_a^b j_n(x, z_{s,n}, y_{n-1}, x_{n-1})\, dz_{s,n}\, dy_{n-1}\, dx_{n-1}.$$

III. GENERAL THEOREMS

This section contains five general theorems, most of which correspond to theorems that have proved useful in experimental work with the finite case. It is assumed that the reinforcement distribution F_n, which is selected by the experimenter, is always continuous and piecewise differentiable in all variables. Under these assumptions and those of Axiom C1 on the smearing distributions, no questions of integrability arise. Proofs of the first theorems are rather explicit in order to indicate the role of the axioms.

THEOREM 1 *(General Response Theorem):*

$$(1) \qquad r_n(x) = \sum_{s \in S} \sigma_n(s) \int_a^b k_s(x; z_s)\, g_{s,n}(z_s)\, dz_s.$$

Proof: Mainly by virtue of Axiom S1, which asserts that exactly one stimulus is sampled on each trial,

$$(2) \qquad r_n(x) = \sum_S \int_a^b j_n(x, s, z_s)\, dz_s$$

$$= \sum_S \int_a^b j_n(x \mid s, z_s)\, j_n(s \mid z_s)\, j_n(z_s)\, dz_s .$$

In view of Axioms C1 and R1,

$$(3) \qquad j_n(x \mid s, z_s) = k_s(x; z_s);$$

from Axiom S2, the independence-of-path assumption on sampling,

$$(4) \qquad j_n(s \mid z_s) = \sigma_n(s);$$

and on the basis of the notation introduced in the last section,

$$(5) \qquad j_n(z_s) = g_{s,n}(z_s).$$

The theorem follows immediately from (2)–(5). Q.E.D.

The next theorem asserts the Markov property, which is essential for further deductive developments of the theory. It is a straightforward matter to generalize this theorem to more complicated reinforcement distributions which depend on the actual responses or reinforcements on several preceding trials; the generality of the present theorem is sufficient for our purposes here.

THEOREM 2 *(Markov Theorem): If the reinforcement distribution $F(y)$ on trial n is independent of n and depends only on the immediately preceding response on trial n, then the sequence of random variables $\langle \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n, \ldots \rangle$ is a continuous-state Markov process.*

Proof: By direct probability considerations for $t_1, \ldots, t_m > 1$,

$$(6) \qquad j_n(z_n \mid z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) =$$

$$\sum_i \int_a^b \int_a^b \sum_{s \in S}$$

$$\times j_n(z_n \mid \delta_{i,n-1}, y_{n-1}, x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m})$$

$$\times j_{n-1}(\delta_{i,n-1} \mid y_{n-1}, x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m})$$

$$\times j_{n-1}(y_{n-1} \mid x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m})$$

$$\times j_{n-1}(x_{n-1} \mid s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m})$$

$$\times j_{n-1}(s_{n-1} \mid z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m})\, dy_{n-1}\, dx_{n-1} .$$

Now by Axiom C2, if $\delta_{i,n-1}=1$, then

$$j_n(z_n \mid \delta_{i,n-1}, y_{n-1}, x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = 1,$$

provided the vector $z_n=y_{n-1}$ in its coordinate for stimulus s; otherwise $j_n(z_n \mid \ldots)=0$. And if $\delta_{i,n-1}=0$, then $j_n(z_n \mid \ldots)=1$ if $z_n=z_{n-1}$; otherwise $j_n(z_n \mid \ldots)=0$. For any of these cases, the value of $j_n(z_n \mid \ldots)$ is not affected by $z_{n-t_1}, \ldots, z_{n-t_m}$. Second, by virtue of Axiom C5,

$$j_{n-1}(\delta_{i,n-1} \mid y_{n-1}, x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = j_{n-1}(\delta_{i,n-1}).$$

Third, on the basis of the hypothesis of the theorem,

$$j_{n-1}(y_{n-1} \mid x_{n-1}, s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = f(y_{n-1} \mid x_{n-1}).$$

Fourth, in view of Axioms R1 and R2,

$$j_{n-1}(x_{n-1} \mid s_{n-1}, z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = j_{n-1}(x_{n-1} \mid s_{n-1}, z_{n-1}).$$

Finally, in view of Axiom S2,

$$j_{n-1}(s_{n-1} \mid z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = \sigma_{n-1}(s_{n-1}).$$

When all these results of applying the independence-of-path assumptions are substituted in (6), and the summations and integrations are performed on the result, we have

$$j_n(z_n \mid z_{n-1}, z_{n-t_1}, \ldots, z_{n-t_m}) = j_n(z_n \mid z_{n-1}),$$

the desired result. Q.E.D.

Some readers may feel that the above theorem could have been assumed as an axiom, but this is to misunderstand the character of the theorem in the context of the general stimulus-sampling theory formulated by the axioms. The axioms on which this theorem is based are of a general nature and are concerned with fundamental aspects of the postulated psychological process of learning. In contrast, the theorem is relatively restricted, dealing as it does with only a small class of the possible schedules of reinforcement.

We turn now to some recursion theorems for various quantities; of particular interest is the one for response probabilities. It is possible to

state and prove these theorems under the general assumption of N stimuli in the set S. However, both computations and notation become rather cumbersome, so that at this stage of development of the theory it is a reasonable simplification to impose the following

RESTRICTIVE HYPOTHESIS: *There is exactly one stimulus element in S.*

Probabilities enter the theory for a continuum of responses in so many different ways that it is difficult to distinguish empirically between models with different numbers of stimuli when the stimulation is constant. And in the case of discrimination experiments, each stimulating situation may be treated as a single stimulus, with the result that on any trial there is exactly one stimulus available for sampling, although the set S may contain more than one element. As a matter of fact, this restrictive hypothesis of a single stimulus is already a practical necessity for complicated reinforcement situations in the finite case (see, for instance, Atkinson and Suppes, 1958).

We begin with a recursion for the distribution g_n of the smearing parameter z of the single stimulus. (On the assumption of a single stimulus we drop the subscript s.)

THEOREM 3:

$$g_{n+1}(z) = (1 - \theta)\, g_n(z) + \theta f_n(z). \tag{7}$$

Proof: By Axiom C2, if conditioning is effective, then $z_{n+1} = y_n$, and thus the distribution of z_{n+1} is that of y_n, which is f_n. On the other hand, if conditioning is not effective, then $z_{n+1} = z_n$, and thus the distribution of z_{n+1} is simply g_n. By Axiom C2 the probability of the first alternative is θ, and that of the second $1-\theta$, which yields the theorem. Q.E.D.

In the familiar notation of the finite case, where $A_{i,n}$ is response i on trial n and $E_{j,n}$ is reinforcing event j on trial n, (7) corresponds to:

$$P(A_{i,n+1}) = (1 - \theta)\, P(A_{i,n}) + \theta P(E_{i,n}). \tag{8}$$

For the response density r_n we have

THEOREM 4:

$$r_{n+1}(x) = (1 - \theta)\, r_n(x) + \theta \int_a^b k(x; y)\, f_n(y)\, dy. \tag{9}$$

Proof: We have at once from Theorem 1

$$r_{n+1}(x) = \int_a^b k(x; z)\, g_{n+1}(z)\, dz .$$

Applying Theorem 3 to the right-hand side, we have

$$\begin{aligned} r_{n+1}(x) &= \int_a^b k(x; z)\, [(1-\theta)\, g_n(z) + \theta f_n(z)]\, dz \\ &= (1-\theta) \int_a^b k(x; z)\, g_n(z) + \theta \int_a^b k(x; z)\, f_n(z)\, dz \\ &= (1-\theta)\, r_n(z) + \theta \int_a^b k(x; y)\, f_n(y)\, dy , \end{aligned}$$

where the variable of integration is changed in the second integral on the right. Q.E.D.

Robert R. Bush suggested that it is of interest to see what happens when the interval $[a, b]$ is cut into a finite number of parts and the resulting finite response case is studied. For simplicity, we may divide the interval into exactly two parts. Let $a<c<b$, and call $X_{1,n}$ a response on trial n in the interval $[a, c]$, and $X_{2,n}$ a response on trial n in $[c, b]$. Clearly

$$\begin{aligned} P(X_{1,n}) &= R_n(c) - R_n(a) = R_n(c), \\ P(X_{2,n}) &= R_n(b) - R_n(c) = 1 - R_n(c). \end{aligned}$$

And by integrating (9) of Theorem 4, we have at once

THEOREM 5:

$$\begin{aligned} P(X_{1,n+1}) &= (1-\theta)\, P(X_{1,n}) + \theta \int_a^c \int_a^b k(x; y)\, f_n(y)\, dx\, dy , \\ P(X_{2,n+2}) &= (1-\theta)\, P(X_{2,n}) + \theta \int_c^b \int_a^b k(x; y)\, f_n(y)\, dx\, dy . \end{aligned} \tag{10}$$

The recursions for $X_{1,n}$ and $X_{2,n}$ may be regarded as a generalization of (8) for the finite case when a continuous smearing of the effects of rein-

forcement is postulated. By further specialization, it is possible to get an exact analog of (8). Let us suppose that there are only two points of reinforcement, one the midpoint y_1 of the interval $[a, c]$, and the other the midpoint y_2 of the interval $[c, b]$. Suppose, moreover, that the smearing densities around these two points of reinforcement are strictly positive only in the subinterval $[a, c]$ or $[c, b]$ as the case may be. Define then

$$Y_{1,n} = \int_a^c k(x; y_1)\, dx, \quad Y_{2,n} = \int_c^b k(x; y_2)\, dx,$$

and under these suppositions (10) becomes

$$P(X_{i,n+1}) = (1 - \theta)\, P(X_{i,n}) + \theta P(Y_{i,n}),$$

an exact analog of (8). (Naturally, weaker suppositions will also yield such an analog, but the present example is illustrative of one method for obtaining the finite case from the continuous one.)

The suppositions just made to yield (8) may also be used to yield the standard theory of the finite case at a deeper level, for (8) is only a recursion in the mean probabilities of responses and in itself does not justify derivation of any sequential statistics like the probability of two successive A_1 responses. However, these matters will not be pursued further here.

In connection with this comparison of models, it may also be remarked that the response density recursion (9) of Theorem 4 is exactly the same as that obtained in Suppes (1959b) for the continuous-response linear model. Consequently, the results in Suppes (1959b) for various kinds of contingent reinforcement (and *a fortiori* noncontingent reinforcement) follow at once in the present theory.

IV. NONCONTINGENT REINFORCEMENT

For noncontingent reinforcement schedules – that is, those for which the distribution $F(y)$ is independent of n and the past – we first use the response density recursion (9) to prove some simple, useful results which do not explicitly involve the smearing distribution of the single stimulus element and which also hold in the linear model but were not stated in Suppes (1959b). There is, however, one necessary preliminary concerning

derivation of the asymptotic response distribution in the stimulus-sampling theory.

THEOREM 6: *In the noncontingent case*

$$r(x) = \lim_{n\to\infty} r_n(x) = \int_a^b k(x; y) f(y)\, dy. \tag{11}$$

Proof: Because in the noncontingent case $f_n(y)=f(y)$, we have at once from Theorem 3

$$g(z) = \lim_{n\to\infty} g_n(z) = f(z). \tag{12}$$

The theorem immediately follows from (12) and Theorem 1. Q.E.D.

We now use (11) to establish the following recursions. In the statement of the theorem $\mathscr{E}(X_n)$ is the expectation of the response random variable X_n; $\mu_r(X_n)$ is its rth raw moment; $\sigma^2(X_n)$ is its variance; and X is the random variable with density r.

THEOREM 7:

$$r_{n+1}(x) = (1-\theta)\, r_n(x) + \theta r(x), \tag{13}$$

$$\mathscr{E}(X_{n+1}) = (1-\theta)\, \mathscr{E}(X_n) + \theta \mathscr{E}(X), \tag{14}$$

$$\mu_r(X_{n+1}) = (1-\theta)\, \mu_r(X_n) + \theta \mu_r(X), \tag{15}$$

$$\sigma^2(X_{n+1}) = (1-\theta)\, \sigma^2(X_n) + \theta\sigma^2(X) + \theta(1-\theta) \times [\mathscr{E}(X_n) - \mathscr{E}(X)]^2. \tag{16}$$

Proof: Because $f_n(y)=f(y)$ in the noncontingent case, (13) follows at once from (9) and (11), i.e., from Theorems 4 and 6. Multiplying both sides of (13) by x^r and integrating over the interval $[a, b]$, we obtain (15), of which (14) is a special case. As for (16), we infer it from the following:

$$\begin{aligned}
\sigma^2(X_{n+1}) &= \mu_2(X_{n+1}) - \mathscr{E}(X_{n+1})^2 \\
&= (1-\theta)\, \mu_2(X_2) + \theta\mu_2(X) - (1-\theta)^2\, \mathscr{E}(X_n) \\
&\quad - 2\theta(1-\theta)\, \mathscr{E}(X_n)\, \mathscr{E}(X) - \theta^2 \mathscr{E}(X)^2 \\
&= (1-\theta)\, [\mu_2(X_n) - \mathscr{E}(X_n)^2] + \theta[\mu_2(X) - \mathscr{E}(X)^2] \\
&\quad + (\theta - \theta^2)\, \mathscr{E}(X_n)^2 - 2(\theta - \theta^2)\, \mathscr{E}(X_n)\, \mathscr{E}(X) \\
&\quad + (\theta - \theta^2)\, \mathscr{E}(X)^2 \\
&= (1-\theta)\, \sigma^2(X_n) + \theta\sigma^2(X) + \theta(1-\theta) \\
&\quad \times [\mathscr{E}(X) - \mathscr{E}(X_n)]^2. \quad \text{Q.E.D.}
\end{aligned}$$

Because (13)–(15) are first-order difference equations with constant coefficients we have as an immediate consequence of the theorem:

COROLLARY 7.1:

$$r_n(x) = r(x) - [r(x) - r_1(x)]\,(1-\theta)^{n-1}, \tag{17}$$

$$\mathscr{E}(X_n) = \mathscr{E}(X) - [\mathscr{E}(X) - \mathscr{E}(X_1)]\,(1-\theta)^{n-1}, \tag{18}$$

$$\mu_r(X_n) = \mu_r(X) - [\mu_r(X) - \mu_r(X_1)]\,(1-\theta)^{n-1}. \tag{19}$$

Although the linear and (one-element) stimulus-sampling models both yield (13)–(19), predictions in the two models are already different for one of the simplest sequential statistics, namely, the probability of two successive responses in the same or different subintervals.

For two subintervals $[a, c]$ and $[c, b]$, we have the following theorem for the stimulus-sampling model. The result generalizes directly to any finite number of subintervals.

THEOREM 8: *For noncontingent reinforcement*

$$\lim_{n\to\infty} P(a \leqq X_{n+1} \leqq c,\ a \leqq X_n \leqq c) = \theta R(c)^2 + (1-\theta) \times \int_a^c \int_a^c \int_a^b k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz, \tag{20}$$

$$\lim_{n\to\infty} P(a \leqq X_{n+1} \leqq c,\ c \leqq X_n \leqq b) = \theta R(c)\,[1 - R(c)] + (1-\theta) \int_a^c \int_c^b \int_a^b k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz, \tag{21}$$

where

$$R(c) = \lim_{n\to\infty} R_n(c).$$

Proof: We first establish (20). To begin with,

$$P(a \leqq X_{n+1} \leqq c,\ a \leqq X_n \leqq c) = \int_a^c \int_a^c j_{n+1}(x_{n+1}, x_n)\, dx_{n+1}\, dx_n.$$

Applying the axioms in the usual way to the right-hand side, we obtain

$$\begin{aligned}
&\int_a^c \int_a^c j_{n+1}(x_{n+1}, x_n)\, dx_{n+1}\, dx_n \\
&\quad = \int_a^c \int_a^b \sum_i \int_a^b \int_a^c \int_a^b j_{n+1}(x_{n+1}, z_{n+1}, \delta_{i,n}, y_n, x_n, z_n) \\
&\qquad \times dx_{n+1}\, dz_{n+1}\, dy_n\, dx_n\, dz_n \\
&\quad = \int_a^c \int_a^b \sum_i \int_a^b \int_a^c \int_a^b j(x_{n+1} \mid z_{n+1})\, j(z_{n+1} \mid \delta_{i,n}, y_n, x_n, z_n) \\
&\qquad \times j(\delta_{i,n})\, f(y_n)\, j(x_n \mid z_n)\, j(z_n)\, dx_{n+1}\, dz_{n+1}\, dy_n\, dx_n\, dz_n \\
&\quad = \int_a^c \int_a^b \int_a^c \int_a^b [k(x_{n+1}; y_n)\, \theta f(y_n)\, k(x_n; z_n)\, g_n(z_n) \\
&\qquad + k(x_{n+1}; z_n)(1-\theta)\, k(x_n; z_n)\, g_n(z_n)] \\
&\qquad \times dx_{n+1}\, dy_n\, dx_n\, dz_n .
\end{aligned}$$

Now

$$\lim_{n\to\infty} g_n(z) = f(z),$$

whence at asymptote, by rearranging the right-hand side and relettering variables, we obtain

$$\begin{aligned}
&\lim_{n\to\infty} P(a \leqq \mathbf{X}_{n+1} \leqq c,\ a \leqq \mathbf{X}_n \leqq c) \\
&\quad = \theta\left(\int_a^c \int_a^b k(x; y)\, f(y)\, dx\, dy\right) \\
&\qquad \times \left(\int_a^c \int_a^b k(x'; z)\, f(z)\, dx'\, dz\right) \\
&\qquad + (1-\theta) \int_a^c \int_a^c \int_a^b k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz .
\end{aligned}$$

But the first term on the right is just $\theta R(c)^2$, which when substituted in yields (20).

The argument establishing (21) proceeds along exactly the same lines, with functions of x_n now integrated over the interval $[c, b]$. Q.E.D.

For comparative purposes the corresponding results for the linear model are derived in the Appendix.

The theorem just proved may be used to develop a reasonably good method of estimating the learning parameter θ. The sequence of response random variables $\langle A_1, A_2, \ldots, A_n, \ldots \rangle$, where

$$A_n = \begin{cases} 1 & \text{if response on trial } n \text{ is in interval } [a, c], \\ 2 & \text{otherwise,} \end{cases}$$

is a chain of infinite order. If it were a first-order Markov chain, (20) and (21) could be used to obtain a maximum likelihood estimate of θ. The estimate θ^* proposed here is formally identical with the latter, but of course it is not the maximum likelihood estimate. I shall call it the *pseudo-maximum likelihood* estimate.

Let $a_1, a_2, \ldots, a_n$ represent a finite sequence of values of the response random variables $A_1, A_2, \ldots, A_n$ from trial 1 to trial n. Let s be the number of subjects. Then, granted statistical independence of the subjects, the maximum likelihood estimate of θ is the number $\hat{\theta}$ (if it exists) such that for all θ'

$$(22) \qquad \prod_{\sigma=1}^{s} f^{(\sigma)}(a_1, a_2, \ldots, a_n; \hat{\theta}) \geqq \prod_{\sigma=1}^{s} f^{(\sigma)}(a_1, a_2, \ldots, a_n; \theta'),$$

where $f^{(\sigma)}(a_1, a_2, \ldots, a_n; \hat{\theta})$ is the probability of the sequence of responses $a_1\ a_2, \ldots, a_n$ for subject σ when the learning parameter is $\hat{\theta}$.

As should be clear from preceding remarks, the pseudo-maximum likelihood estimate of θ is the number θ^* such that for all θ'

$$(23) \qquad \prod_{\sigma=1}^{s} \prod_{m=2}^{n} f^{(\sigma)}(a_m \mid a_{m-1}; \theta^*) f^{(\sigma)}(a_1; \theta^*)$$
$$\geqq \prod_{\sigma=1}^{s} \prod_{m=2}^{n} f^{(\sigma)}(a_m \mid a_{m-1}; \theta') f^{(\sigma)}(a_1; \theta').$$

To simplify notation, let $p_{ij}(\theta)$ be the probability of going from state i to state j $(i, j = 1, 2)$ with parameter θ; let n_{ij} be the number of actual transi-

tions from state i to state j, summed over trials and subjects (the n_{ij} are tabulated from experimental data); let $p_i(\theta)$ be the probability of being in state i on trial 1; and let n_i be the number of subjects in state i on trial 1. We then want to find the θ that maximizes

$$\prod_{i,\,j} p_i^{n_i}(\theta)\, p_{ij}^{n_{ij}}(\theta).$$

It is usually easier to work with the log of this expression, so we seek to maximize

$$L^*(\theta) = \sum_i \left(n_i \log p_i(\theta) + \sum_j n_{ij} \log p_{ij}(\theta) \right). \tag{24}$$

In most cases $L^*(\theta)$ has a local maximum, so we can find θ^* as an appropriate solution of

$$\frac{dL^*(\theta)}{d\theta} = \sum_i \left(\frac{n_i p_i'(\theta)}{p_i(\theta)} + \sum_j \frac{n_{ij} p_{ij}'(\theta)}{p_{ij}(\theta)} \right) = 0, \tag{25}$$

where p' is the derivative of p with respect to θ.

Now on the basis of (20) and (21), at asymptote we have

$$p_{11}(\theta) = \theta R(c) + \frac{(1-\theta)}{R(c)} \int_a^c \int_a^c \int_a^b \times k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz \tag{26}$$

and

$$p_{21}(\theta) = \theta R(c) + \frac{(1-\theta)}{1-R(c)} \int_a^c \int_c^b \int_a^b \times k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz, \tag{27}$$

and $p_i(\theta)$ is independent of θ. Also, of course, $p_{12}(\theta) = 1 - p_{11}(\theta)$, and $p_{22}(\theta) = 1 - p_{21}(\theta)$. Moreover,

$$p_{11}'(\theta) = R(c) - \frac{\alpha}{R(c)}, \quad p_{22}'(\theta) = R(c) - \frac{\beta}{1-R(c)}, \tag{28}$$

where

$$\alpha = \int_a^b K(c; z)^2 f(z)\, dz \tag{29}$$

$$= \int_a^c \int_a^c \int_a^b k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz$$

and

$$\beta = \int_a^c \int_c^b \int_a^b k(x; z)\, k(x'; z)\, f(z)\, dx\, dx'\, dz \tag{30}$$

$$= \int_a^b K(c; z)\,(1 - K(c; z))\, f(z)\, dz = R(c) - \alpha.$$

Applying (26)–(30) to (25) and using the fact that $p_i(\theta)$ is independent of θ, we obtain:

$$\frac{dL^*(\theta)}{d\theta} = \frac{n_{11}\left(R(c) - \dfrac{\alpha}{R(c)}\right)}{\theta R(c) + \dfrac{(1-\theta)\,\alpha}{R(c)}} + \frac{n_{12}\left(\dfrac{\alpha}{R(c)} - R(c)\right)}{1 - \theta R(c) - \dfrac{(1-\theta)\,\alpha}{R(c)}} \tag{31}$$

$$+ \frac{n_{21}\left(R(c) - \dfrac{\beta}{1 - R(c)}\right)}{\theta R(c) + \dfrac{(1-\theta)\,\beta}{1 - R(c)}}$$

$$+ \frac{n_{22}\left(\dfrac{\beta}{1 - R(c)} - R(c)\right)}{1 - \theta R(c) - \dfrac{(1-\theta)\,\beta}{1 - R(c)}} = 0.$$

Solving (31), we have

THEOREM 9: *If $r_1(x) = r(x)$ for all x in $[a, b]$, then the estimate θ^* is a solution of the quadratic equation*

$$N\theta^2 + [(N - n_{11})\,A + (n_{11} + n_{22})\,B + (N - n_{22})\,C]\,\theta \tag{32}$$
$$+ n_{22}AB + (n_{12} + n_{21})\,AC + n_{11}BC = 0,$$

where

$$A = \alpha/[R(c)^2 - \alpha], \quad B = -[R(c) - \alpha]/[R(c)^2 - \alpha],$$
$$C = [1 + \alpha - 2R(c)]/[R(c)^2 - \alpha], \quad N = \sum_{i,j} n_{ij}.$$

Moreover, if $R(c) = \frac{1}{2}$, *then*

$$\theta^* = -\frac{A(n_{12} + n_{21}) + B(n_{11} + n_{22})}{N}.$$

Note that the hypothesis of the theorem simply requires that we start counting trials at asymptote. The statistical properties of the estimator θ^* need investigation; it can be shown to be consistent.

I conclude the treatment of noncontingent reinforcement with two expressions dealing with important sequential properties of stimulus-sampling models. The first gives the probability of a response in the interval $[a_1, a_2]$ given that on the previous trial the reinforcing event occurred in the interval $[b_1, b_2]$.

THEOREM 10:

$$(33) \quad P(a_1 \leqq \mathbf{X}_{n+1} \leqq a_2 \mid b_1 \leqq \mathbf{Y}_n \leqq b_2) = (1 - \theta)[R_n(a_2) - R_n(a_1)] + \frac{\theta}{F(b_2) - F(b_1)} \int_{a_1}^{a_2} \int_{b_1}^{b_2} k(x; y) f(y)\, dx\, dy.$$

Proof: By the usual expansion

$$P(a_1 \leqq \mathbf{X}_{n+1} \leqq a_2 \mid b_1 \leqq \mathbf{Y}_n \leqq b_2) = \frac{1}{F(b_2) - F(b_1)} \times \int_{a_1}^{a_2} \int_{a}^{b} \sum_i \int_{b_1}^{b_2} \int_{a}^{b} j_{n+1}(x_{n+1}, z_{n+1}, \delta_{i,n}, y_n, z_n) \times dx_{n+1}\, dz_{n+1}\, dy_n\, dz_n.$$

And the right-hand side is

$$\frac{1}{F(b_2) - F(b_1)} \left[(1 - \theta) \int_{a_1}^{a_2} \int_{b_1}^{b_2} \int_{a}^{b} k(x; z)\, g_n(z)\, f(y)\, dx\, dy\, dz + \theta \int_{a_1}^{a_2} \int_{b_1}^{b_2} \int_{a}^{b} k(x; y)\, f(y)\, g_n(z)\, dx\, dy\, dz \right].$$

Now in the first and second terms, respectively, we have

$$\int_{b_1}^{b_2} f(y)\,dy = F(b_2) - F(b_1) \quad \text{and} \quad \int_{a}^{b} g_n(z)\,dz = 1.$$

Using these two results, we obtain the theorem at once. Q.E.D.

The expression to which we now turn gives the probability of a response in the interval $[a_1, a_2]$ given that on the previous trial the reinforcing event occurred in the interval $[b_1, b_2]$ and the response in the interval $[a_3, a_4]$.

THEOREM 11:

$$\begin{aligned}(34)\qquad P(a_1 \leqq X_{n+1} \leqq a_2 \mid b_1 \leqq Y_n \leqq b_2,\ a_3 \leqq X_n \leqq a_4) \\ = \frac{(1-\theta)}{R_n(a_4) - R_n(a_3)} \int_{a_1}^{a_2}\int_{a_3}^{a_4}\int_{a}^{b} \\ \times k(x; z)\, k(x'; z)\, g_n(z)\,dx\,dx'\,dz \\ + \frac{\theta}{F(b_2) - F(b_1)} \int_{a_1}^{a_2}\int_{b_1}^{b_2} k(x; y)\, f(y)\,dx\,dy.\end{aligned}$$

Proof: It is first useful to observe that for noncontingent reinforcement

$$\begin{aligned}&P(b_1 \leqq Y_n \leqq b_2,\ a_3 \leqq X_n \leqq a_4) \\ &\quad = P(b_1 \leqq Y_n \leqq b_2 \mid a_3 \leqq X_n \leqq a_4)\, P(a_3 \leqq X_n \leqq a_4) \\ &\quad = P(b_1 \leqq Y_n \leqq b_2)\, P(a_3 \leqq X_n \leqq a_4) \\ &\quad = [F(b_2) - F(b_1)]\,[R_n(a_4) - R_n(a_3)].\end{aligned}$$

Applying the usual expansion to the left-hand quantity in (34), we obtain

$$\begin{aligned}&\frac{1}{[F(b_2) - F(b_1)]\,[R_n(a_4) - R_n(a_3)]} \\ &\quad \times \int_{a_1}^{a_2}\int_{a}^{b} \sum_i \int_{b_1}^{b_2}\int_{a_1}^{a_2}\int_{a}^{b} j_{n+1}(x_{n+1}, z_{n+1}, \delta_{i,n}, y_n, x_n, z_n) \\ &\qquad\qquad \times dx_{n+1}\,dz_{n+1}\,dy_n\,dx_n\,dz_n,\end{aligned}$$

from which, using particularly Axioms C2 and C5, we have

$$\frac{1}{[F(b_2)-F(b_1)]\,[R_n(a_4)-R_n(a_3)]} \times \Bigg[(1-\theta)\int_{a_1}^{a_2}\int_{b_1}^{b_2}\int_{a_3}^{a_4}\int_{a}^{b} k(x;z)\,k(x';z)\,g_n(z)\,f(y) \times dx\,dy\,dx'\,dz + \theta\int_{a_1}^{a_2}\int_{b_1}^{b_2}\int_{a_3}^{a_4}\int_{a}^{b} k(x;y)\,f(y)\,k(x';z)\,g_n(z)\,dx\,dy\,dx'\,dz\Bigg].$$

Now in the first term of this last expression we may integrate out the function $f(y)$ to obtain $F(b_2)-F(b_1)$, which cancels the corresponding quantity in the denominator. Similarly, in the second term we may integrate out $k(x';z)\,g_n(z)$ to obtain $R_n(a_4)-R_n(a_3)$, which for this term cancels the corresponding quantity in the denominator. Putting these results together, we have exactly the theorem. Q.E.D.

It may be noticed that by applying Corollary 7.1 more explicit results are easily obtained from both Theorems 10 and 11.

V. SIMPLE DISCRIMINATION

It is of some interest to sketch how the present theory may be applied to simple discrimination situations in which on each trial exactly one stimulus s_i is presented, and associated with each s_i is a reinforcement distribution f^i. (Readers who do not like the idea of exactly one stimulus being presented may think of each s_i as being a particular *pattern* of stimuli.) Let the probability of presentation of s_i on any trial be ω_i, with

$$\sum_{i=1}^{N} \omega_i = 1, \quad \omega_i \neq 0$$

for $i=1,\ldots,N$, and ω_i independent of trial number and any behavior on preceding trials.

The tree of the Markov process in the states (z^1, z^2) for $N=2$ and $\omega_i=\frac{1}{2}$ is given in Figure 1.

Corresponding to Theorem 1, we have by the same sort of proof for

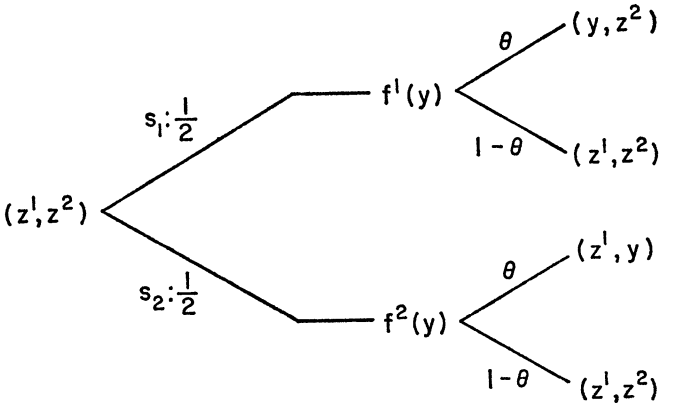


Fig. 1.

arbitrary N

$$r_n(x) = \sum_{i=1}^{N} \omega_i \int_a^b k_{s_i}(x; z^i)\, g_{s_i, n}(z^i)\, dz^i. \tag{35}$$

Corresponding to Theorem 3, we have

$$g_{n+1}(z^i) = (1-\theta)\, g_n(z_i \mid \boldsymbol{S}_n = s_i) + \theta f^i(z^i); \tag{36}$$

and by virtue of Axiom C4 for $i \neq j$ and $\boldsymbol{S}_n = s_j$,

$$g_{n+1}(z^i) = g_n(z_i), \tag{37}$$

whence it easily follows that

$$\lim_{n\to\infty} g_n(z^i) = f^i(z^i). \tag{38}$$

We then have also that

$$\lim_{n\to\infty} P(a_1 \leqq \boldsymbol{X}_n \leqq a_2 \mid \boldsymbol{S}_n = s_i) = \int_{a_1}^{a_2} \int_a^b k_{s_i}(x; y)\, f^i(y)\, dx\, dy. \tag{39}$$

The results (35)–(39) and some other related ones that are easily obtained, although simple in character, permit application of the theory developed in this paper to simple discrimination experiments with a

continuum of responses. On the other hand, it is obvious that the present theory must be modified and extended in fundamental ways to deal with discrimination experiments that have a continuum of stimuli as well as responses.

APPENDIX[1]

Our purpose is to derive for the linear model of Suppes (1959b) the analogs of (20) and (21). A brief description of the linear model will make the present discussion nearly self-contained. An experiment may be represented by a sequence $(\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2, \ldots, \boldsymbol{X}_n, \boldsymbol{Y}_n, \ldots)$ of response and reinforcement random variables. The theory is formulated for the probability of a response on trial $n+1$ given the entire preceding sequence of responses and reinforcements. For this sequence we use the notation s_n (not to be confused with the notation for the value of the sampling random variable in the main body of the paper). Aside from continuity and piecewise-differentiability assumptions, the single axiom of the linear model is

$$J_{n+1}(x \mid y_n, x_n, s_{n-1}) = (1-\theta)\, J_n(x \mid s_{n-1}) + \theta K(x; y_n), \tag{40}$$

where J_n is the joint distribution and K is the smearing distribution.

We first need to define the cross-moments

$$W(a_1, a_2, a_3, a_4, n) = \int_{a_1}^{a_2} \int_{a_3}^{a_4} \int_{s_{n-1}} \times j_n(x \mid s_{n-1})\, j_n(x' \mid s_{n-1})\, j(s_{n-1})\, dx\, dx'\, ds_{n-1}, \tag{41}$$

where the subscript s_{n-1} on the third integration sign indicates integration over the $2(n-1)$-Cartesian product of the interval $[a, b]$ for the sequence s_{n-1}. The cross-moments defined by (41) generalize the moments $W^2_{a_1, a_2, n}$ of Suppes (1959b).

Assuming henceforth *noncontingent reinforcement*, it follows by simple extension of some results in Suppes (1959b) that

$$\begin{aligned} \lim_{n\to\infty} P(a_1 \leqq \boldsymbol{X}_{n+1} \leqq a_2,\ a_3 \leqq \boldsymbol{X}_n \leqq a_4) \\ = (1-\theta) \lim_{n\to\infty} W(a_1, a_2, a_3, a_4, n) \\ + \theta[R(a_2) - R(a_1)]\,[R(a_4) - R(a_3)]. \end{aligned} \tag{42}$$

To obtain an explicit answer we must compute the limit on the right, which we now proceed to do.

By virtue of the definition of s_{n-1}, the right-hand side of (41) may be rewritten, and we have

$$
\begin{aligned}
(43)\qquad & W(a_1, a_2, a_3, a_4, n) \\
&= \int_{a_1}^{a_2}\int_{a_3}^{a_4}\int_{a}^{b}\int_{a}^{b}\int_{s_{n-2}} j_n(x \mid y_{n-1}, x_{n-1}, s_{n-2}) \\
&\quad \times j_n(x' \mid y_{n-1}, x_{n-1}, s_{n-2})\, j(y_{n-1}, x_{n-1}, s_{n-2}) \\
&\quad \times dx\, dx'\, dy_{n-1}\, dx_{n-1}\, ds_{n-2}.
\end{aligned}
$$

Applying the axiom (40) to the right-hand side of (43) and simplifying, we obtain

$$
\begin{aligned}
(44)\qquad & W(a_1, a_2, a_3, a_4, n) \\
&= (1-\theta)^2 \int_{a_1}^{a_2}\int_{a_3}^{a_4}\int_{s_{n-2}} j_{n-1}(x \mid s_{n-2}) \\
&\quad \times j_{n-1}(x' \mid s_{n-2})\, j(s_{n-2})\, dx\, dx'\, ds_{n-2} \\
&\quad + 2\theta(1-\theta) \int_{a_1}^{a_2}\int_{a_3}^{a_4}\int_{a}^{b}\int_{s_{n-2}} j_{n-1}(x \mid s_{n-2})\, j(s_{n-2}) \\
&\quad \times k(x'; y_{n-1})\, f(y_{n-1})\, dx\, dx'\, dy_{n-1}\, ds_{n-2} \\
&\quad + \theta^2 \int_{a_1}^{a_2}\int_{a_3}^{a_4}\int_{a}^{b} k(x, y_{n-1})\, k(x', y_{n-1})\, f(y_{n-1}) \\
&\quad \times dx\, dx'\, dy_{n-1}.
\end{aligned}
$$

Now the first term on the right of (44) is simply $(1-\theta)^2 W(a_1, a_2, a_3, a_4, n-1)$, the second term is

$$2\theta(1-\theta)\,[R_{n-1}(a_2) - R_{n-1}(a_1)]\,[R(a_4) - R(a_3)],$$

and the integral of the third term is a direct generalization of β as defined by (30). Moreover, it is independent of n; and we may thus define, for

ease of notation,

$$(45) \qquad \gamma(a_1, a_2, a_3, a_4) = \int_{a_1}^{a_2} \int_{a_3}^{a_4} \int_{a}^{b} k(x, y)\, k(x'; y)\, f(y)\, dx\, dx'\, dy.$$

In these terms, (44) becomes:

$$(46) \qquad \begin{aligned} W(a_1, a_2, a_3, a_4, n) &= (1-\theta)^2\, W(a_1, a_2, a_3, a_4, n-1) \\ &\quad + 2\theta(1-\theta)\,[R_{n-1}(a_2) - R_{n-1}(a_1)]\,[R(a_4) - R(a_3)] \\ &\quad + \theta^2 \gamma(a_1, a_2, a_3, a_4). \end{aligned}$$

It then easily follows from (46) that

$$(47) \qquad \begin{aligned} &\lim_{n\to\infty} W(a_1, a_2, a_3, a_4, n) \\ &\quad = W(a_1, a_2, a_3, a_4) \\ &= \frac{2(1-\theta)\,[R(a_2) - R(a_1)]\,[R(a_4) - R(a_3)] + \theta\gamma(a_1, a_2, a_3, a_4)}{2-\theta}. \end{aligned}$$

Combining (42) and (47), we then have the following theorem.

THEOREM: *In the linear model*

$$(48) \qquad \begin{aligned} &\lim_{n\to\infty} P(a_1 \leqq \boldsymbol{X}_{n+1} \leqq a_2,\ a_3 \leqq \boldsymbol{X}_n \leqq a_4) \\ &\quad = \theta\,[R(a_2) - R(a_1)]\,[R(a_4) - R(a_3)] + (1-\theta) \\ &\times \left[\frac{2(1-\theta)\,[R(a_2) - R(a_1)]\,[R(a_4) - R(a_3)] + \theta\gamma(a_1, a_2, a_3, a_4)}{2-\theta}\right]. \end{aligned}$$

To obtain the direct analog of (20), (48) specializes to:

$$\begin{aligned} \lim_{n\to\infty} P(a \leqq \boldsymbol{X}_{n+1} \leqq c,\ a \leqq \boldsymbol{X}_n \leqq c) &= \theta R(c)^2 + (1-\theta) \\ &\times \left[\frac{2(1-\theta)\,R(c)^2 + \theta\alpha}{2-\theta}\right], \end{aligned}$$

where α is defined by (29). The analog of (21) may be obtained in like fashion.

NOTE

[1] I am indebted to Raymond W. Frankmann for useful comments on the subject of this Appendix.

17. ON AN EXAMPLE OF UNPREDICTABILITY IN HUMAN BEHAVIOR*[1]

Scriven's example of essential unpredictability in human behavior goes like this. Assume a rational predictor P who wants to infer from information available to him the choice behavior of an individual C, where (i) C is choosing rationally and intelligently, (ii) C is *a contrapredictive* relative to P, i.e., C wishes to falsify any prediction made by P about his, C's choice behavior, (iii) C knows the information available to P. It is easy to make these conditions more precise and then to prove that P cannot predict C's behavior. Scriven fills out the bare bones of this example with many illustrations and comments, which I shall not attempt to sketch. His example of essential unpredictability is one I accept as in the main correct. The critical tenor of my remarks is to place the example in perspective from the standpoint of the theory of games and to challenge the significance of the example for the development of a quantitative and predictive behavioral science.

Consider the following simple game played under two different conditions. Following Scriven, let us call the two players C and P. We shall require that player P (the predictor) move first. Player P chooses action b_1 or b_2. Player C then moves, choosing action a_1 or a_2. Related to Scriven's example, we may interpret b_1 as a prediction by P that C will choose a_1, and b_2 the prediction by P that C will choose a_2. The payoff matrix for the game we may take to have the following simple form

C \ P	b_1	b_2
a_1	-1	1
a_2	1	-1

We shall make it zero-sum by assuming that when C obtains 1 unit, P loses 1 unit, and vice versa. (The exact payoff outcomes are not essential

* Reprinted from *Philosophy of Science* **31** (1964), 143–148.

for the discussion, and it is not necessary to assume the game is zero-sum or even constant sum, but we may as well for simplicity.)

From the standpoint of the theory of games, the fundamental question concerning the rules of this game is whether C does or does not know P's choice when he makes his own move. If C is informed of P's choice, then the game is one of perfect information and according to a theorem that originates with Zermelo, the game is strictly determined. (A game is strictly determined when optimal strategies for both players are nonrandom pure strategies.) Zermelo's theorem asserts that this is the case for games of perfect information. As is obvious for this simple game, when played under conditions of perfect information, C is in a position always to win. This is certainly a trivial and obvious conclusion.

If the rules of the game are now changed in a fundamental way so that P's move is not known to C, then the game is no longer one of perfect information, and pure, nonrandom strategies are no longer optimal. For the outcome payoffs indicated, the optimal, minimax strategies for both P and C are to choose one of their two alternatives with probability $\frac{1}{2}$. If changes are made in the outcomes to upset the perfect symmetry of the present matrix, then the exact numerical randomization that should be followed for optimality will be something different from $\frac{1}{2}$. But these minor quantitative variations do not seem important enough to pursue. Moreover, the restriction to two choices each for P and C is not essential; it is a straightforward matter to consider the case of n alternatives, and it is technically but not conceptually complicated to consider a continuum of possible choices for each of the two players.

Scriven's example of essential unpredictability of human behavior seems to me to be simply a special case of such a simple two-person competitive game. His second condition on C is the condition that assures us that C is playing the game, namely, the condition that C is a contrapredictive relative to P.

To show why I think that Scriven is making claims that are too large for his example, I want to distinguish first various cases of prediction for the behavior of players in a game. For those who prefer Scriven's language of contrapredictives, the translation is simple and direct. Before considering these cases, it is important to emphasize that prior to the analysis of the predictability of C's moves, we must fix upon the game C is playing. We can scarcely be expected to analyze or make assertions about the pre-

dictability of C's moves if after we have made the predictions, we are told that C is not playing this game, but some other game. So I take it in the following analysis that C and P are playing a game with moves and outcomes as specified above. If, in fact, they are not playing this game but some other game, then what is to be said about predictability must be changed. This talk about games should not be taken in too restrictive a sense, for it is a way of saying what C is motivated to do. Consider, for example, the first case I shall describe. The game is the game of perfect information already described. P's problem is to predict what move C will make. Now as a side bet, with another person P', P says that he will predict, separate from the game, what C's moves will be. In this game of perfect information, if C is playing it with serious intent, it is perfectly clear that on the basis of side bets P can indeed make such a prediction in a deterministic fashion. Namely, when P chooses move b_1, C will choose a_2 and when P chooses b_2, C will choose a_1. That is, when P predicts in the game that C will choose a_1, then in actual fact, C will choose a_2 in the game, etc.

There is nothing surprising or paradoxical in this. As most parents know, a situation very close to this arises in the age of most children. Namely, the parents will tell or ask the child to do something and in a very high probability of cases the initial response of the child will be negative. *Sotto voce* or in conversation with another person, the parent can reliably predict this response on the part of the child. We are able to make such predictions by first specifying the game that C, (here, the child) is playing. If C or the child is concerned with the actions, side bets, motives, etc., of person P' he can begin to play another game and behave in a different way, but it is a matter of observation to decide what game he is playing. By these remarks I don't mean to suggest that if we begin to analyze complicated social behavior on the part of candidates for office, lawyers in courtrooms, or lovers in parked cars, we can precisely describe the game the participants are playing and readily predict their moves. Our understanding of the details of behavior are not yet that good. My point is just that the conceptual schema introduced by Scriven does not raise any new or essentially surprising difficulties.

Holding these remarks in mind, let us look at some of the possible cases of prediction that may arise in the context of simple games.

Case 1A: C and P are playing a game of perfect information. P's

moves in the game, intended to be predictions of C's moves, are bad indeed. The game is as defined above, and P is always wrong.

Case 1B: P is making side bets about Case 1A with person P'. The side bets are always correct. If C does want to change the game and now play a game of perfect information defined against both P and P', then it will in general be possible for P to make new side bets with person P'' predicting with probability 1, C's behavior in the enlarged game, provided of course that C is playing the new enlarged game with the same seriousness he played the original game.

Case 2A: C and P are now playing the game in the condition of imperfect information described above. In this case the optimum strategy for both players is to use a randomized strategy and no deterministic predictions within the game or as side bets are possible. If essential unpredictability is defined in terms of strictly deterministic predictions then the existence of this case certainly supports Scriven's thesis. Simple competitive games of imperfect information played well by both players provide immediate examples of essential unpredictability in the sense of strictly deterministic predictions.

Case 3A: We give up the hope for deterministic predictions, as has already been done in much of empirical science, and ask what are the possibilities of probabilistic predictions. In the game of imperfect information played by C and P it is now possible for P to make a side bet of a probabilistic sort concerning the behavior of C. For the payoffs as defined above, and in the environment in which C is playing the game seriously, P would do well to bet that C will use a minimax randomized strategy. P can profitably balance his bets to come out well in either case. If C does use this strategy, P will win the side bet. If C does not use this strategy, P can win at the game by taking advantage of C's departure from a minimax strategy.

Case 3B: C becomes aware that P is making side bets about his, C's probabilistic behavior in the original game and decides to take the falsification of P's side bets more seriously than the original game. In this case we enter a new game and the same analysis can be repeated, although the strategy space is now a much enlarged one. Having fixed this new game, P can in the same way enter upon side bets concerning C's behavior in the new game.

This enumeration of cases is not meant to be exhaustive. More im-

portantly, the talk about games and side bets can be replaced by talk about motives and the study by *P* of *C*'s motivations and goals. My point thus far has been twofold. In the first place, I agree with Scriven that the deterministic prediction of human behavior is impossible in many situations. Simple competitive games provide perhaps the best examples of such situations. But, secondly, I maintain that probabilistic prediction is not only possible but feasible for a very wide class of situations. If classical physics occupied the position it did at the end of the nineteenth century, modern behavior theory would undoubtedly be criticized and judged inferior for its probabilistic rather than its deterministic character. In actual fact, however, the kind of probabilistic behavior theory which I would now like briefly to discuss, is in general methodology very similar to quantitative theories in physics. Modern physical theories are no longer deterministic but thoroughly probabilistic in character. It is not possible to argue effectively that probabilistic theories are needed for specifically human behavior, whereas fully deterministic theories are appropriate and adequate for nonhuman or inanimate behavior. To substantiate the claim that a predictive behavior theory and predictive behavior laws can be developed, what is wanted, it may be said, is the kind of predictive quantification found in physics: From a knowledge of the present state of the organism and its environment we should be able to predict its future state, at least in a probabilistic sense, for the not too distant future and for a moderately diverse even if restricted set of environmental conditions.

Mathematical behavior theory, as developed in recent years by psychologists, provides perhaps the most important example at present of the kind of theory just described. It would be out of place to describe in detail the nature of this theory, but its main features can be sketched as well as some of the problems it successfully handles. It should be emphasized from the start that modern behavior theory is thoroughly probabilistic rather than deterministic. Abandoning the development of a deterministic theory has undoubtedly been one of the main reasons for the considerable number of quantitative successes of the new theory. It is recognized that the very large number of underlying physiological mechanisms cannot yet be connected in an explicit and detailed way to overt behavior. On the other hand causal relationships of a probabilistic character, holding between behavioral variables, have been postulated and approximately verified.

A causal quantitative analysis of human behavior is the sort of thing often considered impossible in principle. It will be useful to consider one simple but fundamental example of a learning experiment that may be very satisfactorily analyzed within the framework of mathematical behavior theory.

The task for a subject in this experiment is to learn the arbitrary associations set by the experimenter between a given list of nonsense syllables and two telegraph keys. On each trial the subject is shown a nonsense syllable, for example, *XUH*, and then asked to punch that key with which it is associated. After one of the keys is punched, the trial is terminated by the experimenter flashing a light, let us say, or using some other similarly simple device, to show the subject which key was in fact the correct one. Of course, on the first trial it is necessary for the subject simply to guess which is the correct response to make, that is, which key to press, but as would be expected, the subject soon uses the information furnished by the lights above the keys. Even the most gross empirical analysis of the data from such experiments shows that the relation between receiving the information as to which key is correct and behaviorally learning to respond correctly is not deterministic. For example, neither the number of errors nor the trial number of last error is the same for all the nonsense syllables. On the other hand, a very simple probabilistic behavior theory gives an excellent account of the data. The intuitive idea is that each nonsense syllable may be represented as a single stimulus. At the beginning of the experiment no stimulus is conditioned or connected by an association bond to its correct response. It is postulated that on each trial there is a constant probability that the appropriate association bond will be formed. It is also postulated that until this association bond is formed the subject is simply guessing the correct answer. After the correct association is formed, the subject makes the correct response on all subsequent trials. From these postulates it is possible to derive a large number of probabilistic predictions about the behavior of subjects, and these predictions have been confirmed to a remarkably accurate degree in a number of experiments. (See, e.g., Bower, 1961; Suppes and Ginsberg, 1963.)

Now those who challenge the very idea of quantifying human behavior will have at least two objections to this simple learning experiment. First, they will wish to point out that in the usual experiment of this kind data

are analyzed for an entire group of subjects together and not in terms of individual behavior. Their line of criticism is then, "Yes, we will admit it is possible to develop some sort of theory for the gross characteristics of group behavior, but it is not possible to do this for the fine nuances and differences of individual behavior. Those characteristics that are special for a given individual can in no sense be accounted for by the kind of theory you describe, which is wholly devoted to the average or norm of the group." However, this view represents both a misunderstanding of the experiment and the theory used to explain it. For several kinds of reasons, many of which are administrative rather than scientific, it is often convenient to run learning experiments for a short period but with a large number of subjects. The theory however applies just as well to the learning behavior of a single subject over a number of hours. In our laboratory at Stanford we have been performing such extensive experiments on individual subjects and, as a result, analyzing and testing the theory in terms of individual behavior. A typical example of such an experiment is one in which the subject is asked to learn a large number of vocabulary items from a foreign language. In one experiment, for instance, the subject hears a Russian word and then is shown three English words. His task is to select the English word that has approximately the same meaning as the Russian word he has heard. The experimental setup and the theory used to analyze the data are very simple and direct extensions of the experiment with nonsense syllables already described. The important point is that we can ask in a direct and meaningful way if the behavior of an individual subject over the course of fifteen or twenty days satisfies the theory.

The second and perhaps more important objection to the experiment with nonsense syllables, or the second experiment with Russian vocabulary, is that although principles of stimulus-response association may be used to explain the simplest sort of learning, it is not possible to use these principles to develop a quantitative theory for more complicated human behavior. It will perhaps be claimed that no simple principles of association between stimuli and responses can possibly account for the learning of even the most elementary concepts, in particular, mathematical concepts, the learning of which must go beyond any completely simple principles of stimulus association. For the past several years we have conducted extensive experimental investigations on the learning of mathematical

concepts by young children. A typical instance of one of these experiments is the following. Children of kindergarten age are asked on each trial to push one of three keys. Above the first key is shown a small triangle, above the second a small quadrilateral, and above the third a small pentagon. On each trial the child is shown a large figure that is either a triangle, quadrilateral, or pentagon and is asked to push the appropriate key. The child cannot learn the concept of triangle, quadrilateral or pentagon in this experiment by any simple principles of stimulus association, because on each trial he is shown a different figure, that is, the same triangle, the same quadrilateral or the same pentagon is never shown on more than one trial. On each occasion when a quadrilateral, for example, is shown, the child is presented with a new quadrilateral having a different shape and orientation from any of the previous ones he has seen. On the other hand, the kind of theory described above gives a very good quantitative account in probabilistic terms of the data of this experiment, provided that the role played earlier by the nonsense syllables or Russian words is now played by the three simple geometrical concepts. In other words, the association is established between the concept and the appropriate response rather than between a single stimulus and the appropriate response.

In claiming that we are adequately predicting quantitative aspects of the behavior of children in learning mathematical concepts, I would not want to be interpreted as saying that we have anything like the final story on such behavior, for there are many deep questions concerned with how the concepts are formed, or with how the concept-response association is established, which we cannot yet answer. For the present purpose, however, the important point is that we do have sufficient theory to provide a very satisfactory quantitative analysis of the behavior.

Contrary to Scriven's casual claim that geography and geology rather than physics are the appropriate models for psychology, I wish to assert, and have tried to show, that we are already well on the way to a mathematical behavior theory adequate for the analysis of many simple experiments and of such a character that it has the kind of mathematical viability and predictive "feel" about it that we expect of physical theories. Let me conclude by saying that the kind of behavior theory I have described has very good philosophical roots. The associationist psychology of Hume is an obvious precursor. Hume, I feel, would be very happy with

the all-or-none learning laws now well confirmed in a variety of experiments concerned with stimulus-response associations or the formation of simple concepts.

NOTE

[1] These comments were read on May 2, 1963 at the annual meeting of the Western Division of the American Philosophical Association at Columbus, Ohio. They were directed at Scriven's paper, 'An essential unpredictability in human behavior'. In E. Nagel & B. B. Wohlman (Eds.), *Scientific Psychology*. New York: Basic Books, 1965. Pp. 411–425. I have very briefly abstracted his main argument in order to make my remarks intelligible without familiarity with his paper.

18. BEHAVIORISM*

I. WHAT IS BEHAVIORISM

I should like to begin by characterizing in a very preliminary way my own conception of behaviorism. Before beginning this characterization, there is something I want to say about the kind of definition I expect to develop. It is philosophically important to be quite clear about the kinds of things or concepts for which it is possible to offer a precise definition and those for which it is not. Consider, for example, the definition of a physical concept like that of mass. It depends on an exact characterization of mechanics or some other branch of physics. On the other hand, the problem is quite different if we are asked to give a definition of physics or of psychology. The source of difficulty in the latter kind of case is that we do not have a well-defined and sufficiently large and flexible formal framework within which we can formulate a definition of physics or of psychology. Of course, it is not simply a problem of the breadth or flexibility of our general framework, but also a problem of the very vagueness and looseness of what we want to mean by physics or psychology. The concept of behaviorism is in many respects much closer to the vague concept of what is physics or what is psychology than to the much more precise concept of mass or of a prime number. For this reason, I shall not attempt in this preliminary discussion of behaviorism to sketch a possible formal definition. I shall, however, have something to say about the characterization of behaviorism as a formal theory.

One initial way to distinguish behaviorism from other approaches to the study of human beings is in terms of the vocabulary used. In behavioristic discussions of human actions or attitudes there continually recur words like 'stimulus', 'response', 'conditioning', 'discrimination', and 'reinforcement'. On the other hand, those who are critical of a behaviorist approach, or those who feel it is not adequate to account for

* This paper has not been previously published. Except for minor revisions it was written during the period 1963–65.

all kinds of human behavior, will emphasize such words as 'intention', 'belief', 'purposive behavior', 'rule-following behavior'. Many of those who like to use these latter terms in a systematic way will favor Brentano's thesis that intentional sentences are required for the description of psychological phenomena, but not for the description of physical phenomena. This formulation of a thesis in terms of sentences is most characteristic of philosophers, but highly unusual for psychologists. An early behavioristic psychologist like John B. Watson, for instance, would scarcely understand the talk about intentional sentences. For him the issue is very clear cut between introspective or subjective psychology of the sort represented by James and Freud, on the one hand, and behavioristic psychology on the other. There has been in the past, though not so much currently, a very considerable literature on behaviorism written by psychologists. Probably none of this literature formulates criteria of behaviorism in terms of the kind of words that occur in sentences, or the kind of sentences that are uttered by scientists. Aversion to talking about sentences is not peculiar to psychologists but is common to scientists of all stripes. For example, physicists who debate the merits of field vs. non-field theories of matter, or who argue about contact-action vs. action-at-a-distance theories of electrodynamics, never formulate the issues in terms of the kind of sentences being uttered. This is true even for such discussions in mathematics, with the exception of those conducted by professional logicians.

Whatever the merits of the linguistic approach to the analysis of behaviorism, it is certainly widespread among philosophers, and there is a very common tendency to convert talk about intentional sentences into talk about intensional sentences. Thus, statements like "John believes that there are lions in Alaska" are not only intentional but also intensional; for the truth conditions of such belief sentences, it is commonly, and I think correctly said, do not satisfy the ordinary extensional truth-functional logic. We mean by this that knowing whether or not there are lions in Alaska does not enable us to determine the truth or falsity of the sentence asserting that John believes that there are lions in Alaska.

The widespread and subtle use of intentional sentences in ordinary talk is not something I see any reason for attempting to exorcise. The task for the behaviorist presumably is to provide an analysis of the truth conditions for such sentences in nonintentional terms. A large and subtle

literature of more than two decades shows clearly enough that this is not a simple or straightforward matter. All the same, I am not at all pessimistic about such an analysis ultimately being given. Later on, I shall attempt to indicate how I think nonintentionally formulated truth conditions for belief sentences can be given. For the present it is only to the point to mention it as a problem.

An apparently very different approach to the issues between the behaviorist and the intentionalist is for the behaviorist to ask the intentionalist at what point in the developing complexity of psychological phenomena, the phenomena become intentional in character. For instance, does the simplest sort of conditioning of a paramecium require intentional concepts for an adequate description? It is interesting to find writers like Chisholm wavering on this point. With the increasing progress of biology and experimental psychology it is surely a difficult thesis to maintain that every kind of conditioning of living organisms requires intentional concepts for their adequate description. On the other hand, if the simplest sorts of conditioning do not require such intentional concepts, it is not easy to say just when intentional concepts do enter. Yet as often occurs with such scientific problems, a case can surely be made by the intentionalist even if he is not able to classify precisely all psychological phenomena into two parts, one of which requires intentional concepts for adequate description and the other of which does not. He can admit that the position of the line which he would draw to make a distinction between the two kinds of phenomena is vague; still he can properly maintain that the concept of intention, and possibly also the concept of the recognition of intention by one organism in relation to another, is central to more complex psychological phenomena and cannot be eliminated or reduced to purely behavioristic terms. It is, of course, part of behaviorism to claim that such a reduction can in fact be made.

The committing of an intentional action and the recognition by another organism that an intentional action has been performed are in many cases closely and intimately related to the use of language. It is another aspect of behaviorism to maintain that linguistic behavior can be analyzed in the same terms as nonlinguistic behavior, without the introduction of any new fundamental or primitive concepts.[1]

Although I do not want to attempt to give a formal definition of behaviorism, the discussion of the kind of issues already mentioned can

be pursued much more thoroughly and deeply if a rather definite theoretical commitment about the nature of behaviorism is made. What I propose to do is to formulate a theoretical framework within which the analysis of the issues will be made. I would not claim that this theoretical framework encompasses all that is often meant by behaviorism, or even that it is adequate to the analysis of all problems that constitute central challenges to the behavioristic philosopher or scientist. In the next section I shall give a sketch of the theory, but before turning to that sketch there are certain preliminary distinctions I would like to make.

One distinction treacherously easy to forget is that between making an analysis in terms of a theory and making an analysis in terms of known experimental or empirical facts. Consider, for example, the problem of formulating in extensional behavioristic terms truth conditions for belief sentences. On the one hand, we can attempt this in an informal and intuitive fashion by attempting to describe in a rough way the kind of empirical facts we think can be used to provide the basis for such an analysis. By anecdote, illustration, and reference to the vaguely defined thing called the paradigm case, we can attempt to elucidate how we think such an analysis would go. From a formal standpoint this approach will inevitably be unsatisfactory. An alternative is to attempt to proceed within a well-defined theory. In this case, we would ultimately want an analysis possessing the same kind of formal clarity and rigor that are characteristic of Tarski's definition of truth for mathematical languages. What we gain in precision in this case will perhaps be lost in terms of generality and detailed analysis of particular cases. At this stage a schematic formal characterization is the very most that could be expected, but even in schematic form the formal analysis has the virtue of making clear the weak points as well as the strong points of the proposed behavioristic analysis. In the present case there are, it seems to me, certain difficulties besetting a formal analysis that do not usually arise. If, for example, one presents a formal axiomatization of some branch of physics, the formal properties of the physical concepts occurring in the formal statement of the theory are usually well enough known to permit a reasonable evaluation of whether or not the formalization of the theory is intuitively adequate. It is understood, of course, that the formal statement of theory does not itself make clear how the theory is to be interpreted in relation to experiment. The job of connecting the formal theory to

experiment is itself an elaborate and complicated business requiring a detailed methodological theory in its own right. The peculiar difficulty on this score in the discussion of psychological phenomena, particularly in connection with the use of intentional concepts, is that the intentionalist may claim that the behaviorist's extensionally stated formalism will require intentional concepts in order to provide an adequate interpretation in terms of experiments. I do not think there is any simple way to meet this objection. It can be met adequately only by elaborating a methodological theory of the relation between theory and experiment, and it would take us too far afield to attempt to state in detail such a theory. In the meantime, I am prepared to accept intentional criticisms of my use of theory. (For those interested in how I would begin to formulate the methodological theory to exorcise at this new level the intentionalist ghost, I mention the first two articles reprinted in this volume, 'A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences' and 'Models of Data'.)

My second preliminary point is that the kind of details required of a formal theory of behavior are rather different if we are pursuing particular scientific problems rather than philosophical problems. For the purpose of philosophical analysis many of the particular details of the theory can be omitted; or put another way, detailed formulation of particular axioms about conditioning or responses is not especially relevant to the problems of philosophical analysis. A typical example of a scientific issue, not particularly relevant to the problems of philosophical analysis about behaviorism, is the currently much discussed issue as to whether learning occurs on an all-or-none or incremental basis. From the standpoint of the precise axioms of learning or of conditioning it is a critical matter for many experiments; but it is hard to see that this issue in the psychological literature has much bearing on the philosophical issues posed to behaviorists by the intentionalists. On the other hand, many of the definitions needed for a detailed analysis of intentional acts are not the sort that are of much interest to experimental psychologists. This lack of interest is not because the definitions are too banal or too vague, but because they move in a direction of detail and precision which is either unfamiliar to psychologists or at the present development of scientific work uninteresting to them. The point I am trying to make is that the formulation and use of the theory are rather different when we are dealing with particular

scientific issues than when we are dealing with fundamental philosophical issues. It might be thought by some people that if behaviorism had yet received an adequate theoretical formulation, there would be no problem of the kind I am posing; for there would be one well-defined formulation of the theory, and this theory would be appropriate for all purposes either scientific or philosophical. I wish indeed that this were the case, but as in all areas of currently developing science, the status of theory is not so well defined nor so beautifully etched in detail. It is only for theories long established and now static in character, like classical mechanics or classical electromagnetism, that such an agreed-upon, detailed formulation can be given. Behavioristic psychology is as yet too new a science and at present too rapidly developing to hope to be able to formulate a theory adequate for all purposes. I would like to formulate the theory so that it is general enough to be used with some precision for the analysis of philosophical problems, and yet contains axioms which are in themselves not the sort that will easily be refuted by new experiments performed the day after tomorrow. I know that it is not possible to achieve this objective, but in principle this is what I would hope to do.

II. SKETCH OF FORMAL THEORY OF BEHAVIOR

The key concepts of the theory are those of stimulus, response and reinforcement. The detailed axioms of the theory, which we shall not consider here, are based on the following postulated sketch of the sequence of events on a given learning trial.[2] First, a certain set of stimuli is presented to the organism. From this set, the organism samples a certain subset. On the basis of the conditioning connections or associations, sampled stimuli and possible responses, a response is made – in the detailed theory it is postulated that the probability of response is simply the proportion of sampled stimuli connected to this response. After a response is made, a reinforcement is given indicating which response was in fact the correct one. On the basis of this reinforcement the sampled stimuli may become reconditioned, that is, a new association between stimuli and responses is set up and the organism enters a new state of conditioning ready for the next trial. Before we go any further, some comment is needed about this talk of trials. Philosophers will be inclined to ask what the discussion of trials has to do with the problems of analysis

confronting us. The answer is that if we attempt a formal characterization of the theory we are naturally led to a characterization closest to that of the majority of experiments performed to support the theory. There is no logical requirement that the concept of conditioning in learning be imbedded in a theory formulated in terms of discrete trials. It is in fact possible to give a continuous-time formulation, but this is a matter of technical rather than conceptual interest. Talk about discrete trials simplifies the job of constructing formal models for the analysis of experiments.

It will be useful to consider an example of the sort of thing the theory ought to cover. We may begin with the familiar Pavlovian conditioning of a hungry dog. This example, as we shall see later, is particularly interesting because Chisholm has claimed it is one of the simplest kinds of psychological phenomena to require intentional concepts for an adequate explanation. In this conditioning experiment the dog comes to salivate when a bell is sounded, and it is common to say that the bell has become the sign of food. One of the things we want to do within our theoretical framework is to offer a formal definition of one stimulus' being a sign of another. To begin with, putting matters in an informal way, I would say that the bell is a sign of food to the dog after a series of learning trials if essentially the following four conditions are satisfied. First, before any learning takes place, the response of salivation is made by the dog in the presence of the food with approximately probability one. Second, before any learning takes place, the probability of the response of salivation taking place upon the presentation of the bell but not the food is approximately zero. Third, there are a series of training trials in which the food and the stimulus bell are presented simultaneously. On these trials the dog responds to the joint stimuli by giving the response of salivation with approximately probability one. Fourth, after the series of training trials, the dog responds by salivating upon presentation of the stimulus bell alone, with approximately probability one.

I shall later want to make a number of comments about this example because we shall use it in the analysis of the necessity of intentional concepts in describing or explaining psychological phenomena. For the present, however, I wish to use it only to suggest a schema for defining the notion of one stimulus' being a sign of another. For the purposes of providing a formal definition I shall use some familiar simple apparatus

from probability theory. We shall suppose that there is an underlying sample space that represents in a formal way the possible outcomes of the experiments we wish to perform. As is customary in probability theory, it is usually not necessary to describe in detail the formal structure of this sample space, but rather to postulate only the probability laws followed by events defined as subsets of the sample space (or, alternatively, random variables defined on the sample space). In the present case the events we shall consider will all have a clear intuitive meaning. I shall use notation that is familiar in the psychological literature. Let R_n be the event of the response of salivation on trial n, let US_n be the event of the appearance of food, the unconditioned stimulus on trial n, let CS_n be the event of the bell's ringing, the conditioned stimulus on trial n; $\neg US_n$ is the absence of event US on trial n.

With this notation available, we may define CS as a sign of US on trial n. The four conditions in the definiens correspond to the four conditions stated in the particular case of the dog salivating.

(1) $P(R_1 \mid US_1) \approx 1,$

(2) $P(R_1 \mid CS_1 \,\&\, \neg\, US_1) \approx 0,$

(3) For $1 < m \leqslant n$, $P(CS_m \,\&\, US_m) \approx 1 \,\&$
$P(R_n \mid (CS_n \,\&\, US_n) \approx 1,$

(4) For $n' > n$, $P(R_{n'} \mid CS_{n'} \,\&\, \neg\, US_{n'}) \approx 1.$

From the standpoint of philosophical problems confronting behaviorism, there are several general things to be said about this definition and the particular example of canine conditioning, but I shall weave these remarks into the comments on Chisholm's view of intentionalists. However, I do want to note several particular things about this characterization of the salivation experiment. In the first place the definition is so set up that one kind of food familiar to the dog will not be a sign for the other, for according to the second condition, the dog will not salivate in the presence of the sign before the training trials. This would not be true of a different kind of food. Also, conditions (1) and (2) formulate a clear difference between the stimulus and the sign. The fourth condition, on the other hand, excludes most of the stimuli the dog samples in his environment, for he will not salivate upon sampling most of these stimuli.

III. CHISHOLM ON INTENTIONALITY

What I have said thus far about behaviorism will not really satisfy those philosophers who are concerned to maintain the view that much specifically human behavior cannot be adequately expressed or explained in behavioristic terms. The central concept of an intentional or purposeful action is most often cited as an example of a concept that cannot be reduced to the behavioristic notions of stimulus, response, conditioning, and the like. In the remainder of this paper I would like to examine some of the issues surrounding this controversy.

I cannot claim to deal with all aspects of the controversy or even to understand some of them. There are two related but different matters that I feel are particularly relevant. One is the discussion of the impossibility of extensionally or behaviorally defining intentional terms; the other concerns the inappropriateness of causal as opposed to "reason" explanations.

A meticulous and careful defense of the claim that intentional terms used to describe intentional actions cannot be defined extensionally is to be found in the writings of Roderick Chisholm, and it will be sufficient to examine some of his views to express my own attitude to this aspect of the controversy.

Chisholm (1957, pp. 170–171) states three criteria for recognizing intentional sentences. First, a simple declarative sentence is intentional if it uses a substantival expression in such a way that neither the sentence nor its contradictory implies there is anything designated by the substantival expression. For example, "Mr. Bailey hopes to find a three-headed calf to add to his collection of circus animals", "Hilbert wanted to find a decision procedure for the whole of mathematics".

Secondly, a non-compound sentence containing a propositional clause is intentional if neither the sentence nor its contradictory implies that the propositional clause is true or that it is false. For example, "John believes that there are polar bears in Africa", "At one time Hilbert believed that a finitistic consistency proof could be found for the whole of mathematics".

The third criterion is the familiar one that a sentence is intentional when its truth value is disturbed by the substitution of one name or description for another, even though the original phrase and the sub-

stituted phrase designate the same object. Sentences asserting necessity provide familiar examples: "It is necessary that nine is greater than seven. The number of planets is equal to nine. However, it is not necessary that the number of planets is greater than seven".

It is certainly possible to refine the statement of these three criteria, or perhaps to quibble either about their adequacy – whether they cover all cases – or about their classifying as intentional some sentences many people regard as extensional. However, I think they will provide tools of the appropriate degree of precision.

Chisholm then formulates Brentano's thesis that intentional sentences are required for the description of psychological phenomena, but not for the description of physical phenomena. He says the following about the invention of a psychological terminology to describe activities like perceiving in nonintentional sentences.

> Instead of saying, for example, that a man *takes* something to be a deer, we could say 'His perceptual environment is deer-inclusive.' But in doing so, we are using technical terms – 'perceptual environment' and 'deer-inclusive' – which, presumably, are not needed for the description of nonpsychological phenomena. And unless we can re-express the deer-sentence once again, this time as a nonintentional sentence containing no such technical terms, what we say about the man and the deer will conform to our present version of Brentano's thesis [p. 173].

Chisholm then goes on to examine three methods of showing that Brentano's thesis is wrong. He first examines the attempt by Ayer and others to describe psychological attitudes in terms of linguistic behavior. Secondly, he examines an approach in terms of the psychological or behavioral concept of "sign behavior". He cites as typical instances of this viewpoint the work of the psychologist Charles Osgood and the philosopher Charles Morris. Thirdly, he examines the attempt to define intentional concepts in terms of the concept of expectation.

There are two sorts of things I want to say about Chisholm's defense of Brentano's thesis. In the first place, I try to show that his criticism of the "theory of sign behavior" is far too simple and crude. But, more importantly, in the second place, I try to argue that Brentano's thesis and similar doctrines about intentionality are essentially irrelevant to the development of behaviorism as a scientific theory.

Concerning nonintentional definitions of 'sign' Chisholm says that such definitions of 'sign' depend upon the substitution of one stimulus for another, or in more standard psychological terminology on the relation

between an unconditioned and a conditioned stimulus. Chisholm focuses his criticisms of this kind of definition of 'sign' on the difficulty of characterizing the respect or degree of similarity between the sign and the stimulus for which it is the substitute. Chisholm summarizes his argument as follows:

> Shall we say that V is a sign of R provided that V has *all* the effects which R would have had? If the bell is to have all the effects which the food would have had, then, as Morris notes, the dog must start to eat the bell. Shall we say that V is a sign of R provided that V has the effects which *only* R would have had? If the sign has effects which only the referent can have, then the sign *is* the referent and only food can be a sign of food. The other methods of specifying the degree or respect of similarity required by the substitute-stimulus definition, so far as I can see, have equally unacceptable consequences [pp. 179–180].

Before stating criticisms of Chisholm's analysis, let me say parenthetically that if we were to take his remarks seriously we would be denying the possibility of an objective scientific description of the ubiquitous psychological phenomena of transfer and generalization in learning. The central weakness of his criticism is to present us with essentially only two alternatives: either the sign produces *all* the effects that the original stimulus itself does, or it produces, in some vague fashion, only *some* of the effects. In the first case the sign needs to be the same kind of event as the original stimulus, and in the second case, any two stimuli share some property with the original stimulus. Another primary difficulty of Chisholm's analysis of the Pavlovian example is that he does not take any account of the fact that a sign must be *learned* as a sign of a stimulus. His *analysis* provides no room for changes in the conditioning of the organism.

I have indicated above how an extensional definition of one thing's being a sign for another can be given. The identification of the events R_n, US, CS and $\neg US$ on any trial on which they occur is as extensional as the similar identification of events in physical experiments, and so also is the concept of probability – the same sense of probability can be used and is used for the analysis of psychological and physical experiments.

I realize that my definition is still partially schematic, but I see no essential difficulty in making it as elaborate and detailed as necessary. The conditions for elaboration are not different from those necessary to spell out in satisfactory detail physical or chemical experiments. The point of the constructive definition of being a sign is simply to show that once even a rather meager quantitative apparatus is introduced, partic-

ularly of a probabilistic kind, then Chisholm's criticisms seem to fall very wide of the mark. In connection with the next to last sentence of the quotation from Chisholm, I note that my definition does not require that the signs have effects only the referents can have, and therefore be led into the absurdity that only food can be a sign of food. We turn our attention only to a certain subset of responses of the organism taking place at that time, just as in a physical experiment, for example, one dealing with the behavior of magnets, we examine only some behavior of the magnet and do not concern ourselves with such problems as the shadow cast by the magnet on the laboratory bench, etc. In the salivation experiment, we do not examine the responses that take place in the organism as the food is ingested, nor do we take account of the motions of the mouth in chewing the food, etc. For other purposes it may be desirable to consider such responses, but in every experiment, whether it is an experiment on the behavior of organisms, or the behavior of inanimate things, we are never concerned with all aspects of that behavior, but only with a very selected portion. In this respect, too, it seems that Chisholm's criticisms have gone badly astray.

I turn now to my argument against the relevance, for the development of behaviorism as a scientific theory, of Brentano's thesis or similar arguments in favor of the irreducibility of intentional concepts. I maintain that it is not essential to provide an adequate definition in behavioristic terms of intentional notions in order to develop a quantitative theory of behavior. And this is not because the intentional concepts are wrong or inapplicable to the discussion of behavior, but rather because an introduction of new distinctions and concepts does not require as a necessary prolegomenon the analysis without remainder of the concepts already in the field. The view that such a definitional analysis is needed is based on a kind of completeness claim that cannot be supported in empirical domains. Systematic terms or concepts of any empirical theory are fantastically incomplete or schematic. Whether we look at a theory of learning or a theory of mechanics the point is the same. The concepts of the theory are only loosely connected to any actual experiment. The point I am making is not new; it has been made by many others and I have tried to amplify it elsewhere.[3] From the standpoint of sophisticated common sense or the previous theory in the field, the concepts of a new theory may be unbelievably crass and crude in their analysis of the nuances of experience.

This is certainly true of the 17th-century theories of mechanics in relation to their Aristotelian predecessors. It was not taken as a responsibility of these new theories to analyze out in terms of the simple Cartesian and Newtonian concepts the many subtle distinctions introduced by Aristotle and the Scholastics in describing physical phenomenon. Indeed, this would have been a hopeless enterprise. Even so simple a distinction as that of Aristotle's between natural and violent motion would seem to have no clear definition in the Cartesian or Newtonian theories. The Aristotelian doctrines of the potential and the actual and of form and matter had even less chance of being explained.

A small-scale example of such irreducibility is to be found in the development of theories of choice and quantitative theories of belief. One appropriate way of describing modern theories of subjective probability is in terms of the attempt to develop a quantitative theory of belief – partial belief, if you will. It is not feasible to describe here the work that has been done on this subject by Ramsey, de Finetti, Savage, and others. The only point I want to make is that the rather considerable quantitative development of this theory has been made in blithe independence of Brentano's thesis. The problems that still beset the theory do not seem primarily to involve problems of intentionality. Consider, for example, the analysis of the degree of belief in terms of the two-place relation *equal to or less probable than.* If we use the ordinary formalism familiar in probability theory, then this relation is said to hold between two events, although in older writers it would be said to hold between two propositions. I think it is possible to object, on intentional grounds, to the use of the event-language for it may be maintained that the *description* of an event is an important determiner of the degree of belief we assign to an event. My answer to this is twofold. On the one hand, if we are interested in studying beliefs as opposed to the assertion of belief statements, a good case can be made for concentrating mainly on non-verbal behavior. The oft-emphasized rubric is that the true indication of a person's beliefs are the actions he takes and not the statements he makes. Numerous recent writers on subjective probability have emphasized this behavioristic point about measuring degrees of belief. This viewpoint is very well expressed but unfortunately thereafter abandoned in the opening lines of Hare's *The Language of Morals* (1952):

If we were to ask of a person 'What are his moral principles?' the way in which we could

be most sure of a true answer would be by studying what he *did*. He might, it is true, profess in his conversation all sorts of principles, which in his actions he completely disregarded; but it would be when, knowing all the relevant facts of a situation, he was faced with choices or decisions between alternative courses of action, between alternative answers to the question 'What shall I do?', that he would reveal in what principles of conduct he really believed. The reason why actions are in a peculiar way revelatory of moral principles is that the function of moral principles is to guide conduct [p. 1].

In view of these opening lines, it is somewhat surprising to find that almost the entire remainder of Hare's book is devoted to an analysis of the language of morals and not to the development of any theory of moral decisions or actions themselves.

In line with Hare's dictum, the surest way to understand a man's moral principles is to study what he does. Over the past 15 years there have been a large number of experimental studies that are more or less relevant to formal theories of preference and choice (for a review of this experimental literature see Luce and Suppes, 1965). It is characteristic of this literature that it is directly concerned with testing behavioristic theories of choice, and secondly, that problems of intention of the sort raised by Chisholm and other writers do not impinge in any systematic way on the design and execution of the experiments. I do not mean to suggest that this experimental literature makes it a point specifically to deny the correctness or appropriateness of intentional concepts in describing much human behavior. It is rather, as I have already indicated, that these difficult and subtle concepts are bypassed and ignored in the new formal developments of a theory of behavior. As a sample of the kind of empirical findings that are coming out of these studies, let me mention just one that is of some generality. The study of betting behavior both in experimental settings and in real-life settings at the race track show that under a wide variety of circumstances there is a very strong tendency on the part of most people to underestimate high probabilities and to overestimate low probabilities. This kind of finding says something important about the belief structures of average people, and yet obviously does not depend in any way on intentional notions. Perhaps another way of putting a criticism of intentionality theses like those of Chisholm's is that it is not made clear what the relation of the thesis about intentionality is to the scientific study of intentional behavior. The sense in which a scientific theory of human behavior must be intentional for Chisholm is not at all clear.

My second point of emphasis is to stress my conviction that the

peculiar character of intentional contexts to be found in belief statements and in other kinds of modal statements that may be cited in support of Brentano's thesis will disappear once a properly detailed behavioristic analysis of language is given. As a very preliminary indication of how I would conceive doing this, we may consider belief statements about the dog's response to the ringing of the bell. Without quibbling about exact perceptual details, I think we may agree that the event of the bell's ringing in the dog's presence is identical with the event of certain sound waves of a specified range of intensity and frequency reaching the auditory receptor organs of the dog. We do not raise problems about substitutibility in asserting, on the one hand, that the dog believes that the ringing of the bell in his presence is a sign that food is to follow, and the parallel assertion that the dog believes that the reception of certain sound waves of a given frequency and intensity range is to be followed by his receiving food. At least, I do not find any difficulties of substitutivity here; and the central reason, I feel sure, is the fact that the dog is not a language user. Yet I would maintain that belief statements about the dog or other non-human mammals are unexceptional. The difficulty with belief statements about human users of complicated language is that a variety of signs are used to encode beliefs, and in a fully detailed analysis it is necessary not only to describe the event about which the belief is held but also the encoding signs. I would defend the thesis that it is impossible to have a belief without such specific encoding. (Whether the organism is explicitly conscious of the encoding is an irrelevant matter.) The essential point is that in terms of the specific encoding one should in principle be able to offer a general definition of truth for belief statements. Thus, if John says, "I believe that there are lions in Alaska" the truth of this statement would be defined in the classical Tarskian manner except that the model is not now a direct model of the real world, but a model of the encoded beliefs of John. It is quite true that the methods for determining whether or not belief statements are true or not will usually be highly indirect, but of course this is true for many statements of other sorts as well. Secondly, the problems of vagueness in belief are not different from the problems of vagueness in ordinary statements. There is no more difficulty in principle in deciding on the truth of a vague statement about belief or a precise statement about vague beliefs than there is in discriminating between the truth of "John ran quickly" and "John ran very

quickly". Perhaps the following analogy will make what I am trying to say about the truth of belief statements clearer. The method of testing the truth of belief statements I am suggesting would be similar to the following test by a computer. The computer has certain information about the world stored in memory. When a sentence is handed to the computer it then applies a Tarskian definition of truth to check the truth of the new statement in terms of the information stored in memory. What is important for the truth of the belief sentence for the computer is the information stored in memory, not correct knowledge about the world. I certainly do not think that what I have said here in this brief way about a nonintentional definition of truth for belief sentences has been sufficiently detailed to solve the many puzzling questions raised about belief statements in the recent literature. I do think it points in the right direction, and in particular, is much closer to the intuitive content of belief sentences than linguistic accounts that involve translating belief sentences in ordinary parlance into statements about belief in certain sentences.

IV. TYPES OF EXPLANATION

I mentioned earlier that an important aspect of controversies surrounding the development of a quantitative theory of human actions concerns the inappropriateness of causal as opposed to "reason" explanations. I had hoped to have time in this paper to devote a fairly detailed effort to refuting the kind of claims that are typified in the arguments given by R. S. Peters in his book on philosophical psychology (1958). Briefly put, Peters' argument is that the rule-following purposive model of human behavior is always required for an adequate explanation of a human action. He admits that causal explanations are relevant and can on occasion state necessary conditions for an action. His argument is that we can never give sufficient conditions in causal terms for a human action, because "we can never specify an action exhaustively in terms of movements of the body or within the body" [p. 12]. As an example of a human action that cannot exhaustively be described in terms of physical movements, Peters mentions the act of signing a contract. He points out the many different ways in which the pen may be held, how the size of the writing or the time taken to finish the signature may vary, etc. His entire position is well summarized in the passage following this example.

So we could never give a sufficient explanation of an action in causal terms because we could never stipulate the movements which would have to count as dependent variables. A precise functional relationship could never be established. Of course, just as we could stipulate a general range of movements necessary to define signing a contract, so also we could lay down certain very general *necessary* conditions. We could, for instance, say that a man could not sign a contract unless he had a brain and nervous system. Such physiological knowledge *might* enable us to predict *bodily movements*. And *if* we had bridging laws to correlate such physiological findings with descriptions of actions we might *indirectly predict* actions. But we would *first* have to grasp concepts connected with action like 'knowing what we are doing' and 'grasp of means to an end'. As such concepts have no application at the level of mere movement, such predictions would not count as sufficient *explanations of actions* [pp. 13–14].

It seems to me that this passage reflects a profound misunderstanding of the nature of scientific method in the physical sciences as well as in the psychological and biological sciences. If we were to take these strictures correctly, no causal explanation in macroscopic physics would be acceptable; in fact, no adequate causal explanation could be given of any physical phenomena at the macroscopic or microscopic level involving motions and interactions of a large number of particles, because we are not now able, and probably never shall be able, to state a precise functional relationship between the motions of the individual particles and the observed macroscopic phenomena. Consider, for example, the thermodynamical and mechanical explanation of the formation of clouds on the windward side of a mountain. This explanation is given in terms of the upward motion of an incredibly large number of air and water vapor molecules. Some general characteristics of this motion can be stated, for example, the mean velocity of a molecule, but it is utterly hopeless to attempt to give any account of "precise functional relationships" between the motions of individual particles and the cloud we can all observe.

Perhaps my deepest objection to what Peters says is that, like Chisholm, he does not seem to recognize the highly schematic character of the causal explanation of any phenomena, animate or inanimate. It is, I would claim, never possible to give a direct characterization of sufficient conditions for the occurrence of a phenomenon. The concept of sufficiency is relative to our description of the phenomenon, and the adequacy of a causal explanation must also be judged relative to that description. There is indeed no such thing as an ultimate causal analysis of any phenomenon. Behaviorism and quantum physics are in the same causal boat afloat on a probabilistic sea.

NOTES

[1] For a detailed discussion of this point see the final article in this volume.
[2] See the last article of this volume for a detailed and formal set of axioms.
[3] Article 2 in this volume.

19. ON THE BEHAVIORAL FOUNDATIONS OF MATHEMATICAL CONCEPTS*

I. INTRODUCTION

The title of this paper will perhaps mean different things to different people. Philosophers and mathematicians interested in the foundations of mathematics and the philosophy of language may think I intent to pursue a systematic pragmatics built around such notions as Ajdukiewicz' concept of acceptance. Actually, I am going in a different direction. What I want to do is outline present applications of mathematical learning theory to mathematical concept formation. The aims of this paper are primarily constructive, that is, to contribute to the development of a scientific theory of concept formation. Before I turn to this subject, however, I want to comment on two general aspects of the teaching of mathematical concepts.

The first concerns the much-heard remark that the newer revisions of the mathematics curriculum are particularly significant because of the emphasis they place on *understanding* concepts as opposed to the perfection of *rote* skills. My point is not to disagree with this remark, but to urge its essential banality. To understand is a good thing; to possess mere rote skill is a bad thing. The banality arises from not knowing what we mean by *understanding*. This failure is not due to disagreement over whether the test of understanding should be a behavioral one. I am inclined to think that most people concerned with this matter would admit the central relevance of overt behavior as a measure of understanding. The difficulty is, rather, that no one seems to be very clear about the exact specification of the behavior required to exhibit understanding. Moreover, apart even from any behavioral questions, the very notion of understanding seems fundamentally vague and ill defined.

To illustrate what I mean, let us suppose that we can talk about under-

* Reprinted from *Mathematical Learning* (Monographs of the Society for Research in Child Development, 30, Serial No. 99) (ed. by L. N. Morriset and J. Vinsonhalers), 1965, pp. 60–96.

standing in some general way. Consider now the concept of triangularity. Does understanding this concept entail the understanding that the sum of the interior angles is 180°, or that triangles are rigid whereas quadrilaterals are not, or the ability to prove that if the bisectors of two angles of a triangle are equal then the triangle is isosceles? This example suggests one classical philosophical response to our query, that is, to understand a concept means, it is said, to know or believe as true a certain set of propositions that use the concept. Unfortunately, this set is badly defined. It is trivial to remark that along these lines we might work out a comparative notion of understanding that is a partial ordering defined in terms of the inclusion relation among sets of propositions that use the concept. Thus, one person understands the concept of triangularity better than a second if the set of propositions that uses the concept and is known to the first person includes the corresponding set for the second person. (Notice that it will not do to say simply that the first person knows more propositions using the concept, for the second person might know fewer propositions but among them might be some of the more profound propositions that are not known by the first person; this situation corresponds to the widely held and probably correct belief that the deepest mathematicians are not necessarily the best mathematical scholars.)

But this partial ordering does not take us very far. A more behavioral line of thought that, at first glance, may seem more promising is the response of the advocates of programmed learning to the charge that the learning of programmed material facilitates rote skills, but not genuine understanding of concepts. They assert that if the critics will simply specify the behavior they regard as providing evidence of understanding, the programmers will guarantee to develop and perfect the appropriate repertory of responses. This approach has the practical virtue of sidestepping any complex discussion of understanding and supposes, with considerable correctness, no doubt, that without giving an intellectually exact analysis of what to understand a concept means, we still can obtain a rough consensus at any given time of what body of propositions we expect students to master about a given concept. This is the appropriate practical engineering approach, but it scarcely touches the scientific problem.

In this paper I do not pretend to offer any serious characterization of what it means to understand a concept. I do think that the most promising

direction is to develop a psychological theory of concept transfer and generalization. The still relatively primitive state of the theory of the much simpler phenomena of stimulus transfer and generalization do not make me optimistic about the immediate future. For immediate purposes, however, let me sketch in a very rough way how the application of ideas of transfer and generalization can be used to attack the banality mentioned earlier in the standard dichotomy of understanding vs. rote skill.

We would all agree, I think, that such matters as learning to give the multiplication tables quickly and with accuracy are indeed rote skills. But there is also what I consider to be a mistaken tendency to extend the label "rote skill" to many parts of the traditional mathematics curriculum at all levels. The body of mathematical material tested, for example, by the British Sixth Form examinations is sometimes so labeled by advocates of the newer mathematics curriculum. In terms of the accepted notion of rote skill developed and studied by psychologists, this is a mistake, for the production of a correct response on these examinations cannot be explained by any simple principle of stimulus-response association. Moreover, the problems of transfer involved in solving typical British Sixth Form examination problems, in comparison with the kind of examination set by advocates of the newer mathematics curriculum may, in fact, require more transfer of concepts; at least, more transfer in one obvious way of measuring transfer, that is, in terms of the number of hours of training spent in relation to the ability to solve the problems by students matched for general background and ability. I recognize that these are complicated matters and I do not want to pursue them here. Also, I am fully in sympathy with the general objectives of the newer mathematics curriculum. I am simply protesting against some of the remarks about understanding and rote skills that occur in the pedagogical conversations and writings of mathematicians.

The second general point I want to mention briefly is of a similar sort. I have in mind the many current discussions of the efficacy of the discovery method of teaching. Such discussions seem to provide yet one more remarkable example, in the history of education, of a viewpoint achieving prominence without any serious body of empirical evidence to support or refute its advocates. From the standpoint of learning theory, I do not even know of a relatively systematic definition of the discovery method. I do not doubt that some of its advocates are themselves remarkably capable

teachers and able to do unusual and startling things with classes of elementary-school children. The intellectual problem, however, is to separate the pedagogical virtuosities of these advocates' personalities from the systematic problem of analyzing the method itself. Workable hypotheses need to be formulated and tested. I know that a standard objection of some advocates of the discovery method is that any quick laboratory examination of this teaching method vs. a more standard immediate reinforcement method, particularly as applied to young children, is bound not to yield an unbiased test. The results and the implications of the methods, it is said, can only be properly evaluated after a long period. I rather doubt that this is the case but, if it is so, or if it is propounded as a working hypothesis by advocates of the method then, it seems to me, it is their intellectual responsibility to formulate proper tests of a sufficiently sustained sort.

I realize that my remarks on this subject have the character of *obiter dicta*. On the other hand, in a more complete treatment of mathematical concept formation in young children, I would consider it necessary to probe more deeply into the issues of motivation, reinforcement and concept formation that surround the controversy between the discovery method and other more classical methods of reinforcement. Some experimental results on methods of immediate reinforcement are reported in Section III, 'Some Concept Experiments with Children'.

I turn now to the specific topics I would like to develop more systematically. In the next section, a version of stimulus-sampling learning theory is formulated that holds considerable promise for providing a detailed analysis of the behavioral processes involved in the formation of mathematical concepts. In the following section, I report in somewhat abbreviated form six experiments dealing with mathematical concept formation in young children. A particular emphasis is placed on whether the learning process in this context is represented better by all-or-none or incremental conditioning. The final section is concerned with behavioral aspects of logical inference and, in particular, of mathematical proofs.

II. FUNDAMENTAL THEORY

The fundamental theory I shall apply in later sections is a variant of stimulus-sampling theory first formulated by Estes (1950). The axioms

given here are very similar to those found in Suppes and Atkinson (1960). I shall not discuss the significance of the individual axioms at length because this has been done in print by a number of people. The axioms I may mention, however, are based on the following postulated sequence of events occurring on a given trial of an experiment: The organism begins the trial in a certain state of conditioning. Among the available stimuli a certain set is sampled. On the basis of the sampled stimuli and their conditioning connections to the possible responses, a response is made. After the response is made, a reinforcement occurs that may change the conditioning of the sampled stimuli. The organism then enters a new state of conditioning ready for the next trial. The following axioms (divided into conditioning, sampling, and response axioms) attempt to make the assumptions underlying such a process precise (they are given in verbal form, but it is a routine matter to translate them into an exact mathematical formulation):

Conditioning Axioms

C1. *On every trial each stimulus element is conditioned to at most one response.*

C2. *If a stimulus element is sampled on a trial, it becomes conditioned with probability* c *to the response (if any) that is reinforced on that trial; if it is already conditioned to that response, it remains so.*

C3. *If no reinforcement occurs on a trial, there is no change in conditioning on that trial.*

C4. *Stimulus elements that are not sampled on a given trial do not change their conditioning on that trial.*

C5. *The probability* c *that a sampled stimulus element will be conditioned to a reinforced response is independent of the trial number and the preceding pattern of events.*

Sampling Axioms

S1. *Exactly one stimulus element is sampled on each trial.*

S2. *Given the set of stimulus elements available for sampling on a trial, the probability of sampling a given element is independent of the trial number and the preceding pattern of events.*

Response Axioms

R1. *If the sampled stimulus element is conditioned to a response, then that response is made.*

R2. *If the sampled stimulus element is unconditioned, then there is a probability p_i that response i will occur.*

R3. *The guessing probability p_i of response i, when the sampled stimulus element is not conditioned, is independent of the trial number and the preceding pattern of events.*

Although not stated in the axioms, it is assumed that there is a fixed number of responses and reinforcements and a fixed set of stimulus elements for any specific experimental situation.

Axioms C5, S2, and R3 are often not explicitly formulated by learning theorists, but for the strict derivation of quantitative results they are necessary to guarantee the appropriate Markov character of the sequence of state-of-conditioning random variables. Axioms of this character are often called *independence-of-path assumptions.*

The theory formulated by these axioms would be more general if Axiom S1 were replaced by the postulate that a fixed number of stimuli are sampled on each trial or that stimuli are sampled with independent probabilities, and if Axiom R1 were changed to read that the probability of response is the proportion of sampled stimulus elements conditioned to that response, granted that some conditioned elements are sampled. For the experiments to be discussed in the next section this is not an important generalization and will not be pursued here. (From the historical standpoint the generalizations just mentioned actually were essentially Estes' original ones.) Nowadays, they are referred to as the assumptions of the component model of stimulus sampling. Axiom S1 as formulated here is said to formulate the pattern model, and the interpretation is that the organism is sampling on a given trial the pattern of the entire stimulating situation, at least the relevant pattern, so to speak. This pattern model has turned out to be remarkably effective in providing a relatively good, detailed analysis of a variety of learning experiments ranging from rats in T-mazes to two-person interaction experiments.

There is one other general remark I would like to make before turning to the discussion of particular experiments. The kind of stimulus-response theory just formulated is often objected to by psychologists interested in cognitive processes. I do not doubt that empirical objections can be found to stimulus-response theory when stated in too simple a form. I am prepared, however, to defend the proposition that, at the present time, no other theory in psychology can explain in the same kind of quantitative

detail an equal variety of learning experiments, including concept formation experiments. I should also add that I do not count as different, cognitive formulations that are formally isomorphic to stimulus-sampling theory. In our recent book Atkinson and I (Suppes and Atkinson, 1960) attempted to show how the hypothesis language favored by many people (e.g., Bruner *et al.*, 1956) can be formulated in stimulus-sampling terms. For example, a strategy in the technical sense corresponds precisely to a state of conditioning and a hypothesis to the conditioned stimulus sampled on a given trial, but details of this comparison are not pertinent here.

III. SOME CONCEPT EXPERIMENTS WITH CHILDREN

I now turn to the application of the fundamental theory, stated in the preceding section, to a number of experiments that are concerned with concept formation in young children. It would be possible, first, to describe these experiments without any reference to the theory, but, in order to provide a focus for the limited amount of data it is feasible to give in this survey, it will be more expedient to specialize the theory initially to the restricted one-element model, and report on data relevant to the validity of this model.

We obtain the one-element model by extending the axioms given in the preceding section in the following respect: we simply postulate that there is exactly one stimulus element available for sampling on each trial and that at the beginning of the experiment this single element is unconditioned.

This special one-element model has been applied with considerable success by Bower (1961) and others to paired-associate experiments, that is, to experiments in which the subject must learn an arbitrary association established by the experimenter between, say, a nonsense syllable as single stimulus and a response, such as one of the numerals 1–8 or the pressing of one of three keys. The most important psychological implication of this one-element model is that in the paired-associate situation the conditioning takes place on an all-or-none basis. This means that prior to conditioning the organism is simply guessing the correct response with the probability p_i mentioned in Axiom R3, and that the probability of conditioning on each trial in which the stimulus is presented is c. Once

the stimulus is conditioned the correct response is made with probability one.

In an earlier paper, Rose Ginsberg and I (Suppes and Ginsberg, 1963) analyzed a number of experiments, including some of those reported here, to exhibit a simple but fundamentally important fact about this all-or-none conditioning model. The assumptions of the model imply that the sequence of correct and incorrect responses prior to the last error form a binomial distribution of Bernouilli trials with parameter p. This null hypothesis of a fixed binomial distribution of responses prior to the last error admits, at once, the possibility of applying many powerful classical statistics that are not usually applicable to learning data. What is particularly important from a psychological standpoint is this hypothesis' implication that the mean learning curve, when estimated over responses prior to the last error, is a horizontal line. In other words, no effects of learning should be shown prior to conditioning. Ginsberg and I analyzed experiments concerned with children's concept formation, animal learning, and probability learning, and with paired-associate learning in adults from this standpoint. I shall not propose to give as extensive an analysis of data in the present paper as we attempted there, but I will attempt to cite some of the results on this question of stationarity, because of its fundamental importance for any psychological evaluation of the kind of processes by which young children acquire concepts.

Other features of the experiments summarized below will be mentioned seriatim, particularly if they have some bearing on pedagogical questions. One general methodological point should be mentioned, however, before individual experiments are described. In many of the experiments, the stimulus displays were different on every trial so that there was no possibility of establishing a simple stimulus-response association. How is the one-element model to be applied to such data? The answer represents, I think, one of our more important general findings: *a very good account of much of the data may be obtained by treating the concept itself as the single element.* The schema, then, is that a simple concept-response association is established. With the single exception of Experiment I, we have applied this interpretation to the one-element model in our experiments.

Experiment I. Binary Numbers

This experiment is reported in detail in Suppes and Ginsberg (1962a).

Five- and six-year-old subjects were required to learn the concepts of the numbers 4 and 5 in the binary number system, each concept being represented by three different stimuli; for example, if the stimuli had been chosen from the Roman alphabet, as in fact they were not, 4 could have been represented by *abb*, *cdd*, and *eff*, and 5 by *aba*, *cdc*, and *efe*. The child was required to respond by placing directly upon the stimulus one of two cards. On one card was inscribed a large Arabic numeral 4 and on the other a large Arabic numeral 5. All the children were told on each trial whether they made the correct or incorrect response, but half of them were also required to *correct* their wrong responses. Thus, in this experiment, in addition to testing the one-element model, we were concerned with examining the effect upon learning of requiring the subject to correct overtly a wrong response. There were 24 subjects in each of the two experimental groups. From test responses, after each experimental session, it seemed evident that whereas some subjects in both groups learned the concept as such, others learned only some of the specific stimuli representing the concepts so that, in effect, within each group there were two subgroups of subjects. It is interesting to note that this finding agrees with some similar results in lower organisms (Hull and Spence, 1938), but is contrary to results obtained with adult subjects for whom an overt correction response seems to have negligible behavioral effects (Burke *et al.*, 1954).

The data for both correction and noncorrection groups are shown in Figure 1. It is apparent that there was a significant difference between the

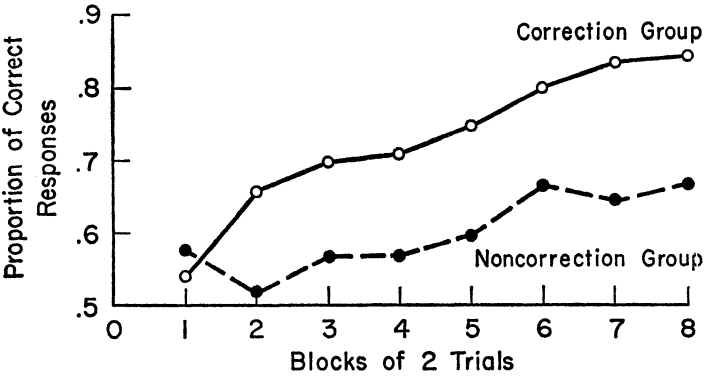


Fig. 1. Proportion of correct responses over all trials (binary-number experiment).

two groups in the rate of learning. The t of 4.00 computed between overall responses of the two groups is significant at the .001 level.

For the analysis of paired associates and concept formation we restricted ourselves to the 24 subjects of the correction group. To begin with, we analyzed the data as if each of the six stimuli, three for each number, represented an independent paired-associate item. In accordance with this point of view, we have shown in Figure 2 the proportion of

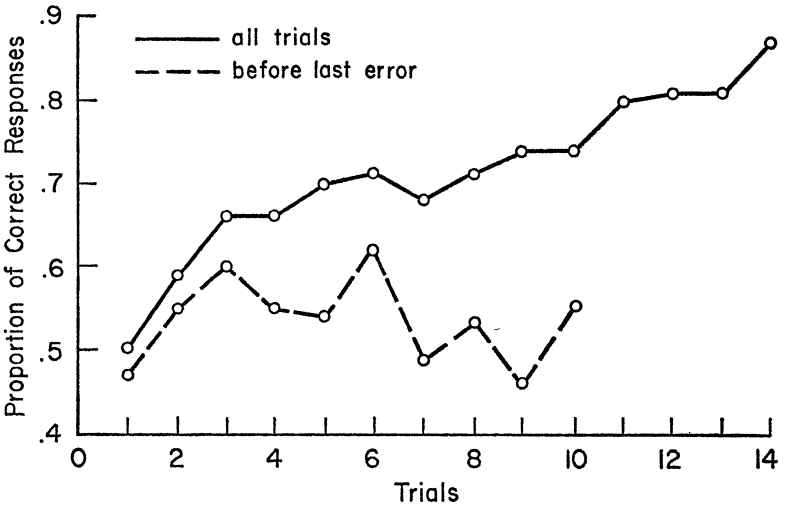


Fig. 2. Proportion of correct responses prior to last error and mean learning curve (binary-number experiment).

correct responses prior to the last error and the mean learning curve for all responses.

The data points are for individual trials. Because a total of only 16 trials were run on each stimulus we adopted a criterion of six successive correct responses, and thus the proportion of correct responses prior to the last error is shown only for the first 10 trials. A χ^2 test of stationarity over blocks of single trials supports the null hypothesis ($\chi^2 = 8.00$, $df = 9$, $P > 0.50$, $N = 844$).

Let us now turn to the question of concept formation. The identification we make has already been indicated. We treat the concept itself as the single stimulus, and in this case we regard the experiment as consisting of

two concepts, one for the number 4 and one for the number 5. (It should be apparent that the identification in terms of the numbers 4 and 5 is not necessary; each concept can be viewed simply as an abstract pattern.)

The criterion for the learning of the concept was correct responses to the last three presentations of each stimulus. On this basis we divided the data into two parts. The data from the group meeting the criterion were arranged for concept-learning analysis – in this case a two-item learning task. The remaining data were assumed to represent paired-associate learning involving six independent stimulus items. For the paired-associate group over the first 10 trials we had 81 cases; for the concept-formation group we had 21 cases with 48 trials in each. The χ^2 test of stationarity was not significant for either group (for the concept subgroup $\chi^2 = 8.36$, $df = 9$, $P > 0.30$, $N = 357$; for the paired-associate subgroup $\chi^2 = 11.26$, $df = 8$, $P > 0.10$, $N = 570$).

To provide a more delicate analysis of this important question of stationarity we can construct Vincent curves in the following manner (cf. Suppes and Ginsberg, 1963). The proportion of correct responses prior to the last error may be tabulated for percentiles of trials instead of in terms of the usual blocks of trials. In Figure 3 the mean Vincent curve for the

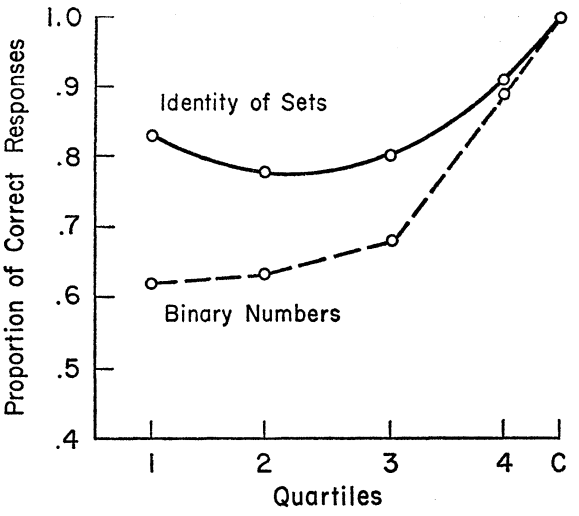


Fig. 3. Vincent learning curves in quartiles for proportion of correct responses prior to last error, binary numbers and identity of sets (Exps. I and II).

subjects in the binary-number experiment who met the concept criterion is shown. The curve is plotted in terms of quartiles. As the mean percentile of each of the four quartiles is 12.5%, 37.5%, 62.5%, and 87.5%, respectively, and C represents the 100% point, the distance between 4, the fourth quartile, and C on the abscissa is one-half of that between the quartiles themselves. The evidence for nonstationarity in the final quartile will be discussed subsequently along with the other Vincent curve shown in this figure.

It should be noted, of course, that the subjects who take longer to meet the criterion are weighted more heavily in the Vincent curves. For example, suppose one subject has 16 responses prior to his last error whereas another subject has only 4. The first subject contributes 4 responses to each quartile whereas the second subject contributes only 1. This point will be discussed in more detail below. I turn now to the second experiment.

Experiment II. Equipollence and Identity of Sets

This experiment was performed with Rose Ginsberg and has been published in Suppes and Ginsberg (1963). The learning tasks involved in the experiment were equipollence of sets and the two related concepts of identity of sets and identity of ordered sets.

The subjects were 96 first graders run in 4 groups of 24 each. In Group 1 the subjects were required to learn identity of sets for 56 trials and then equipollence for a further 56 trials. In Group 2 this order of presentation was reversed. In Group 3 the subjects learned first identity of ordered sets and then, identity of sets. In Group 4 identity of sets preceded identity of ordered sets. Following our findings in Experiment I, that is, that learning was more rapid when the child was required to make an overt correction response after an error, we included this requirement in Experiment II and most of the subsequent experiments reported below. Also, in this experiment and those reported below, no stimulus display on any trial was repeated for an individual subject. This was done in order to guarantee that the learning of the concept could not be explained by any simple principles of stimulus-response association, as was the case for Experiment I. For convenience of reference we termed concept experiments in which no stimulus display was repeated *pure* property or *pure* concept experiments.

The sets depicted by the stimulus displays consisted of one, two, or three elements. On each trial two of these sets were displayed. Minimal instructions were given the subjects to press one of two buttons when the stimulus pairs presented were "the same" and the alternative button when they were "not the same".

Our empirical aims in this experiment were several. First, we wanted to examine in detail if the learning of simple set concepts by children of this age took place on an all-or-none conditioning basis. Second, as the two sequences of learning trials on two different concepts for each group would indicate, we were interested in questions of transfer. Would the learning of one kind of concept facilitate the learning of another, and were there significant differences in the degree of this facilitation? Third, we were concerned with considering the question of finding the behavioral level at which the concepts could be most adequately defined. For example, in learning the identity of sets could the learning trials be satisfactorily analyzed from the standpoint of all trials falling under a single concept? Would it be better to separate the trials on which identical sets were presented from those on which nonidentical sets were presented in order to analyze the data in terms of two concepts? Or would a still finer

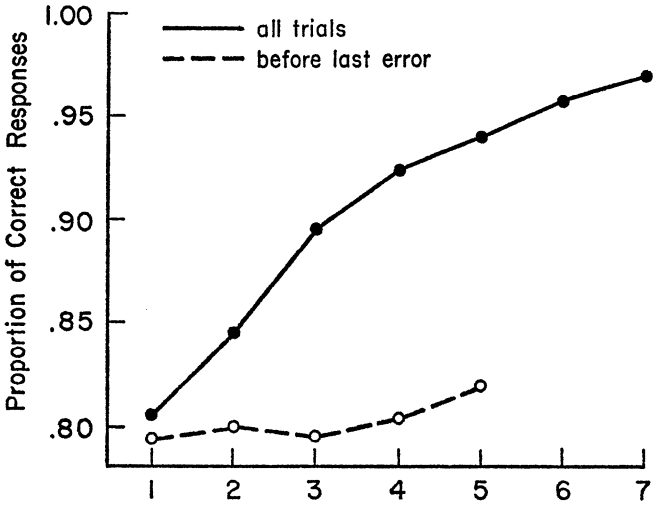


Fig. 4. Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of sets, $N = 48$, Groups 1a and 4a (Exp. II).

division of concepts in terms of sets identical in terms of order, sets identical as nonordered sets, equipollent sets and nonequipollent sets, be desirable?

In somewhat summary fashion the experimental results were as follows: The mean learning curves over all trials for all four groups are shown in Figures 4–7. As is evident from these curves the number of errors on the concept of identity of ordered sets was extremely small. From the high proportion of correct responses even in the first block of trials it is evident that this concept is a very natural and simple one for children. Learning curves for trials before the last error are also shown in these figures. To identify the last error prior to conditioning, we adopted a criterion of 16 successive correct responses. For this reason, these curves are only shown for the first 40 trials. The combined curve for Groups 1a and 4a is clearly stationary. This is also the case for 2b, 3a, 3b and 4b.[1] The results of the χ^2 test of stationarity for blocks of 4 trials are shown in Table I and confirm these graphic observations. Only the curve for 1b approaches significance. (No computation was made for 3a because of the small number of errors; the number of subjects in the final block of 4 trials is shown in the right-hand column of the table.)

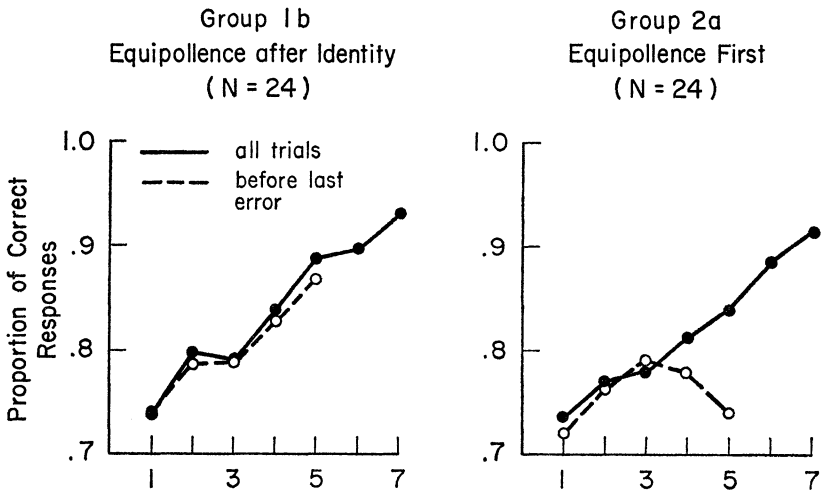


Fig. 5. Proportion of correct responses over all trials and before last error in blocks of eight trials, equipollence of sets, Groups 1b and 2a (Exp. II).

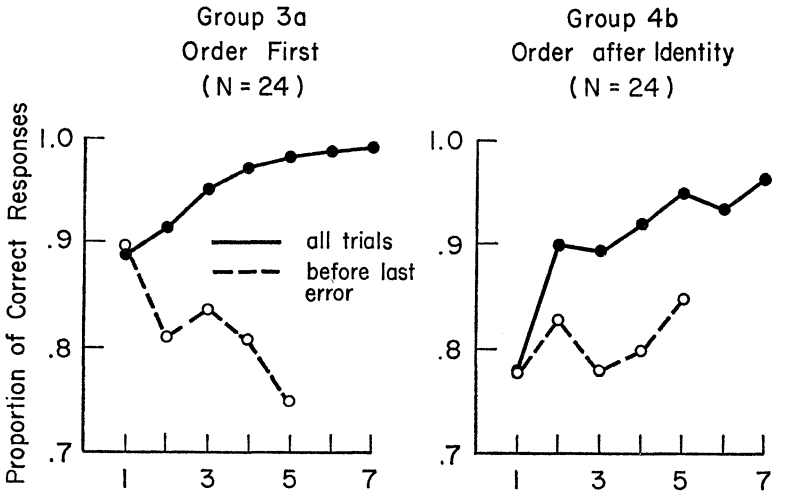


Fig. 6. Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of ordered sets, Groups 3a and 4b (Exp. II).

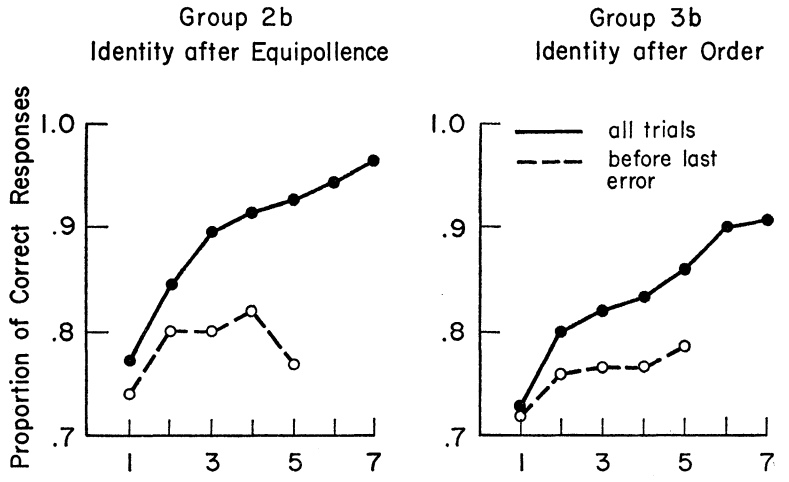


Fig. 7. Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of sets, Groups 2b and 3b (Exp. II).

TABLE I

Stationarity results for equipollence and identity of sets experiment (Exp. II)

Group	χ^2	*df*	$p>$	*Ss* in last block
1a & 4a	4.95	9	0.80	9
1b	16.69	9	0.05	12
2a	4.79	9	0.80	11
3a	– Too few errors –			1
4b	4.89	9	0.80	5
2b	5.96	9	0.70	5
3b	3.49	9	0.90	10

I shall restrict myself to one Vincent curve for this experiment. The 48 subjects of Groups 1 and 4 began with the concept of identity of sets. Of the 48 subjects, 38 met the criterion of 16 successive correct responses mentioned above. The Vincent curve for the criterion subjects is shown in Figure 3. Evidence of nonstationarity in the fourth quartile is present as in the case of Experiment I.

Examination of the mean learning curves over all trials apparently indicates little evidence of transfer. Somewhat surprisingly, the only definite evidence confirms the existence of negative transfer. In particular, it seems clear from Figure 6, there is negative transfer in learning the concept of identity of ordered sets after the concept of identity of unordered sets. Also, from Figures 4 and 7, it seems apparent that there is negative transfer in learning identity of sets after identity of ordered sets, but not after equipollence of sets.

The effects of transfer are actually more evident when we examine the data from the standpoint of two or four concepts. The mean learning curves over all 56 trials for the various concepts are shown in Figures 8–14. The data points are for blocks of 8 trials. The abbreviations used in the legends are nearly self-explanatory. For the learning curves shown at the right of each figure, the O curve is for pairs of sets identical in the sense of ordered sets, the $I\bar{O}$ curve for pairs of sets identical only in the sense of unordered sets, the $E\bar{I}$ curve for pairs of equipollent but not identical sets, and the $\bar{E}$ curve for pairs of nonequipollent sets. These four curves thus represent all pairs of sets in four mutually exclusive and exhaustive classes. The legend is the same for all figures. On the other

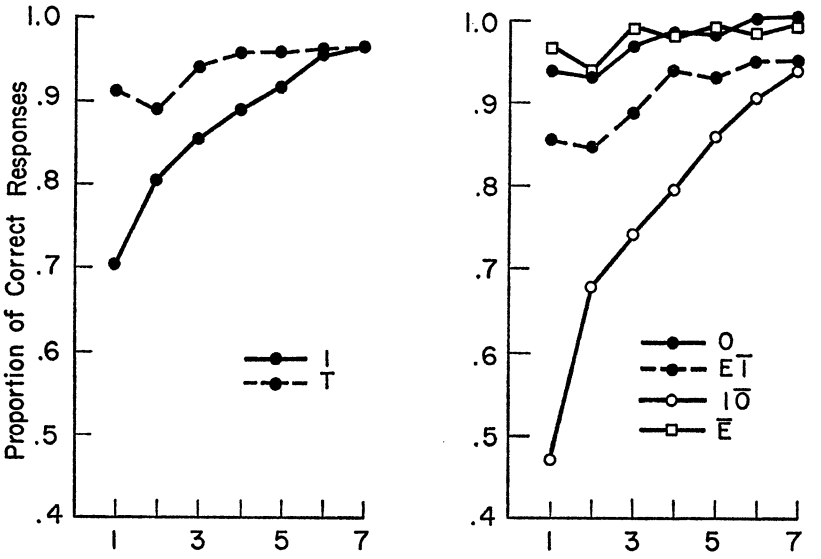


Fig. 8. Proportion of correct responses in blocks of eight trials for two and four concepts, identity of sets ($N = 48$), Groups 1a and 4a (Exp. II).

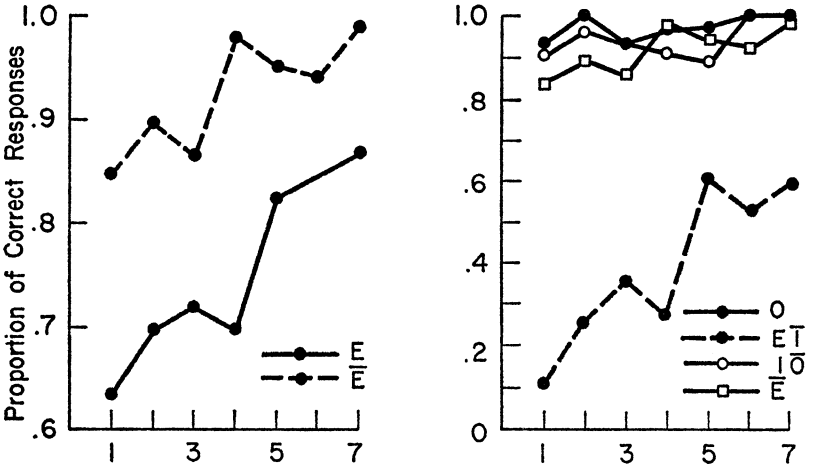


Fig. 9. Proportion of correct responses in blocks of eight trials for two and four concepts, equipollence following identity of sets, Group 1b (Exp. II).

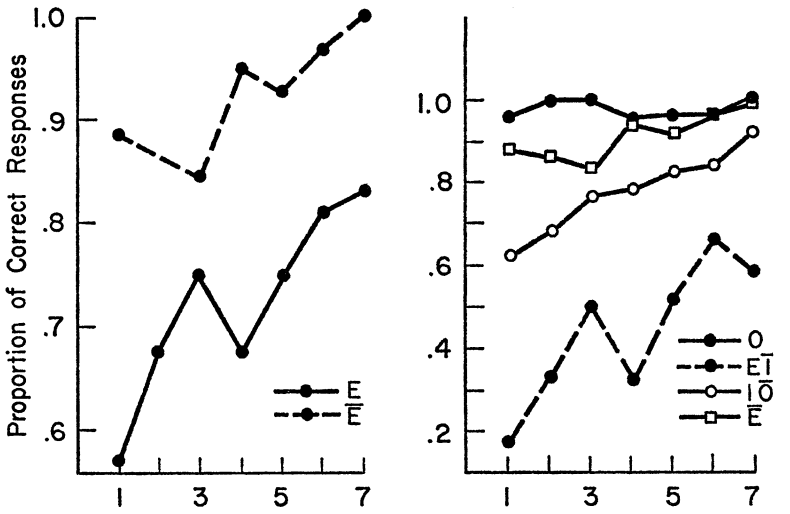


Fig. 10. Proportion of correct responses in blocks of eight trials for two and four concepts, equipollence, Group 2a (Exp. II).

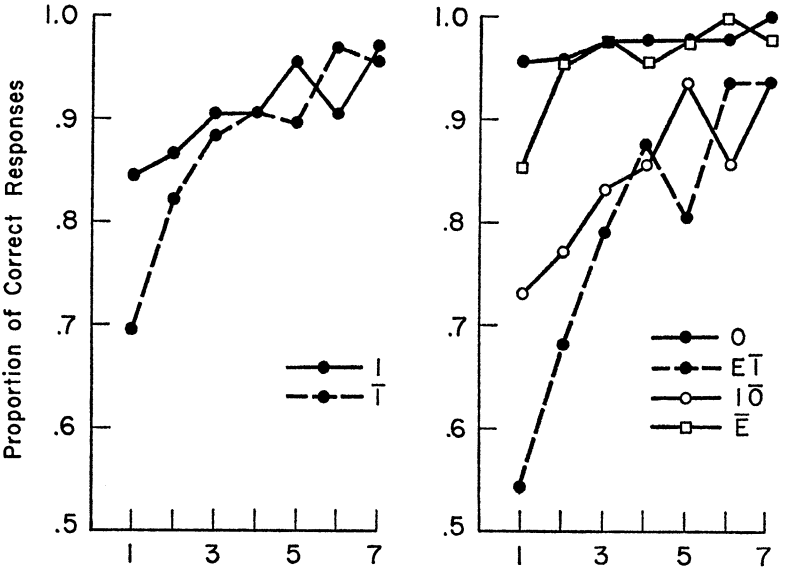


Fig. 11. Proportion of correct responses in blocks of eight trials for two and four concepts, identity following equipollence of sets, Group 2b (Exp. II).

hand, the curves for the two-concept analysis shown at the left of each figure differ in definition according to the problem being learned. In Figure 8 the dichotomy is identical and nonidentical sets (I and $\bar{I}$); in Figure 9 it is equipollent and nonequipollent sets (E and $\bar{E}$), and so forth for the other five figures.

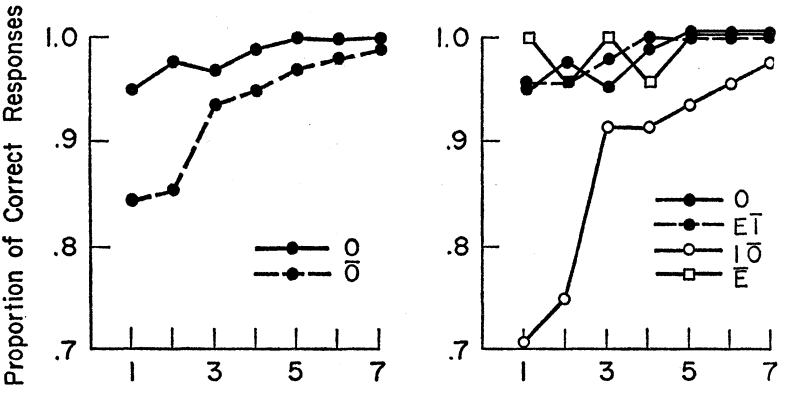


Fig. 12. Proportion of correct responses in blocks of eight trials for two and four concepts, identity of ordered sets, Group 3a (Exp. II).

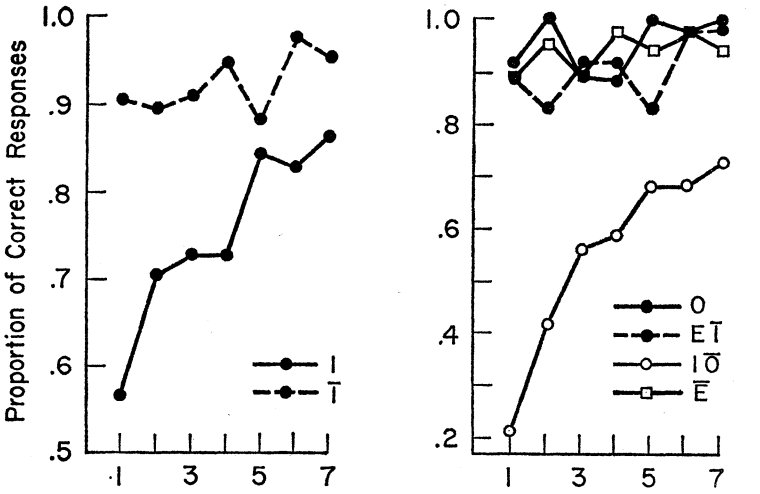


Fig. 13. Proportion of correct responses in blocks of eight trials for two and four concepts, identity of sets, following identity of ordered sets, Group 3b (Exp. II).

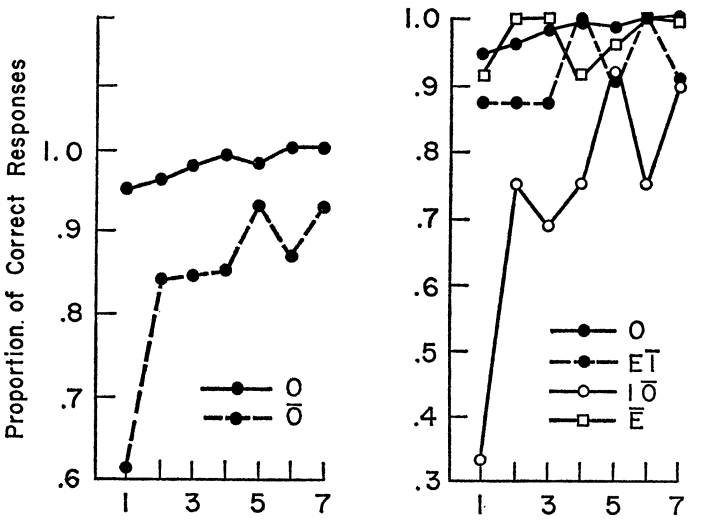


Fig. 14. Proportion of correct responses in blocks of eight trials for two and four concepts, identity of ordered sets following identity of sets, Group 4b (Exp. II).

Before considering questions of transfer, several observations should be made about the individual figures. First, for each of the eight subgroups (1a–4b) the learning curves for the two-concepts and the four-concepts are not homogeneous. A difference in difficulty at either level of analysis can be detected in all cases. Second, contrary to some experimental results in concept formation, the two-concept curves at the left of each figure show that the absence of identity or equipollence is often easier to detect than its presence. The dichotomy of O vs. $\bar{O}$, that is, identity or nonidentity of ordered sets, is the natural one. When the "presence" of a concept disagrees with this natural dichotomy, as it does in the case of identity and equipollence of sets, it is more difficult to detect than the absence of the concept. This conclusion is borne out by Figures 8 and 10 for the groups beginning with identity and equipollence, respectively, as well as for Group 3b (Figure 13), that was trained on ordered sets before identity of sets. This same conclusion even holds fairly well for the second sessions after training on some other concept (Figures 9, 11, 12). Figure 14, which compares O and $\bar{O}$ after training on identity of sets, indicates, I think, the tentative conclusion to be drawn. *Whether the absence or presence of a concept is more difficult to learn*

depends much more on the previous training and experience of a subject than on the concept itself. When we compare Figure 12 with Figure 14 we see that even the difference between O and $\bar{O}$ in Figure 12 is influenced by the prior training or identity, for the difference is greater in Figure 14, and surely this is so because the $I\bar{O}$ cases have to be reversed in going from sets to ordered sets.

Third, examination of the four-concept curves reveals a natural gradient of difficulty. We may apply something rather like Coombs's (1950) unfolding technique to develop an ordinal generalization gradient. The natural or objective order of the classes of pairs of sets is O, $I\bar{O}$, $E\bar{I}$, $\bar{E}$. For any of the three concepts of sameness studied in the experiment, we may, without disturbing this objective ordering, characterize the classes exhibiting presence of the concept and those exhibiting its absence by cutting the ordering into two pieces. On a given *side* of the *cut*, as I shall call it, the nearer a class is to the cut the more difficult it is. Consider, to begin with, Figure 8. The task is identity of sets, and the cut is between $I\bar{O}$ and $E\bar{I}$; we see that, on the one side $I\bar{O}$ is more difficult than O, and on the other side of the cut, $E\bar{I}$ is more difficult than $\bar{E}$. Turning to Figure 9, the task is equipollence and thus the cut is between $E\bar{I}$ and $\bar{E}$; of the three concepts on the $E\bar{I}$ side, $E\bar{I}$ is clearly the most difficult and $I\bar{O}$ is slightly more difficult than O, sustaining the hypothesis of an ordinal gradient. In Figure 10, the task is equipollence again, but in this case without prior training, and the results are as expected but more decisive than those shown in Figure 9. Figure 11, like Figure 8, sustains the hypothesis when the task is identity of sets. In the case of Figure 12, the task is identity of ordered sets and thus $I\bar{O}$, $E\bar{I}$ and $\bar{E}$ occur on the same side of the cut. $I\bar{O}$ is clearly the most difficult, but it is not really possible clearly to distinguish $E\bar{I}$ and $\bar{E}$ in difficulty, for very few errors are made in either class. In Figure 13 the task is identity of sets again, but this time following identity of ordered sets. The proper order of difficulty is maintained but the distinction between $E\bar{I}$ and $\bar{E}$ is not as sharply defined as in Figure 8 or Figure 11. Finally, in Figure 14, the task is identity of ordered sets following identity of sets. The gradients are as predicted by the hypothesis and are better defined than in Figure 12 – no doubt because of the prior training on identity of sets. The existence and detailed nature of these natural gradients of difficulty within a concept task are subjects that seem to be worth considerable further investigation.

I turn now to evidence of transfer in the four-concept analysis. From examination of the over-all, mean learning curves which, in the terminology of the present discussion, are the one-concept curves, we observed no positive transfer but two cases of negative transfer. As might be expected, the four-concept curves yield a richer body of results. I shall try to summarize only what appear to be the most important points. Comparing Figures 8 and 11, we see that for the learning of identity of sets, prior training on equipollence has positive transfer for class $I\bar{O}$ and negative transfer for $E\bar{I}$. The qualitative explanation appears obvious: the initial natural dichotomy seems to be O, $\bar{O}$, and for this dichotomy $I\bar{O}$ is a class of "different" pairs, but the task of equipollence reinforces the treatment of $I\bar{O}$ pairs as the "same"; the situation is reversed for the class $E\bar{I}$, and thus the negative transfer, for under equipollence $E\bar{I}$ pairs are the "same", but under identity of sets they are "different".

Comparing now Figures 8 and 13 in which the task is again identity of sets but the prior training is on identity of ordered sets rather than equipollence, there is, as would be expected by the kind of argument just given, negative transfer for the class $I\bar{O}$. There is also some slight evidence of positive transfer for $E\bar{I}$.

Looking next at Figures 9 and 10, we observe positive transfer for the class $I\bar{O}$ when the task is equipollence and the prior training is on identity of sets. What is surprising is the relatively slight amount of negative transfer for the class $E\bar{I}$.

Finally, we compare Figures 12 and 14, in which the task is identity of ordered sets; in the latter figure this task is preceded by identity of sets and we observe negative transfer for the class $I\bar{O}$, as would be expected. The response curves for the other three classes are too close to probability 1 to make additional inferences, although there is a slight negative transfer for $E\bar{I}$ that cannot be explained by the principles stated above.

It seems apparent from these results that the analysis of transfer in the learning of mathematical concepts may often be facilitated if a fine-scale breakdown of the concepts in question into a number of subconcepts is possible. Needed most is a quantitative theory to guide a more detailed analysis of the transfer phenomena.

Experiment III. Polygons and Angles

This experiment is reported in detail in Stoll (1962), and some of the

data are presented here with her permission. The subjects were 32 kindergarten children divided into two equal groups. For both groups the experiment was a successive discrimination, three-response situation, with one group discriminating between triangles, quadrilaterals, and pentagons, and the other group discriminating between acute, right, and obtuse angles. For all subjects a typical case of each form (that is, one of the three types of polygons or three types of angles) was shown immediately above the appropriate response key. As in the case of Experiment II, no single stimulus display was repeated for any one subject. Stimulus displays representing each form were randomized over experimental trials in blocks of nine, with three of each type appearing in each block. The subjects were run to a criterion of nine successively correct responses, but with not more than 54 trials in any one session.

For the quadrilaterals and pentagons, the guessing probability prior to the last error was essentially the same, $\hat{p}=0.609$ and $\hat{p}=0.600$, respectively. Consequently, the proportions of correct responses for the combined data are presented in blocks of six trials, together with the mean learning curve for all trials, in Figure 15. The corresponding data for the triangles are not

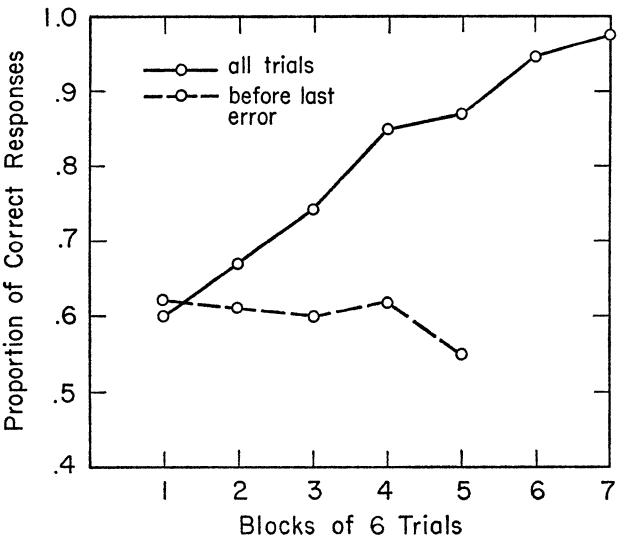


Fig. 15. Proportion of correct responses prior to last error and mean learning curve (quadrilateral and pentagon concepts, Stoll experiment).

presented because the initial proportion of correct responses was quite high and the subjects learned to recognize triangles correctly very easily.

Figure 16 presents the same curves for the combined data for the three types of angles, although the guessing probability varied between the

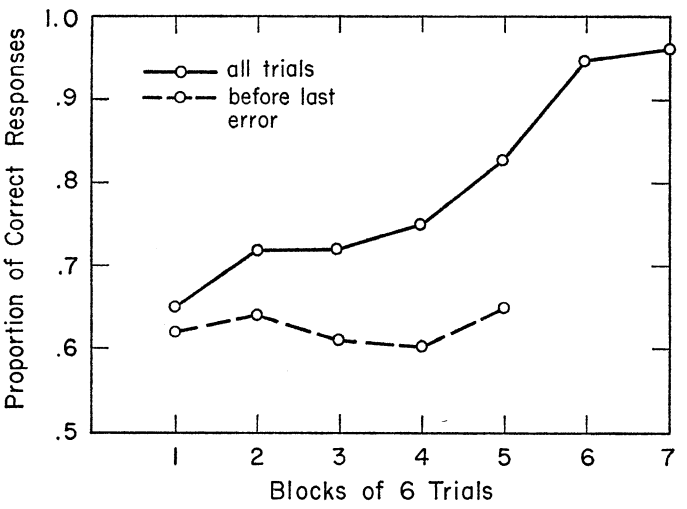


Fig. 16. Proportion of correct responses prior to last error and mean learning curve (acute, right, and obtuse angle concepts, Stoll experiment).

angles. Both figures strongly support the hypothesis of a constant guessing probability prior to conditioning. In the case of the quadrilaterals and pentagons, $\chi^2 = 0.71$, $df = 4$, $P > 0.90$, $N = 548$. In the case of the combined data for the angles, $\chi^2 = 0.97$, $df = 4$, $P > 0.90$, $N = 919$.

The Vincent curves for each concept (except that of the triangle) are shown in Figure 17. The pentagons, quadrilaterals, and right angles have quite stationary Vincent curves, whereas there is a definite increase in the fourth quartile of the Vincent curves for the acute and obtuse angles, and in the case of the obtuse angles there is, in fact, a significant increase in the third quartile. Statistical tests of stationarity of these Vincent curves support the results of visual inspection. Each test has 3 degrees of freedom because the analysis is based on the data for the four quartiles. In the case of the quadrilaterals, $\chi^2 = 1.75$; for the pentagons, $\chi^2 = 1.33$; for the right

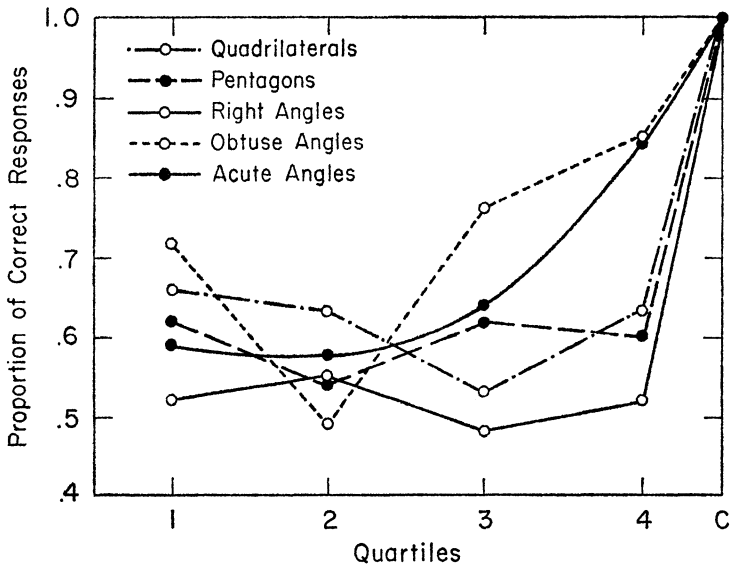


Fig. 17. Vincent learning curves in quartiles for proportion of correct responses prior to last error for Stoll experiment.

angles, $\chi^2 = 0.95$; for the obtuse angles, $\chi^2 = 12.63$; and for the acute angles, $\chi^2 = 16.43$. Only the last two values are significant.

Using responses before the last error, for all concepts except that of triangle, goodness-of-fit tests were performed for (1) stationarity in blocks of six trials, (2) binomial distribution of responses as correct or incorrect in blocks of four trials, and (3) independence of responses, the test made for zero-order vs. first-order dependence. The results of these tests are presented in Table II. The results shown strongly support the adequacy of the one-element model for this experiment.

Experiment IV. Variation in Method of Stimulus Display

In this study conducted with Rose Ginsberg, we compared the rate of learning in two experimental situations, one in which stimulus displays were presented individually in the usual way, and the other in which the same stimulus displays were presented by means of colored slides to groups of four children. The concept to be learned was identity of sets, and in both situations the children were required to respond by pressing

TABLE II

Stationarity, order, and binomial distribution results (Stoll experiment on geometric forms)

	X^2	*df*	*P*>
Quadrilateral, $p = 0.609$:			
Stationarity ($N = 273$)	1.68	4	0.70
Order ($N = 262$)	0.65	1	0.40
Binomial distribution ($N = 65$)	1.77	2	0.40
Pentagon, $p = 0.600$:			
Stationarity ($N = 275$)	2.40	4	0.60
Order ($N = 269$)	1.76	1	0.15
Binomial distribution ($N = 65$)	2.07	2	0.35
Acute angle, $p = 0.674$:			
Stationarity ($N = 338$)	7.96	4	0.05
Order ($N = 348$)	3.17	1	0.05
Binomial distribution ($N = 85$)	2.66	2	0.25
Right angle, $p = 0.506$:			
Stationarity ($N = 313$)	6.34	4	0.10
Order ($N = 326$)	2.41	1	0.10
Binomial distribution ($N = 80$)	10.52	2	0.001*
Obtuse angle, $p = 0.721$:			
Stationarity ($N = 268$)	1.10	4	0.85
Order ($N = 256$)	7.32	1	0.001*
Binomial distribution ($N = 63$)	2.90	2	0.20
Quadrilateral and pentagon, $p = 0.604$:			
Stationarity ($N = 548$)	0.71	4	0.90
Binomial distribution ($N = 130$)	1.77	2	0.40
All angles, $p = 0.624$:			
Stationarity ($N = 919$)	0.97	4	0.90

one of two buttons, depending upon whether the stimulus display on that trial was identical or nonidentical. Of the 64 subjects 32 were from first grade and 32 from kindergarten classes. For the children receiving individual displays the experimental situation was essentially identical with that of Experiment II.

Each group, however, was divided into two subgroups. One subgroup received the stimulus material in random order, and the other in an order based on anticipated difficulty; in particular, presentations of one-element sets came first, then two-element sets, and finally three-element sets.

The mean learning curves for the two subgroups with random presentation are shown in Figure 18. The results suggest that presentation by

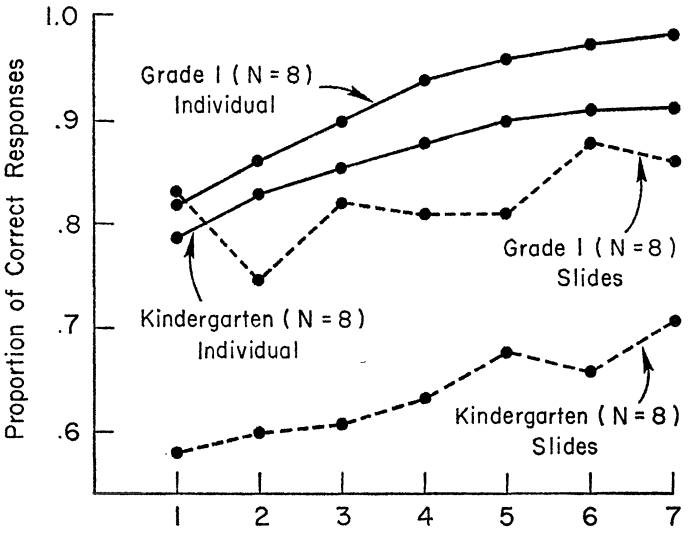


Fig. 18. Proportion of correct responses in blocks of 12 trials, subgroups with random presentation (Exp. IV).

slides is a less effective learning device for younger children, and the younger the child, the more this finding seems to apply. At all levels of difficulty, the kindergarten children learned more efficiently when the stimuli were presented to them in individual sessions. With one- or two-element sets displayed, grade-1 subjects learned only slightly better in the individual session situation than in the slide situation, but when the task was more difficult (stimulus displays of three-element sets) the individual learning situation was clearly the most adequate. In interpreting these results it should be emphasized that the individual session was strictly experimental so that the amount of interaction between subject and experimenter was paralleled in both individual and slide situations.

Why these two experimental situations should produce different results in terms of learning efficiency is not yet clear to us. One possibility is the following: It has been shown, both with lower organisms (Murphy and Miller, 1955) and young children (Murphy and Miller, 1959), that the ideal situation for learning is the contiguity of stimulus, response and reinforcement. In the individual sessions these requirements were met, for the response buttons were 1.5 inches below the stimulus displays and the reinforcement lights were 1.0 inches from the stimuli. On the other hand, in the slide presentations, although the stimulus displays and reinforcements were immediately adjacent to each other, the response buttons were about 3 feet from the screen on which the stimulus display was projected. Experimentally, it has been shown (Murphy and Miller, 1959) that with children of this age group a separation of 6.0 inches is sufficient to interfere with efficient learning.

Experiment V. Incidental Learning

This experiment represents a joint study with Rose Ginsberg. Thirty-six kindergarten children, in 3 groups of 12 each, were run for 60 trials a day on 2 successive days of individual experimental sessions during which they were required to learn equipollence of sets. On the first day, the stimulus displays presented to the subjects on each trial differed in color among the three groups but otherwise were the same. In Group 1, all displays were in one color – black – and in Group 2, equipollent sets were red and nonequipollent sets, yellow. For the first 12 trials in Group 3, equipollent sets were red and nonequipollent sets, yellow; for the remaining 48 trials on that day the two colors were gradually fused until discrimination between them was not possible. On the second day, all sets were presented to all three groups in one color – black.

As is apparent from the brief description of the experimental design, Group 1 simply had two days' practice under the same conditions with the concept of equipollence. In Group 2, the child did not actually need to learn the concept of equipollence but could simply respond to the color difference on the first day. It is well known that such a color discrimination for young children is a simple task. If the child in this group learned anything about equipollence of sets the first day, therefore, we may assume it to have been a function of incidental learning. If incidental learning is effective, his performance on the second day, when the color cue is dropped,

should have been at least better than the performance of children in Group 1 on the first day. In Group 3, where we gave the child the discriminative cue of color difference in the first trial and then very slowly withdrew that cue, the child should have continued to search the stimulus displays very closely for a color stimulus and thus have been obliged to pay close attention to the stimuli.

The mean learning curves for the three groups are shown in Figure 19.

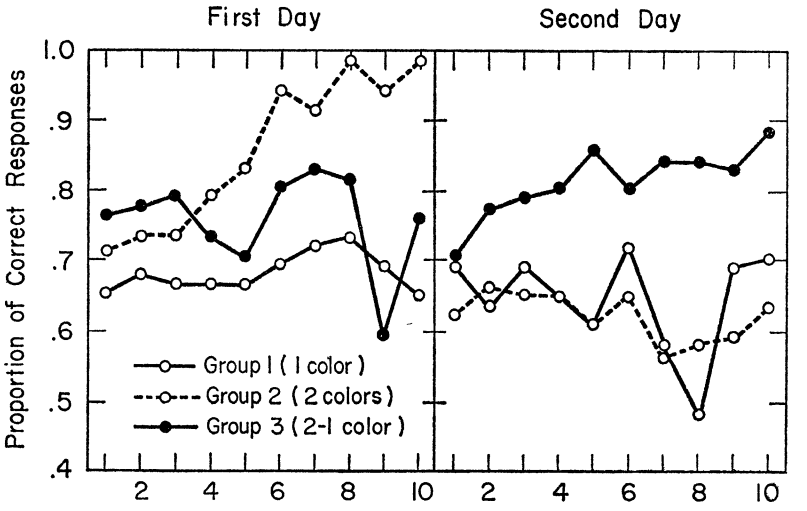


Fig. 19. Proportion of correct responses in blocks of six trials for both days (Exp. V).

Of the three groups only Group 2 approached perfect learning on the first day. In this group, of course, only color discrimination was necessary. Both the other groups did not improve over the first 60 trials, although Group 3 showed some initial improvement when the color cues remained discriminable. On the second day, Group 1 showed no improvement, and the learning curves for this group and Group 2 were practically identical. For Group 3, on the other hand, the results were conspicuously better on the second day than for those of any other group. It is apparent from these curves that the task chosen was relatively difficult for the age of the children, because essentially no improvement was shown by Group 1 over the entire 120 trials. The conditions in Group 3, where the children were

forced to pay very close attention to the stimuli, do seem to have significantly enhanced the learning.

Experiment VI. Variation of Response Methods

This study was made jointly with Rose Ginsberg. Its object was to study the behavioral effects of different methods of response. Specifically, 3 groups, each composed of 20 kindergarten children, were taken individually through a sequence of 60 trials on each of 2 successive days for a total of 120 trials. The task for all 3 groups was equipollence of sets.

In Group 1, the child was presented with pictures of two sets of objects and was to indicate, by pressing one of two buttons, whether the sets "went together" or did not "go together" (were equipollent or non-equipollent).

In Group 2, the child was presented with one display set and two "answer" sets and was required to choose the answer that "went together" with the display set.

In Group 3, the child was presented with one display set and three "answer" sets and was to make his choice from the three possible answers.

This situation has fairly direct reference to teaching methodology in the sense that Group 2 and Group 3 represent multiple-choice possibilities. In Group 1, where the child is required to identify either the presence of the concept or its absence on each trial, the situation is comparable to one in which the child must indicate whether an equation or statement is correct or incorrect.

On the first day, each group of children learned the task described above. On the second day, they were run on an alternative method. Specifically, Group 1 was run under Group 3 conditions and Groups 2 and 3 were run under Group 1 conditions.

The mean learning curves for all groups on both days are shown in Figure 20. It will be noticed that in Group 2, where the subjects were required to choose from one of two available responses, they learned slightly more quickly and to a slightly better level of achievement on the first day than the other groups but, on the second day, when the experimental conditions were shifted, Group 2 subjects did less well than the subjects in the other two groups. The clear superiority of Group 1 on the second day, when they were transferred to Group 3 conditions, indicates some positive transfer from learning to judge whether or not a

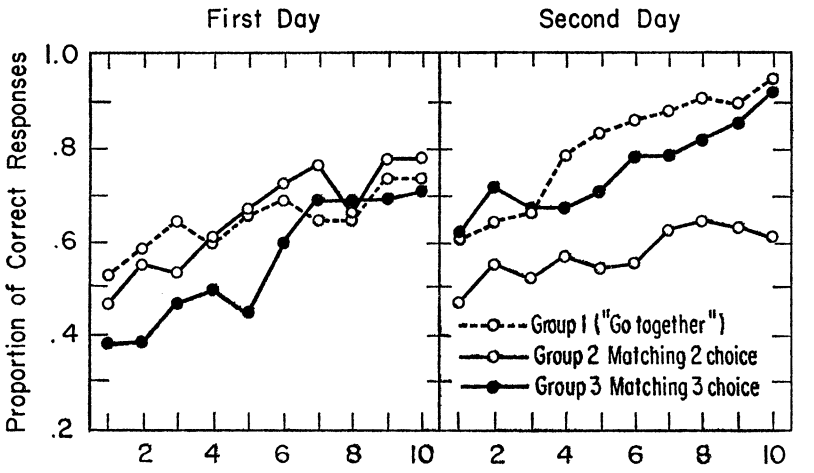


Fig. 20. Proportion of correct responses for two successive days in blocks of six trials for all subjects (Exp. VI).

concept is present to the multiple-choice situation, whereas the results for Groups 2 and 3 on the second day indicate some negative transfer from the multiple-response methods to the presence-or-absence method.

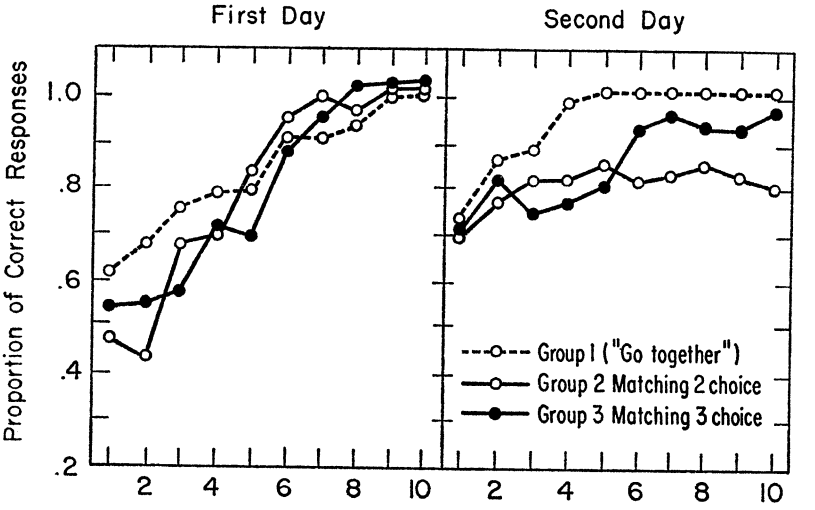


Fig. 21. Proportion of correct responses for two successive days in blocks of six trials for subjects achieving of 12 successive correct responses (Exp. VI).

These results are further supported when we examine separately the data from subjects achieving a criterion of 12 successive correct responses on the first day. The more successful method was clearly that used in Group I, as indicated by the curves in Figure 21. The subjects in this group were conspicuously more successful than the other groups on the second day, making, in fact, no errors from Trial 30 to Trial 60. Group 3 achieved perfect scores on the second day only on the last six trials, and Group 2 never reached that level on the second day, although, like the other criterion subjects, they had achieved perfect learning on the first day.

It seems reasonable to conclude tentatively that the method used with Group 1, where subjects were required to recognize the presence or absence of some property on each trial, is the more successful method in establishing the understanding of a concept well enough to permit transfer to a different response method.

Support for the all-or-none model of conditioning is also to be found in this experiment. In Table III, χ^2 goodness-of-fit tests of stationarity over

TABLE III

Test for stationarity over trials before the final error (Exp. VI)

	Group 1		Group 2		Group 3	
	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
χ^2	4.97	2.41	10.76	4.255	16.07	2.87
df	8	1	9	8	9	7
$P>$	0.70	0.10	0.20	0.80	0.05	0.80

trials before the final error for each group on each day are shown. The six values are all nonsignificant and thus support the basic assumption of the all-or-none models.

Some Tentative Conclusions

On the basis of the six experiments just discussed I would like to draw some tentative conclusions, some of which are important for pedagogical procedures (cf. Suppes and Ginsberg, 1962b). I want to emphasize, however, that I do not wish to claim that the evidence from these experiments

is conclusive enough to establish any one of the six conclusions in any final way, but what I do hope is that the attempt to summarize some of the implications of these experiments will stimulate other research workers to investigate these and related propositions in more adequate detail.

(1) Formation of simple mathematical concepts in young children is approximately an all-or-none process. Evidence indicates, however, that significant deviations from the all-or-none model are present (see the discussion of the two-element model below).

(2) Learning is more efficient if the child who makes an error is required to make an overt correction response in the presence of the stimulus to be learned (Exp. I).

(3) Incidental learning does not appear to be an effective method of acquisition for young children. In Experiment IV the group of children that responded to a color discrimination did not subsequently give any indication of having learned the underlying concepts.

(4) Contiguity of response, stimulus, and reinforcement enhances learning (Exp. V).

(5) In the learning of related mathematical concepts the amount of over-all transfer from the learning of one concept to another is surprisingly small. However, considerable positive or negative transfer between specific subconcepts is often present (Exp. II).

(6) Transfer of a concept is more effective if, in the learning situation, the subject is required to recognize the presence or absence of a concept in a number of stimulus displays, than if learning has involved matching from a number of possible responses (Exp. VI).

Several of these conclusions are at variance with generally accepted results for adult learning behavior. For example, the efficacy of an immediate overt correction response (see Burke *et al.*, 1954, for negative results on this method in adults), the variation of response method, or the relative specificity of the learning of concepts with relatively little transfer. What is much needed is a wider range of systematic studies to isolate the factors of learning in young children which are particularly distinct from common features of adult learning behavior.

Two-Element Model

In the first conclusion mentioned above, we stated that the formation of

concepts is approximately an all-or-none process in young children. On the other hand, the detailed analysis of responses prior to the last error indicates that, in many cases, there is an incremental effect appearing in the last quartile or even, sometimes, in the last two quartiles of the data. This matter is discussed in some detail in Suppes and Ginsberg (1963). I would simply like briefly to mention here what currently appears to be the best extension of the one-element model to account for these results.

The simplest alternative model is the linear incremental model with a single operator. The intuitive idea of this model is precisely the opposite of the all-or-none conditioning model. The supposition is that learning proceeds on an incremental basis. Let q_n be the probability of an error on trial n. Then the model is formulated by the following recursive equation:

$$(1) \qquad q_{n+1} = (1 - \theta)\, q_n,$$

where $O < \theta \leqslant 1$. It is simple to show but somewhat surprising that this purely incremental model has precisely the same mean learning curve as the all-or-none model if we set $c = \theta$. (To obtain this identity of the learning curves we must, of course, consider all responses and not simply responses prior to the last error.) The incremental model differs sharply from the all-or-none model in the kind of learning curve predicted for responses prior to the last error, as is evident from Equation (1). It may be shown, moreover, that the concave upward Vincent curves obtained in several of the experiments discussed above (see Figures 3 and 17) cannot be accounted for by the linear incremental models.

The second simple alternative that will account for these concave-upward Vincent curves is a model that represents a kind of compromise between the all-or-none model and the incremental model. It results from a simple extension of the one-element model, that is, the assumption that associated with each situation are two stimulus elements and, therefore, learning proceeds in two stages of all-or-none conditioning. Each of the two elements is conditioned on an all-or-none basis, but the two parameters of conditioning, one for each element, may be adjusted to produce various incremental effects on the response probabilities. Let σ and τ be the two elements. The basic learning process may be represented by the following four-state Markov process in which the four states (σ, τ), σ, τ, and O represent the possible states of conditioning of the two stimulus elements. Because we do not attempt experimentally to identify the stimuli σ and τ,

	(σ, τ)	σ	τ	O
(σ, τ)	1	0	0	0
σ	$b'/2$	$1-b'/2$	0	0
τ	$b'/2$	0	$1-b'/2$	0
O	0	$a/2$	$a/2$	$1-a$

this Markov process may be collapsed into a three-state process, in which the states are simply the *number* of stimuli conditioned to the correct response. In the matrix shown above a is the probability of conditioning at the first stage and b' is the probability of conditioning at the second stage. The division by $\frac{1}{2}$ in the matrix simply represents the equal probability of sampling one of the two elements. If we consider only the number of stimuli, it is convenient to replace $b'/2$ by b and we obtain the transition matrix shown below:

	2	1	O
2	1	0	0
1	b	$1-b$	0
O	0	a	$1-a$

To complete the description of the process we associate with the sampling of each element σ and τ a guessing probability g_σ and g_τ when the elements are still unconditioned. For the states 0 and 1 of the second matrix shown we then have the guessing probabilities g_0 and g_1 defined in the obvious manner in terms of the sampling probabilities:

$$g_0 = \tfrac{1}{2}g_\sigma + \tfrac{1}{2}g_\tau,$$
$$g_1 = \tfrac{1}{4}g_\sigma + \tfrac{1}{4}g_\tau + \tfrac{1}{2} = \tfrac{1}{2}g_0 + \tfrac{1}{2}.$$

The probabilities g_σ and g_τ are not observable but g_0 is, and g_1 is a simple function of it. This means that we now have a process with three free parameters, the conditioning parameters a and b and the guessing probability g_0. I shall not attempt to report on the detailed application of this two-element model, but we are now in the process of applying it to a number of different experimental situations and hope to report in detail on its empirical validity in the near future.[2†] I would, however, like to remark that a very interesting interpretation of this kind of two-stage

model has recently been given by Restle (1964), who interprets the two stages of learning as conditioning and discrimination. The model he proposes differs in detail from that given here, but for most observable response patterns the differences between the two will not be large.

Before turning to another topic, I would like to emphasize that I do not feel that the analysis of concept formation in terms of the simple one- and two-element models sketched here is fully satisfactory intellectually. It is apparent that these models must be regarded as schemata of the full process that is taking place in concept formation. What is surprising is that they are able to account for response data as well as they do. Theories that postulate more details about the learning process in concept formation are needed to go beyond the present analysis. This, I take it, will be particularly true as we proceed to the analysis of more complicated mathematical concepts, whose learning must rest upon the understanding of simpler concepts.

IV. LOGIC AND MATHEMATICAL PROOFS

Together with several younger associates I have conducted, for several years, pedagogical and psychological experiments on the learning of mathematical logic with elementary-school children. Before turning to a relatively systematic statement of some of our results, I would like to survey briefly what we have attempted.

In the fall of 1956 I brought into my college logic course a selected group of sixth, seventh, and eighth graders (they were, in fact, no more selected than the Stanford students in the course). Their demonstrated ability to master the course and perform at a level only slightly below that of the college students was the initial impetus for further work. The next important step was the extensive study by Shirley Hill of the reasoning abilities of first, second, and third graders. This study was begun in 1959 and completed as her dissertation in 1961. I shall report briefly on this below. In 1960 Dr. Hill and I wrote a text and taught a pilot group of fifth graders a year's course in mathematical logic. The course was structured very similarly to a college logic course except that material was presented more explicitly and at a much slower pace. Students were selected on the basis of ability and interest (the minimum *I.Q.* was 110), and again the positive results were an impetus to further work. Because of the success of this class, the textbook was revised (Suppes and Hill, 1964) and, during

the academic year 1962–63, was taught to approximately 300 selected fifth graders in the Bay Area, with support for the project coming from the Office of Education and the National Science Foundation. These same classes were given a second year of instruction as sixth graders and, in another year, we shall be able to report in detail on their level of achievement. We were also interested in seeing if we could train fifth-grade teachers to teach the course as part of their regular curriculum. To this end, we gave them a special course in logic in the summer of 1961 and all the classes but one were taught by the teachers.

We began experimental psychological studies of how and to what degree children of still younger ages could learn the concepts of formal inference. I shall report briefly on a pilot study with first graders. On the basis of the experience of several of us with the teaching of logic to elementary-school children, we conducted an extensive psychological experiment with fourth-grade children to determine whether it was easier initially to learn rules of sentential inference when the standard interpretations were given, or whether it was easier simply to learn the rules as part of an uninterpreted meaningless game. This last possibility was, of course, most disturbing for a wide variety of mathematicians interested in the teaching of mathematics. I shall not enter here into the many reasons why I think there are good psychological arguments to believe that the initial teaching of inference simply as a game will turn out to be the most effective approach. I am frankly reluctant to formulate any very definite ideas about this highly controversial matter until we have accumulated a much more substantial body of evidence.

I turn now to the two experiments mentioned above on which I want to report briefly.

Experiment VII. Logical Abilities of Young Children

As already remarked, this extensive empirical study constituted Shirley Hill's doctoral dissertation (1961). Dr. Hill gave a test instrument consisting of 100 items to 270 children in the age group 6–8 years (first, second, and third grades). Each of the 100 items consisted of 2 or 3 verbal premises plus a conclusion presented orally as a question. The subject was asked to affirm or deny the conclusion as presented. There were two primary reasons for not asking the children to compose a conclusion: In the first place, children of this age sometimes have difficulty formulating

sentences; this has sometimes been cited as the reason for inappropriate measures of their reasoning abilities. The second reason is, simply, the methodological difficulty of interpreting the correctness or incorrectness of a conclusion given as a free response. The 100 items were equally divided between positive and negative answers. The first part of the test consisted of 60 items that were drawn from sentential logic. Every conclusion or its negation followed from the given premises by the sentential theory of inference. The second part consisted of 40 items that were drawn from predicate logic, including 13 classical syllogisms. The predicate logic items, however, also included inferences using two-place predicates together with existential quantifiers.

Because it is easy for children to give the correct answer to a problem in which the conclusion is generally true or false, every attempt was made to construct the items in such a way that the omission of one premise would make it impossible to draw the correct conclusion. To provide a behavioral check on this aspect of the items a base-line group of 50 subjects was given the test with the first premise of each item omitted. For instance, to quote the illustration given by Dr. Hill (1961, p. 43), the original item might read:

> If that boy is John's brother, then he is ten years old.
> That boy is not ten years old.
> Is he John's brother?

For the base-line group the item would be presented:

> If that boy is not ten years old, is he John's brother?

An example of a badly constructed item would be the following:

> If boys are stronger than girls, then boys can run faster than girls.
> Boys are stronger than girls.
> Can boys run faster than girls?

Naturally almost all children gave the correct answer to this latter item, but their behavioral response actually told us little about their intuitive grasp of principles of logical inference. That Dr. Hill's items were well constructed are attested to by the fact that the base-line group averaged 52.02% correct items, which does not significantly differ from chance. (Note that this percentage is based on 5000 subject items.)

I shall not go into all the facets of Dr. Hill's study here. I mainly want

to report on one or two of the most important conclusions. Let me first mention the results of the three standard groups of ages 6, 7, and 8 years. The 6-year-old group receiving the items described above got 71.18% of the items correct. The 7-year-old group got 79.54% of the items correct, and the 8-year-old group got 85.58% correct. These percentage figures indicate a steady increase with age in the ability to draw correct logical inferences from hypothetical premises. In addition to the fact of increase, it is just as important to note that the 6-year-old children performed at quite a high level, in contradiction to the view of Piaget and his followers that such young children are limited to concrete operations. Dr. Hill's study certainly provides substantial evidence to the contrary.

To avoid any possible confusion, it should be borne in mind that no claim is made that this study shows young children to be able explicitly to state formal principles of inference. What is claimed is that their grasp of the structure of ordinary language is sufficiently deep for them to be able to make *use* of standard principles of inference with considerable accuracy.

I would like to present just two other results of Dr. Hill's study. To avoid the conjecture that children aged six may be able to do the simpler forms of inference quite well, but will do badly on the more difficult inferences involving two-place predicates, the percentage of correct responses for each age group on the 10 types of inferences appearing in the 100-item test are shown in Table IV. The last two categories entitled

TABLE IV

Percentage of correct responses for different principles of inference by age level

Principles of Inference	Percentage of correct responses		
	Age 6	Age 7	Age 8
Modus ponendo ponens	78	89	92
Modus tollendo ponens	82	84	90
Modus tollendo tollens	74	79	84
Law of hypothetical syllogism	78	86	88
Hypothetical syllogism and *tollendo tollens*	76	79	85
Tollendo tollens and *tollendo ponens*	65	77	81
Ponendo ponens and *tollendo tollens*	65	67	76
Classical syllogism	66	75	86
Quantificational logic – universal quantifiers	69	81	84
Quantificational logic – existential quantifiers	64	79	88

'Quantificational Logic – Universal Quantifiers' and 'Quantificational Logic – Existential Quantifiers' refer to inferences that do not fall within the scheme of the classical syllogism. Although these last two categories are more difficult than the simplest *modus ponendo ponens* applications, the performance level of the children aged six is still well above chance, and it is interesting to note that the performance on universal quantifiers is actually slightly better than the performance on sentential inferences using both *ponendo ponens* and *tollendo ponens.*

The second result concerns the attempt to identify some of the more obvious sources of difficulty. The lack of a sharply defined gradient in Table IV suggested further examination of individual items. What turned out to be a major source of difficulty was the inclusion of an additional negation in an inference. Two hypothetical items that illustrate this difference are the following: Consider first as a case of *modus ponendo ponens:*

If this is Room 7, then it is a first-grade room.
This is Room 7.
Is it a first-grade room?

Let us now modify this example, still making it an application of *modus ponendo ponens:*

If this is not Room 8, then it is not a first-grade room.
This is not Room 8.
Is it a first-grade room?

The additional negations in the second item are a source of considerable difficulty to the children. It might be thought that the negations simply

TABLE V

Comparison of increase in error associated with the addition of negation and with compound implications

	Percentage of error out of total possible responses		
Principles of inference	Regular form	Additional negation	Combined implication
Modus ponendo ponens	0.06	0.19	0.17
Modus tollendo tollens	0.12	0.34	
Modus tollendo ponens	0.03	0.25	0.27
Modus tollendo tollens	0.12	0.34	
Law of hypothetical syllogism	0.08	0.22	0.16
Modus tollendo tollens	0.12	0.34	

cause difficulty because they represent an increase in general complexity. To examine this question Dr. Hill compared the cases using a single rule of inference in which negations occurred, with the use of combined implications involving more than one rule of inference. The results are shown in Table V. It is clear from this table that an additional negation adds a greater factor of difficulty than the use of more than one principle of inference.

I have only presented here a few of the results of this important study. A complete statement of the results are included in Hill (1961).

Experiment VIII. Pilot Study of Mathematical Proofs

The details of this pilot study are in Suppes (1965a).[3†] The original study was conducted with the assistance of John M. Vickers, and we are now engaged in a larger study along the same lines. The primary objective of this pilot study was to determine if it is feasible to apply the one-element model, described earlier, to the behavior of young children by constructing proofs in the trivial mathematical system, described as follows: Any finite string of 1's is a well-formed formula of the system. The single axiom is the single symbol 1. The four rules of inference are:

R1. $S \rightarrow S11$
R2. $S \rightarrow S00$
R3. $S1 \rightarrow S$
R4. $S0 \rightarrow S$

where S is a nonempty string. A theorem of the system is, of course, either the axiom or a finite string that may be obtained from the axiom by a finite number of applications of the rules of inference. A general characterization of all theorems is immediate: any finite string is a theorem if and only if it begins with 1. A typical theorem in the system is the following one, which I have chosen because it uses all four rules of inference:

Theorem	101
(1) 1	Axiom
(2) 100	R2
(3) 10	R4
(4) 1011	R1
(5) 101	R3.

The proofs of minimal length in this system are easily found, and the correction procedure was always in terms of a proof of minimal length.

The stimulus discrimination facing the subject on each trial is simply described. He must compare the last line of proof in front of him with the theorem to be proved. This comparison immediately leads to a classification of each last line of proof into one of four categories: additional 1's need to be added to master the theorem (R1); additional 0's need to be added to master the theorem (R2); a 1 must be deleted to continue to master the theorem (R3); or a 0 must be deleted in order to master the theorem (R4). The rule in terms of whether the response should be made is shown in parentheses. When the subject is completely conditioned to all four stimulus discriminations, he will make a correct response corresponding to the application of a rule that will produce a part of a proof of minimal length. For each of the four discriminations with respect to which he is not yet conditioned, there is a guessing probability p_i, $i=1, 2, 3,$ or 4, that he will guess the correct rule and thus the probability $1-p_i$ that he will guess incorrectly. In the analysis of data it was assumed that four independent one-element models were applied, one for each stimulus discrimination. (It is a minor but not serious complication to take account of two possible responses, both correct, i.e., leading to a minimal proof; e.g., in the proof of 1111 we may apply R1 twice and then R3, or R1, R3, and then R1 again.)

The pilot study was conducted with a group of first-grade children from an elementary school near Stanford University. There were 18 subjects in all, divided into 2 groups of 9 each. One group received the procedure just described, including a correction procedure in terms of which a correct response was always shown at the end of the trial. The other group used a discovery method of sorts and was not given a correction procedure on each trial but, at the end of each proof, the subjects were shown a minimal proof or, in the event the subject constructed a minimal proof, told that the proof constructed was correct.

The following criterion rule was used: A subject, according to the criterion, had learned how to give minimal proofs in the system when 4 correct theorems were proved in succession, provided the subject had proved at least 10 theorems. All subjects were given a maximum of 17 theorems to prove, and all subjects, except for 2 in the discovery group, satisfied this criterion by the time the seventeenth theorem was reached.

The 17 theorems were selected according to some relatively definite criteria of structural simplicity from the set of theorems of which the length was greater than 1 and less than 7.

In Table VI, the mean proportion of errors prior to the last error, in

TABLE VI

Observed proportion of errors prior to last error for the correction, discovery, and combined groups (blocks of 12 trials)

	Block				
Group	1	2	3	4	5
Correction	0.28	0.23	0.15	0.00	0.10
Discovery	0.23	0.20	0.40	0.30	0.33
Combined	0.25	0.21	0.30	0.18	0.24

blocks of 12 trials for each group and for the 2 groups combined, are summarized. A trial in this instance is defined as a step, or line, in the proof.

More than 60 trials were necessary in order to prove the 17 theorems, but because very few subjects needed the entire 17 theorems to reach criterion, the mean learning curves were terminated at Trial 60. From this table, it seems that the correction group did better than the discovery group, but I do not think the number of subjects or the total number of trials was adequate to draw any serious conclusions about comparison of the two methods. It is interesting to note that the discovery group had a much more stationary mean learning curve than did the correction group, and in that sense satisfied the one-element model. Of course, these curves are obtained by summing over errors on all four rules. It is very possible that with a larger set of data, for which it would be feasible to separate out the individual rules as the application of the one-element model described above would require, the correction group also would have stationary mean learning curves for data prior to the last error on the basis of the individual rules.

NOTES

[1] 'Group 1a' refers to the performance of Group 1 subjects on the first of their two tasks, 1b to performance on the second task, and similarly for 2a, 2b, 3a, 3b, 4a, and 4b.

[2†] For extensive applications, see Crothers and Suppes (1967).

[3†] Article 20 in this volume.

20. TOWARDS A BEHAVIORAL FOUNDATION OF MATHEMATICAL PROOFS*

I. INTRODUCTION

The logical theory of mathematical proofs has been developed intensively and with great success in this century. I do not need to review for a colloquium audience in Warsaw the main outlines of this development. What is surprising is that so little has been written about the *psychological* theory of mathematical proofs. However interesting they may be as literary documents, I am not willing to count as scientific psychology mathematicians' testimonials of how they made discoveries. It is not that this material is not interesting. It is just that it holds very little promise scientifically. The psychological phenomena which lie at the base of any genuinely new mathematical discovery are surely as complicated and intellectually involved as any in the whole range of human behavior. Introspective accounts of these phenomena are as difficult to work with as basic data, as are the descriptions of nature lovers of sunsets and storms in developing the science of meteorology.

Perhaps to the disappointment of some of you I shall approach the problem of providing a psychological analysis of mathematical proofs by considering examples of an almost ridiculous simplicity. The analysis shall proceed on the assumption that it is possible for sufficiently simple contexts to analyze written and spoken speech acts within the general framework of behavioral psychology (what I shall mean by 'behavioral psychology' will become clearer in the sequel). As an act of faith I would also express the conviction that a still further reduction of behavioral psychology to neurophysiology will ultimately be possible, but I am not hopeful that this reduction will occur in the near future, particularly with reference to complicated intellectual processes.

As W. K. Estes wisely pointed out in the original discussion of this paper

* Reprinted from *The Foundations of Statements and Decisions: Proceedings of the International Colloquium on Methodology of Science, September 18–23, 1961* (ed. by K. Ajdukiewicz), PWN-Polish Scientific Publishers, Warszawa, 1965, pp. 327–341.

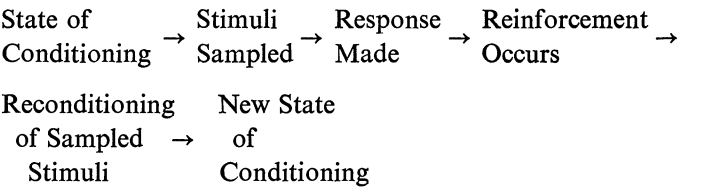
in Warsaw, I also do not believe that a detailed explanation in behavioral terms of the genuine discoveries of mathematicians can be given in the framework I am describing here. The aims of the psychological theory I shall set forth are schematic in the same way that physical theories are schematic. There are only a few phenomena caught in the raw, so to speak, in nature that are subject to any exact explanation and prediction of behavior in terms of existing physical theories, and we all have a good rough estimate of the relative power of physical and psychological theories. On the other hand, I do believe that the kind of theory and analysis I shall be giving of the simplest elements of the learning of mathematical proofs do provide the right sort of framework for the analysis of the most complicated mathematical activity. A phrase I used in this last sentence also provides an important indication of what I wish to mean by a behavioral foundation of mathematical proofs. I do not, it should be clear, mean a behavioral foundation for the written or inscribed proofs themselves, but rather, for the act of learning to give such proofs on the part of students of mathematics. In this sense, a behavioral foundation emphasizes the learning or discovery of proofs.

I will return to these general comments after giving a brief sketch of the relevant psychological theory and an indication of how it may be applied to a simple mathematical system.

II. BRIEF SKETCH OF STIMULUS-SAMPLING THEORY

Stimulus-sampling learning theory was first given a quantitative formulation in 1950 by W. K. Estes, but its basic concepts were developed by a number of psychologists running back to the beginning of the century; particularly important have been the general contributions of such figures as Pavlov, Watson, Hull, and Guthrie. The great merit of Estes is to have shown how these ideas may be cast in a quantitative formulation subject to genuine mathematical analysis. In a highly simplified form the basic ideas run as follows. The organism is presented with a sequence of trials, on each of which he makes a response that is one of several possible choices. In any particular experiment it is assumed that there is a set of stimuli from which the organism draws a sample at the beginning of each trial. It is also assumed that on each trial each stimulus is conditioned to at most one response. The probability of making a given response on any

trial is postulated to be simply the proportion of sampled stimuli conditioned to that response, unless there are no conditioned stimuli in the sample, in which case, there is a 'guessing' probability for each response. Learning takes place in the following way. At the end of a trial a reinforcing event occurs which identifies that one of the possible responses which was correct. With some fixed probability the sampled stimuli become conditioned to this response, if they are not already, and the organism begins another trial in a new state of conditioning. The sequence of events occurring on a given trial may be illustrated by the following diagram.



Note that the trial begins with a certain kind of conditioning and ends with a new state of conditioning. This change of conditioning is the kernel of the learning process.[1]

To illustrate how a quantitative theory may be developed with these ideas we shall consider what is perhaps the simplest possible version. We assume that there is exactly one stimulus element and that this element is sampled on every trial by the subject. In the scheme sketched above, this reduces to triviality the sampling process. Secondly, once the single element is conditioned the response is known whenever the conditioning of that single element is known. A mathematical model that arises from this simple one-element assumption can be described in the following way. On every trial the subject – the learner in the experiment – is in one of two states. Either the single element is conditioned (state C) to the correct response or it is unconditioned (state U). In the simplest applications we wish to consider, which are also the ones pertinent to our subsequent discussion of mathematical proofs, we also assume that of the two responses available exactly one is always correct. (The restriction to two responses is not essential. What is needed is that the correct response in a given situation is well defined. We may classify the other responses simply as incorrect.) We then formulate the mathematical background of

the model in such a way that the subject's behavior forms a Markov process in these two states with the transition matrix indicated below.

	C	U
C	1	0
U	c	$1-c$

The meaning of this matrix is exceedingly simple. When the subject is in the unconditioned state, on each trial there is a probability c that he will move to the conditioned state. Once he becomes conditioned he remains so, as indicated by the probability 1. Secondly, we postulate that the subject guesses the correct response with probability p when he is in the unconditioned state, and responds correctly with probability 1 when he is in the conditioned state. In spite of the simplicity of this model, it is interesting to note that the two states of conditioning, that is, being conditioned or being unconditioned, are not themselves directly observable. In this sense even this simplest formulation of the theory already has a non-trivial theoretical component.

It will be instructive to consider two simple applications of this one-element model to experiments. I first consider a typical paired-associate experiment. The subject is shown a succession of nonsense syllables on printed cards; each nonsense syllable constitutes a stimulus, and on each trial he sees exactly one stimulus. What the subject must learn to do is to make an appropriate response when a stimulus is shown. Typically, he might be asked to respond with one of the numerals 1, 2, 3, or 4. Given a list of twenty nonsense syllables, the experimenter would arbitrarily assign five of them to each of the four numerals. On the first trial the subject is in the unconditioned state; if the experiment had been well designed, the probability p of guessing the correct answer should be 0.25. There is considerable evidence (Bower, 1961; Estes, 1960) to show that when the subject does make a firm association between a nonsense syllable and the correct response, he makes this association on an all-or-none basis and retains it throughout the experiment. The probability c of moving from the unconditioned state may be estimated from the experimental data.

The paired-associate experiment provides a paradigm of the stimulus-response conditioning connection, but it is not necessary for application of this model that the conditioning connection be conceived as holding

between a particular stimulus display and a given response. We have also applied this model extensively to concept formation in children (Suppes and Ginsberg, 1961, 1962a). Here I shall describe briefly an experiment concerned with the concept of identity of sets. The subjects were 48 children of first-grade age (6 or 7 years). On each trial the child's task was to indicate whether two sets were identical or not. There were a total of 56 trials on 28 of which the stimulus display showed identical sets and the remaining 28 nonidentical sets. The subjects were instructed to press one of two buttons when the stimulus pairs presented were 'the same' and the other button when they were 'not the same'. In this experiment no stimulus display on any trial was repeated for individual subjects. In this case the conditioning connection may be postulated as holding at different levels of abstraction. To begin with, we may assume there is a single concept of identity of sets, and the child is learning to establish the appropriate connection between this concept and the two responses. Until this connection is established he is guessing the correct answer with probability p, and after it is made he makes the correct response with probability 1.

The next natural level of analysis is in terms of two concepts, one for the pairs of sets that are identical and one for the pairs of sets that are not identical. The one-element model may then be applied to the subsequences of trials on which identical sets are displayed and again to the complementary subsequences of trials on which nonidentical sets are displayed. As the model has been formulated above and as applied to paired-associate data, it is assumed that the probabilities of conditioning are statistically independent for the two subsequences. For the analysis of any concept formation experiment in terms of more than one concept, it is necessary directly to test this assumption of statistical independence against the data.

At a still more refined level, we may analyze the stimulus displays in terms of pairs of sets that are identical in the sense of ordered sets, pairs of sets that are identical but not in the sense of ordered sets, pairs of sets that are not identical but equipollent, and pairs of sets that are not equipollent. In this case, we consider four concepts rather than one or two.

It is not the purpose of this paper to evaluate the empirical adequacy of any of the alternative ways of analyzing an experiment on identity of sets. It is worth emphasizing, however, that there is no direct way of build-

ing from the individual stimulus displays to these various concepts by simple stimulus connections when no stimulus display is repeated for an individual subject. Admittedly, it is not fully satisfactory intellectually to analyze the learning of concepts simply in terms of a conditioning connection between the concept and the correct response. Theories which postulate more details about the learning process in concept formation are needed to go beyond the present analysis. There are two things to be said about such theories at the present time. In the first place, there seems little doubt but that a good first approximation to the data may be obtained in terms of theories formulated in terms of notions of hypothesis and strategies (compare the discussion in Suppes and Atkinson, 1960, Sec. 1.7; Restle, 1961). On the other hand, at the moment, these theories have little more to offer than the simple one-element model in terms of detailed analysis of actual experimental data.

It is also not appropriate here to consider in detail statistical methods of analyzing the goodness of fit of the simple one-element model to experimental data. However, in order to sketch briefly some results for a pilot experiment at the end of this paper, I recapitulate briefly the ideas set forth in Suppes and Ginsberg (1961). There are just two basic ideas needed for essentially complete statistical analysis of the one-element all-or-none conditioning model. In the first place, the assumption that there is a constant guessing probability p that the subject responds correctly before he is conditioned implies that the sequence of responses prior to the last error of the subject is a sequence of Bernoulli trials with binomial distribution parameter p. Classical statistical tests for stationarity of the response probability, independence from trial to trial, and the actual binomial distribution of responses in blocks of fixed size may be applied to the data considered in terms of responses made prior to the last error. The conditioning parameter c, on the other hand, enters only in terms of the distribution of the last error. Across a group of subjects this last error may be estimated from the essentially geometric distribution of the last error as derived from the theory formulated above. A standard goodness-of-fit test may then be performed to see if the assumption of a homogeneous conditioning parameter for all subjects in a given experiment is acceptable.

Before turning to the specific context of mathematical proofs, there is one further remark about applications of the one-element all-or-none conditioning model which is needed. This is the application to simple

discrimination experiments. Essentially a discrimination experiment is one in which the subject needs to learn to discriminate between two or more stimuli and make the appropriate response to each. It is possible to think of a paired-associate experiment as such a discrimination experiment. On the other hand, because the discrimination itself is not difficult, it is not ordinarily described as such. However, in many cases, the problem of discrimination is one of discriminating between two stimuli that are highly similar in their perceptual characteristics. When it is assumed that the similarity is negligible, or in more technical terms, that there are no common stimuli between the two stimulus displays, the one-element model may be applied to the discrimination experiment in the same way that we have applied it above to a paired-associate experiment. For instance, suppose the subject is a rat in a *T*-maze. At the choice point of the *T*-maze a white card is placed on some trials, and on other trials, a black card. The animal must learn to turn left in order to receive food when the card is white, and to turn right when it is black. We may analyze such an experiment exactly in the manner indicated for paired-associate situations. The ideas of discrimination to be mentioned below will implicitly assume the simple context in which there is no problem of stimulus overlap.

III. AN UTTERLY TRIVIAL MATHEMATICAL SYSTEM

The simple mathematical system we shall analyze in terms of the behavioral ideas just discussed is concerned with production of finite strings of 1's and 0's. Any finite string of 1's and 0's is a well-formed formula of the system. The single axiom is the single symbol 1. The four rules of inference are:

R1. $S \rightarrow S11$
R2. $S \rightarrow S00$
R3. $S1 \rightarrow S$
R4. $S0 \rightarrow S$,

where S is a nonempty string. A theorem of the system is, of course, either the axiom or a finite string that may be obtained from the axiom by a finite number of applications of the rules of inference. A general characterization of all theorems is immediate: any finite string is a theorem if and

only if it begins with 1. A typical theorem in the system is the following one, which I have chosen because it uses all four rules of inference.

Theorem 101
(1) 1 Axiom
(2) 100 R2.
(3) 10 R4.
(4) 1011 R1.
(5) 101 R3.

It is apparent that a shorter proof of this theorem could not be given, and this is generally true of this system. A proof of minimal length of any theorem is easily found, but it is not the case that there is exactly one proof of minimal length. For instance, if we want to prove the theorem 111 we may apply rule R1 twice to obtain 1111 and then remove the last 1 by applying R3; or we may interchange the position of the application of R3 and prove the theorem by first using R1 then R3 and then R1 again. Counting the introduction of the axiom as one line, the two proofs are both of length four in terms of number of steps. (I have not fully formalized the system here by giving a recursive definition of proof, etc., because it is completely obvious how these matters go for a system of this kind; I want to give only enough formal detail to make the mathematical system definite.)

For simple reference in the behavorial analysis to follow let us call this mathematical system, the system *U*.

IV. BEHAVIORAL ANALYSIS OF PROOFS IN THE SYSTEM

Initially, it will be simplest to ignore the possibility of more than one proof of minimal length and consider only an analysis that will always yield exactly one proof of minimal length. The stimulus discrimination facing the subject on each trial is simply described. He must compare the last line of proof in front of him with the theorem to be proved. This comparison immediately leads to a classification of each last line of a proof into one of four categories: additional 1's need to be added to match the theorem (R1); additional 0's need to be added to match the theorem (R2); a 1 must be deleted to continue to match the theorem (R3); or, finally, a 0 must be deleted in order to match the theorem (R4). The rule

that should be applied once the stimulus comparison has been made is indicated in parentheses. When the subject is completely conditioned to all four stimulus discrimination situations, he will make the response corresponding to applying that rule. For each of the four discriminations with respect to which he is not yet conditioned, there is a guessing probability p_i, $i=1, 2, 3,$ or $4,$ that he will guess the correct rule and thus a probability $1-p_i$ that he will guess incorrectly. Also, when the subject is unconditioned, for any one of the discrimination comparisons there is a probability c_i that he will become conditioned on the next trial. On the assumption of statistical independence made earlier, we may then analyze separately the four subsequences of trials on which the four stimulus discrimination categories appear. It is to be emphasized again that the four guessing parameters p_i and the four conditioning parameters c_i are to be estimated from the experimental data.

The example given in the preceding section of two proofs of minimal length shows that the analysis just stated represents a slight oversimplification. For example, if exactly one 1 is needed and two 1's are at the end of the string standing as the last line of proof, it will be just as efficient first to apply R3 and then R1 as to apply R1 and then R3. It does not require serious modification of the behavioral analysis of the system U to take account of this fact. Reinforcement can be given randomly of either of the rules that are correct in terms of rendering a minimal proof and either one of these responses can be counted as correct when made. This leaves the analysis in terms of the two states of conditioning untouched. It does change the relation between the state of conditioning and the probability of a response. The weaker requirement for this discrimination is that the probability of making one of the two correct responses is 1. We could, if we so desired, analyze the system U in such a way that there were more than four discriminating situations in order to take account of the cases in which more than one response was correct. For an extensive experiment this would be desirable. In terms of the pilot study to which I would now like to turn, it is not necessary.

The pilot study was conducted with a group of first-grade children (ages 6 and 7) in an elementary school near Stanford University. Initially, we considered doing the experiment with fourth-grade children (ages 9 and 10), but preliminary testing with a few children of this age indicated that the experimental problem was far too easy for them; most of the fourth-

grade children made no errors at all. There is a variety of evidence to indicate that the reasoning abilities of young children are far superior to their ability to put an argument in written form (see e.g., Hill, 1961). For this reason we attempted to avoid entirely a written context for the exhibiting of proofs in the system *U*. Our procedure was the following.[2]

We considered only theorems which are of length greater than one and less than seven. There are 62 members of this class of theorems. We ordered their proofs according to the following four criteria of simplicity.

(1) If proof P_1 may be obtained from proof P_2 by interchanging 1's and 0's in all lines, then P_1 and P_2 are of equal simplicity.

(2) If $m<n$ then a proof of length m is simpler than a proof of length n.

(3) If P_1 and P_2 are minimal proofs of the same length, P_1 is one of several alternative minimal proofs and P_2 is a unique minimal proof, then P_1 is simpler than P_2.

(4) If P_2 is not simpler than P_1 by virtue of Criteria 2 or 3 and P_1 uses a smaller number of different rules of inference than does P_2, then P_1 is simpler than P_2.

These four criteria arrange the 62 members into 17 equivalence classes of increasing complexity. Within an equivalence class the theorems were randomized and each subject was presented with a sequence of 17 theorems, as determined by random selection from each of the classes.

The four criteria of simplicity are not necessarily the only ones or even the best ones with which to begin. They did provide jointly a workable basis for arranging the theorems of length not greater than six.

The apparatus consisted of a plywood board, approximately 20 in. by 6 in., placed horizontally. The top half of the board was slotted for insertion of cards 20 in. by $2\frac{1}{2}$ in. Theorems were written on these large cards. The bottom half of the board had hooks at $2\frac{3}{4}$ in. intervals to permit hanging 2 in. by $2\frac{1}{2}$ in. cards on which were printed either a 0 or a 1. Each theorem was written on one of the large cards in such a way that it could be matched directly below it by hanging the appropriate small cards on the hooks.

Subjects were instructed as follows:

This is a string of zeros and ones. Whenever I put a string of zeros and ones up here you are to put a string just like it down here. Every string that I shall put up begins with a one, and so we'll leave these ones in the first places.

Now of course if you could hang zeros and ones on the board any way you wanted,

the game would be terribly easy. Why don't you try that now: Make a string of zeros and ones just like this one underneath it.

That's right. Now that was too easy to be much fun, wasn't it? In the game we are going to play now there are some rules about the ways in which you may make strings. These are the rules:

First, you may put two zeros on the end of a string you have made, but you may not put one zero by itself on the end. With ones it's the same way: you may put two ones on the end, but you may not put one one by itself on the end.

If you wish though, you may take the last card off the end of a string, whether it's a one or a zero, and put it back in its box, but you may only remove the last card, none of the others.

So the rules are: You may add two zeros to the end of a string, add two ones to the end of a string, take a zero off the end of a string, or take a one off the end of a string, except that you may not take away this first one (practice). Now would you tell me the rules, just to make sure you understand them?

Now each time you put cards on or take a card away, I want you to tell me what you're doing; what rule it is that you're following. That is, say, "I'm adding two zeros", or "I'm adding two ones", or "I'm taking off the last zero", or "I'm taking off the last one".

Subjects were divided into two groups. Subjects in the *correction* group were corrected for each wrong step in each proof. Subjects in the other group (the *discovery* group) were stopped only when a valid proof was not completed in three times the length of a minimal proof. At the end of each proof, subjects in the discovery group were shown a minimal proof or – in the event that the subject constructed a minimal proof – told that the proof constructed was correct. For theorems which have several alternative proofs, we considered either proof correct and when demonstrating a minimal proof selected randomly.

Subjects made correction responses overtly, that is, in the correction group they removed the cards themselves and put up correct cards (or took down the correct cards) as instructed. In the discovery group they executed a minimal proof following instructions.

Actually not all subjects were required to prove 17 theorems. The following criterion rule was used. We considered that a subject had learned how to give minimal proofs in the system when four correct theorems were proved in succession. However, it was required that each subject prove at least ten theorems. All subjects, except for two in the discovery group, satisfied this criterion by the time the seventeenth theorem was reached. (It should be noted, however, that the nine subjects in each group represent a net figure; approximately this same number of subjects were discarded because of various kinds of problems that arose:

seeming failure to comprehend the instructions at all, lack of attention after the first few theorems, etc.)

Because there were only nine subjects in each group the empirical data are not to be taken seriously and no detailed statistical analysis shall be presented here. I have summarized in Table I below the mean proportion

TABLE I

Observed proportion of errors prior to last error for the correction, discovery and combined groups. Blocks of 12 trials

	Block				
Group	1	2	3	4	5
Correction	.28	.23	.15	.00	.10
Discovery	.23	.20	.40	.30	.33
Combined	.25	.21	.30	.18	.24

of errors prior to the last error in blocks of twelve trials for each group and for the two groups combined. A trial in this instance is defined as a step in a proof and not as an entire proof. There were more than sixty trials, that is, more than a total of sixty lines of proof in the seventeen theorems, but because very few subjects needed the entire seventeen theorems to reach criterion, it has been necessary to terminate the mean curves with the last block ending on trial 60. Several things are to be noted about the data in Table I. In the first place, the correction group seems to have done better than the discovery group, which result is consistent with experiments of a similar character dealing with the effects of immediate reinforcement. Secondly, the discovery group is more or less stationary (i.e., the learning curve in terms of responses prior to the last error is approximately flat). If anything, there is a tendency for the proportion of errors to increase with trials, whereas the correction group is clearly not stationary, and there is a definite tendency for the proportion of errors to decrease with trials. When the two groups are combined, an approximately stationary learning curve is obtained. The problem for future investigation is to discover which of these effects will be observed in larger and more stable bodies of data. In interpreting Table I it should be emphasized again that these figures are based on responses prior to the last error.

Naturally, if the full set of data were considered, the learning curves would approach 1 at trial 60. It is also important to emphasize that the data of Table I are based on considering only a single sequence of trials. There is no analysis of the data into the separate subsequences defined by the various stimulus comparisons. This lumping together of the four sets of stimulus discrimination situations could in itself account for the lack of stationarity for the correction group, because responses were not deleted from consideration after the appropriate rule became conditioned to the stimulus discrimination. The rule for considering responses prior to the last error was invoked only for the whole sequence and not for the subsequences.

There was also some evidence against the independence of responses as shown by the figures given in Table II. Here are shown the conditional

TABLE II

Conditional probability prior to last error of a correct response following a correct response and following an error

Group	After correct R.	After error
Correction	.85	.77
Discovery	.81	.58
Combined	.83	.64

probabilities of a correct response following a correct response and following an incorrect response for the correction, discovery and combined groups. The biggest difference occurs for the discovery group, which has a mean probability of 0.81 that a correct response will follow a correct response in comparison to a mean probability of 0.58 that a correct response will follow an incorrect response.

From this preliminary evidence it is perhaps doubtful that the one-element all-or-none conditioning model will fit very well the fine structural details of experimental data derived from young children learning proofs in the mathematical system *U*. On the other hand, this model does seem to give a pretty good first approximation to actual behavior. To give some idea of the immediate range of alternative possibilities I shall

briefly sketch three other models that might be applied to data similar to those obtained from our pilot study.

The simplest alternative model is the linear incremental model with a single operator. The intuitive idea of this model is precisely the opposite of the all-or-none conditioning model. The supposition is that learning proceeds on an incremental basis. Let q_n be the probability of an error on trial n. Then the model is formulated by the following recursive equation

$$(1) \qquad q_{n+1} = (1-\theta)\, q_n,$$

where $0<\theta\leqslant 1$. It is simple to show but somewhat surprising that this purely incremental model has precisely the same mean learning curve as the all-or-none model if we set $c=\theta$. (To obtain this identity of the learning curves we must consider all responses and not simply responses prior to the last error.) The incremental model does differ sharply from the all-or-none model in the kind of learning curve predicted for responses prior to the last error, as is evident from Equation (1).

The second simple alternative is a model which represents a kind of compromise between the all-or-none model and the incremental model. It assumes that associated with each discrimination situation there are two elements. Each of these (unobserved) elements is conditioned on an all-or-none basis but the two parameters of conditioning may be adjusted to produce various incremental effects on the response probabilities. A model of this kind, as is pointed out in Suppes and Ginsberg (1961), could account fairly well for the kind of data shown in Tables I and II. Probably its main inadequacy for accounting for more extensive data obtained from a large number of subjects would be found in connection with the problem of the assumed independence of the subsequences defined by the four types of stimulus comparisons.

The third alternative is to start with one of the three models already sketched and to introduce in a natural way dependencies among the subsequences. To introduce these dependencies we define a new process whose states are now ordered quadruples. The first coordinate of the quadruple indicates the state of conditioning of Rule R1, the second coordinate the state of conditioning of Rule R2, etc. One natural direction is then to define conditioning parameters c_{ij} for each Rule i, where j is the number of other rules already conditioned. By assuming further that the

parameters depend only on j and not on i, we once again obtain a process with four conditioning parameters but with the parameters defined in an entirely different way. Without a large set of data to analyze and thereby to decide among these various alternatives it does not seem profitable to pursue them in any detail. Experiments are now underway in our laboratory and I hope to be able to report soon which models are most able to account for the fine structure of the data.

V. GENERAL COMMENTS

I would like to conclude with two general comments. In the discussion following the original presentation of this paper in Warsaw, Professor Kalmár appropriately raised the question of how the reinforcement schedule, i.e., the correction procedure, would be defined for more complicated mathematical systems than U, in particular for systems which do not possess a decision procedure. It should be apparent that the behavioral theory outlined above is certainly not yet powerful enough to specify clear recipes for laying out the schedule of reinforcements. At the present time for more complicated systems, for example, the elementary algebra of integers and real numbers, the only practical procedure seems to be to proceed in a manner very similar to that used by Newell and Simon (1956) and Newell *et al.* (1957) in working out a program for proving theorems of elementary logic. Essentially their procedure is to abstract those heuristic principles that seem most useful in giving the set of proofs under consideration. My own conjecture is that in this area we shall find a substantial intersection between the work of mathematical psychologists interested in behavior theory and scientists like Simon who are interested in artificial intelligence and computer simulation of human behavior.

My second general comment is to emphasize that I am under no illusions about the fragmentary character of the behavioral foundations sketched in this paper. The next step forward it seems to me is to provide a theory at the following level of generality. Suppose we retained the problem of proving theorems in systems whose well formed formulas are strings of 1's and 0's. As rules of inference we have various rules of production of the kind given for the system U. The problem is to formulate in sufficiently general terms the behavior theory that will lead to

appropriate conditioning connections in at least a fairly wide class of systems similar to *U*. The weakness of the present theory is easily brought out by considering the problem of using the theory to build a machine to prove theorems in such systems. It is clear that a machine could not be programmed, on the basis of the present theory, in a general way to prove theorems in systems similar to *U*. It is of course a trivial matter to program a machine to prove theorems in systems like *U* if the programming is done for that particular system after the rules of inference of the system are specified. The much deeper problem of programming a machine to accommodate itself to proofs in a variety of systems similar to *U* seems to me to be one of the most pressing problems to solve in order to provide a more adequate behavioral foundation of mathematical proofs. The solution of this problem will be of direct help in constructing a more general theory to predict the "proof-giving" behavior of our young subjects.

NOTES

[1] For an explicit axiomatic formulation of these ideas, see Suppes and Atkinson (1960, p. 5); for a more complete discussion of the technical aspects of axiomatization, see Estes and Suppes (1959b) or Article 23 in this volume.

[2] I am much indebted to John M. Vickers for his contributions to the detailed design and actual execution of this pilot study. Susan Matheson assisted Mr. Vickers in running the experiment.

21. THE PSYCHOLOGICAL FOUNDATIONS OF MATHEMATICS*

I. INTRODUCTION

I would like to say to begin with that it is a pleasure to be here and to participate in this colloquium on models. For the topic of my own lecture today I am somewhat hesitant in view of the fact that Professor Piaget is sitting here and has been writing on this topic for many years. I wish that I had confidence that the kind of things I want to say will turn out to be the right things, the significant things to suggest in investigations on the psychological foundations of mathematics, but I have no such confidence. Secondly, it is perhaps paradoxical considering the subject of this colloquium and my own interests that I shall not have more to say of a direct sort about models, but it seems to me that the problems raised by the learning of mathematics provide an excellent touchstone for testing and evaluating models, particularly with respect to many of the issues we have already discussed. In the cognitive domain mathematics provides one of the clearest examples of complex learning, for the structure of the subject itself provides numerous constraints on the structure of any models that are to be considered adequate to mathematics learning. Therefore I hope to justify, in the context of the present colloquium, my own concern with the psychological foundations of mathematics by emphasizing the importance of the kind of learning found in mathematics for the development of complex models of learning. I would agree wholeheartedly with those two good cognitivists Frank Restle and Herbert Simon that simple stimulus models are certainly not adequate to give a very deep account of mathematics learning. On the other hand, I am equally skeptical of the cognitive models that have as yet been proposed with respect to the central problems of giving such an account, although I have a great deal of respect and appreciation for the kind of

* Reprinted from *Les Modèles et la Formalisation du Comportement* (Colloques Internationaux du Centre National de la Recherche Scientifique), Editions du Centre National de la Recherche Scientifique, Paris, 1967, pp. 213–234.

thing that Restle, Simon and their associates have been concerned with over the past few years.

Before I begin discussing particular psychological issues there is another direction of interest quite apart from models that I want to mention, and that is the relation of the psychological foundations of mathematics to foundations of mathematics in the classical sense, and by the *classical sense* I mean the work in foundations that has been characteristic of this century. The three main positions in the twentieth century on the foundations of mathematics characteristically differ in their conception of the nature of mathematical objects. Intuitionism holds that in the most fundamental sense mathematical objects are themselves thoughts or ideas. For the intuitionist formalization of mathematical theories can never be certain of expressing correctly the mathematics. Mathematical thoughts, not the formalization, are the primary objects of mathematics. Yet the nature of mathematical thinking has scarcely been seriously discussed from a psychological standpoint on the part of any intuitionist.

The second characteristic view of mathematical objects is the Platonistic one that mathematical objects are abstract objects existing independently of human thought or activity. Those who hold that set theory provides an appropriate foundation for mathematics usually adopt some form of Platonism in their basic attitude toward mathematical objects. The philosophy of Bourbaki, for example, is that of Platonism.

The view of mathematical objects adopted by the formalists is something else again. According to an often quoted remark of Hilbert, formalism adopts the view that mathematics is primarily concerned with the manipulation of marks on paper. In other words, the primary subject matter of mathematics is the language in which mathematics is written, and it is for this reason that formalism goes by the name 'formalism'.

In spite of the apparent diversity of these three conceptions of what mathematics is about – and certainly they differ extraordinarily in their conception of the proper object of mathematical attention – there is a very high degree of agreement about the validity of any carefully done piece of mathematics. The intuitionist will not always necessarily accept as valid a classical proof of a mathematical theorem, but the intuitionist will, in general, always agree with the classicist as to whether or not the theorem follows according to classical principles of construction and inference. There is a highly invariant content of mathematics recognized

by all mathematicians, including those concerned with the foundations of mathematics, which is absolutely untouched by radically different views of the nature of mathematical objects. It is also clear that the standard philosophical methods for discussing the nature of mathematical objects do not provide appropriate tools for characterizing this invariant content. A main thesis of this paper is that the classical philosophical discussions of the nature of mathematical objects may fruitfully be replaced by concentration, not on mathematical objects, but on the character of mathematical thinking. There is reason to hope that by concentration on mathematical thinking or mathematical activity we can be led to characterize the invariant content of mathematics. Or, to put it another way, to get at the nature of working mathematics without commitment to a particular philosophical doctrine.

My original title for this paper was 'Behavioral Foundations' rather than 'Psychological Foundations'. The reason for changing is the desire to avoid the charge of attempting to reduce mathematics to the kind of considerations exemplified in Skinner's *Verbal Behavior* (1957). Moreover, it is an increasing conviction of mine that the classical concepts of behaviorism, namely, those of stimulus, response and reinforcement, are not, at least in their standard formulation, nearly adequate for any complicated behavior, and in particular, for the intellectual activity of mathematicians and scientists.

It will perhaps be desirable to make this point somewhat more explicit, particularly because of the considerable interest in this colloquium in the formal properties of models. It would be too substantial a digression to present possible formal axiomatizations of stimulus-response theory and then to analyze in this rather detailed and cumbersome framework the severe limitations on accounting for the formation of new concepts in the repertoire of a subject. The essential idea of the argument that shows how severe these limitations are can be presented within various fragments of stimulus-response theory.

The first thing to be noticed in considering the question of what does the theory say about the formation of new concepts out of old ones is that many details of the learning process are irrelevant. For example, for analysis of this problem it is not essential to know whether learning is mainly all-or-none or incremental. The second thing to note is that unless the theory has sufficient apparatus for defining new concepts in terms of

old ones the theory cannot give a systematic account of how the new concepts are learned.

This logical question of definability is central to my argument, and a simple example of a purely mathematical sort may be useful in clarifying the method by which it may be shown that one concept may not be defined in terms of other concepts.

Consider first the ordinal theory of preference based on a set A of alternatives, a binary relation P of strict preference and a binary relation I of indifference, where P and I are relations on A. A triple $\mathfrak{A}=(A, \mathrm{P}, \mathrm{I})$ is an *ordinal preference pattern* if and only if the following three axioms are satisfied for every x, y and z in A:

Axiom 1: *If* $x \mathrm{P} y$ *and* $y \mathrm{P} z$ *then* $x \mathrm{P} z$;

Axiom 2: *If* $x \mathrm{I} y$ *and* $y \mathrm{I} z$ *then* $x \mathrm{I} z$;

Axiom 3: *Exactly one of the following:* $x \mathrm{P} y$, $y \mathrm{P} x$, $x \mathrm{I} y$.

The Italian mathematician Alessandro Padoa formulated in 1900 a principle that may be used to show in a rigorously definite way that one concept of a theory is not definable in terms of the others. The principle is simple to formulate: find two models of the theory such that the given concept is different in the two models, but the remaining concepts are the same in both models. It is easy to show that if the given concept were now definable in terms of the other concepts then it would be possible to derive a formal contradiction within the theory. (For a more detailed discussion of these matters, see Chap. 8 of my *Introduction to Logic*.) Thus to show that the concept P of strict preference cannot be defined in terms of the concept of the set A of alternatives and the concept I of indifference, it is sufficient to consider the following two models $\mathfrak{A}_1$ and $\mathfrak{A}_2$ of the theory.

$$A_1 = A_2 = \{1, 2\}$$
$$\mathrm{I}_1 = \mathrm{I}_2 = \{(1, 1), (2, 2)\}$$
$$\mathrm{P}_1 = \{(1, 2)\}$$
$$\mathrm{P}_2 = \{(2, 1)\}$$

Note that two trivial numerical examples of ordinal preference patterns are sufficient to establish the undefinability of the concept of strict preference. On the other hand, it is easy to offer a definition of indifference in terms of strict preference:

$$x \mathrm{I} y \quad \text{if and only if} \quad \text{not } x \mathrm{P} y \text{ and not } y \mathrm{P} x.$$

If this example is kept explicitly in mind, it will be easier to appreciate the point I want to make about any current variant of stimulus-response theory of concept formation. One way or another the theory must be rich enough to make possible the formal definability of the new concept to be learned. I can see no other way of giving a formal account of learning the new concept. If the machinery does not exist within the theory for characterizing the new concept, then the theory cannot give an adequate account of how the new concept is formed by the subject. In this connection it is important to emphasize how incomplete are all standard learning-theoretic accounts of concept formation. Current theories simply do not postulate mechanisms of concept formation which are adequate to even the most primitive and simple concepts, let alone ones of any mathematical complexity.

To illustrate this failure, we may consider some examples of the sort often studied experimentally. In line with earlier remarks, I shall ignore detailed assumptions about learning and give a schematic account that is compatible with any one of several fully worked-out learning models. As a matter of notation, let S be the basic set of stimuli, and given concepts may be represented as partitions $C_1, \ldots, C_n$ of S. In general, each C_i is a partition of S, although in many familiar experimental examples the concepts are just two-valued and thus lead to concepts that may be represented as subsets of S. Let new concepts be represented as partitions $K_1, \ldots, K_m$ of S. The first general point to note is that if we are simply given an $m+n+1-$tuple $\mathfrak{S}=(S, C_1, \ldots, C_n, K_1, \ldots, K_m)$ then no questions about generating the concepts K_j from the given concepts C_i can be definitely settled. It is necessary also to specify what operations may be performed on the given C_i, or what additional structure is imposed on the basic set S of stimuli. It is a matter of the postulated psychological theory of concept formation to impose this additional structure.

In familiar experiments on concept identification it is assumed that the intersection, union and complement of two-valued concepts can be formed, but these Boolean operations are weak. Certainly they are not adequate to give an account of the formation of any complex mathematical concepts. For example, if we assume in an experiment, for purposes of theoretical analysis at least, that an individual has the concepts of shape, size and color, with indefinitely many values for each concept, we cannot in terms of the Boolean operations, or their generalizations to

partitions, define or characterize any of the intuitively simple comparative concepts of greater size, more saturation of color, etc.

Although the matter cannot be pursued in detail here, it should be all too obvious to those familiar with the psychological literature of concept formation that the structures of the mechanisms of concept formation as yet proposed are far too simple, as a direct application of Padoa's method will show, to account for the formation of the great variety of mathematical concepts.

Because of their central importance for any theory of concept formation in mathematics, the three topics I shall concentrate on in the remainder of this paper are *abstraction*, *imagery* and *algorithms*.

II. ABSTRACTION

It has long been customary, although probably less so now than previously, to talk about abstract set theory or abstract group theory. To a psychologist or philosopher concerned with the nature of mathematics, it is natural to ask what is the meaning of 'abstract' in these contexts. There is, I think, more than one answer to this query. One possibility is that abstract often means something very close to 'general', and the meaning of 'general' is that the class of models of the theory has been considerably enlarged. The theory is now considered abstract because the class of models of the theory is so large that any simple imagery or picture of a typical model is not possible. The range of models is too diverse.

In the case of group theory, for example, one intuitive basis was the particular case of groups of transformations. In fact, the very justification of the postulates of group theory is often given in terms of Cayley's theorem that every group is isomorphic to a group of transformations. It has been maintained that the "basic" properties of groups of transformations have been correctly abstracted in the abstract version of the axioms just because we are able to prove Cayley's theorem. So we can see that another sense of *abstract*, closely related to the first, is that certain intuitive and perhaps often complex properties of the original objects of the theory have been dropped, as in the case of groups, sets of natural numbers, or sets of real numbers, and we are now prepared to talk about objects satisfying the theory which may have a very much simpler internal

structure. This meaning of *abstract*, it may be noted, is very close to the etymological meaning.

Under still another, closely related sense of the term, a theory is called *abstract* when there is no one highly suggestive model of the theory that most people think of when the theory is mentioned. In this sense, for example, Euclidean plane geometry is not abstract, because we all immediately begin to think of figures drawn on the blackboard as an approximate physical model of the theory. In the case of group theory the situation is different. It would indeed be an interesting question to ask a wide range of mathematicians what is called to mind or what imagery is evoked when they read or think about, let us say, the associative axiom for groups or the axiom on the existence of an inverse. Or as another suitable example, what sort of stimulus associations or imagery do they have in thinking about the axiom of infinity in set theory? It is my own conjecture that the combinatorial, formalist way of thinking is much more prevalent than many people would like to admit. Many mathematicians, particularly those with an algebraic tendency, have as the immediate sort of stimulus imagery the mathematical symbols themselves and think very much in terms of recombining and manipulating these symbols.

It is interesting to note that the classical search for a representation theorem for a theory can very well be thought of as an effort to make the abstract theory more intuitive. The formal idea of a representation theorem can be put as follows. We begin by discussing the class or category M of all models of the theory. We then seek a subclass or subcategory R of models of the theory such that given any model in M there exists an isomorphic model in the representing class R. We may of course always obtain a trivial representation theorem by simply taking $R=M$, but the satisfying representation theorems are just those that are able to select as the class R an intuitively clear and relatively simple class of models. Cayley's theorem is a good example. Another classic example is Stone's representation theorem for Boolean algebras. Many of us would have had a feeling that we did not quite understand what the abstract theory of Boolean algebras came to if Stone's theorem had proved not to be true. The motivation for Boolean algebras is mainly thought of in terms of the algebra of sets, but if the abstract theory has models of Boolean algebras that are not isomorphic to algebras of sets, what indeed are we to make of the structure of these abstract algebras? Stone's theorem

shows that we do not have any worries on this score, but in the history of mathematics and science many negative examples can be mentioned, in which the move to a more abstract theory was not buffered by the proof of an appealing representation theorem, but these matters cannot be pursued in further detail here.

III. IMAGERY

Mathematicians classify each other as primarily geometers, algebraists, or analysts. The contrast between the geometers and algebraists is particularly clear in folklore conversations about imagery. The folklore version is that the geometers tend to think in terms of visual geometrical images and the algebraists in terms of combination of symbols. I do not know to what extent this is really true, but it would be interesting indeed to have a more thorough body of data on the matter. To begin with, it would be desirable to have some of the simple association data which exist in such abundance in the experimental literature of verbal learning. Such association data would be an interesting supplement to the kind of thing discussed and reviewed in Hadamard's little book on the psychology of mathematics.

I tend to think of the concepts of imagery and abstraction as closely related. I could in fact see attempting to push a definition of abstraction as the measure of the diversity of imagery produced by a standard body of mathematics and stimulus material in a given population.

As one kind of investigation connected with imagery in abstraction, the following sort of modification of the standard association experiment is of considerable interest. With a standard body of mathematical material we would set students to work proving theorems from the axioms of different mathematical systems. It would, of course, be interesting to take axioms from different domains; for example, to compare Euclidean geometry and group theory. As the subjects proceeded to prove theorems we would at each step ask for their associations. Two sorts of questions would be of immediate interest. What is the primary character of the associations given? Secondly, what kinds of dependence exist between the association given at different stages in the proof of a given theorem, or in proofs of successive theorems of a given system? As far as I know, no investigations of this sort have yet been conducted. On the

other hand, such experiments should not be difficult to perform and the results might be of interest.

I have undoubtedly put the matter too simply. One main problem is to distinguish between associations that play an essential and important role in obtaining the proof, and those which are more or less accidental accompaniments of the central activity of finding the proof. For example, a person may read a theorem about geometry, written in English words, and as he begins to search for a proof of this theorem, he associates to simple geometrical figures – in particular, to the sort of figure useful for setting up the conditions of the theorem. At the same time that he has this geometrical association, he may have associations about his wife, his mother, or his children. We would not want to think of these latter associations as playing the same sort of role in finding proofs. In other words, we want to see to what extent a chain of associations may be identified, which is critical for the heuristic steps of finding a proof. It is also important, I am sure, to separate the geometrical kind of case from the other extreme – as a pure case, the kind of thinking that goes on when one is playing a game such as chess or checkers. What kind of associations are crucial for finding a good move in chess, checkers, or, to pick a different sort of example, bridge?

An experiment we have conducted in our laboratory has some bearing on these questions. This experiment concerned the possible differences between learning rules of logical inference in a purely formal way and as part of ordinary English. The three rules studied were

Det	Sim	Com
$P \rightarrow Q$	$P \wedge Q$	$P \wedge Q$
P	P	$Q \quad P$
Q		

(Here $\rightarrow$ is the sign of implication and $\wedge$ the sign of conjunction, but subjects were not told this when they began the formal part of the experiment.) Group 1 received the formal part first (FA) and the interpreted logic in ordinary English (IB). Group 2 reversed this order: IA then FB. Note that A stands for the first part of the experiment and B for the second part. Schematically then:

Group 1. FA + IB
Group 2. IA + FB.

The formal (F) and interpreted (I) parts of the experiment were formally isomorphic.

TABLE I

Comparisons of errors on different parts of logic experiment

Comparison	t	df	Significance
FA > FB	1.94	46	0.1
IA > IB	3.28	46	0.01
FA ≠ IA	1.47	46	–
FB ≠ IB	0.08	46	–
FA + FB ≠ IA + IB	1.15	94	–
FA + IB ≠ IA + FB	1.07	94	–

Some of the results are shown in Tables I and II. The subjects were fourth graders with an *I.Q.* range from 110 to 131; there were 24 subjects in each group.

TABLE II

Vincent learning curves in quartiles for logic experiment

	Probability of error in each quartile			
Group	1	2	3	4
FA	0.40	0.36	0.39	0.24
IB	0.32	0.32	0.30	0.19
IA	0.48	0.41	0.33	0.28
FB	0.21	0.21	0.28	0.14

Perusal of Tables I and II indicates that the order of presentation, formal material first or last, does not radically affect learning. There is, however, some evidence in the mean trials of last error that there was positive transfer from one part of the experiment to the other for both groups. For example, the group that began with the formal material had a mean trial of last error of 14.1 on this part, but the group who received this material as the second part of their experiment had a smaller mean trial of last error of 10.9. In the case of the interpreted part, the group beginning with it had a mean trial of last error of 18.3, but the group that

received this material after the formal part had a mean trial of last error of 7.7, a very considerable reduction. Now one way of measuring the amount of transfer from one concept or presentation of mathematical material to a second is to consider the average mean trial of last error for both concepts in the two possible orders. If we look at the logic experiment from this standpoint there is a significant difference between the group beginning with the formal material, completely uninterrupted as to meaning, and the group beginning with the interpreted material. The average trial of last error on both parts of the experiment for the group beginning on the formal part is 10.9 and that for the group beginning on the interpreted part is 14.6. In a very tentative way these results favor an order of learning of mathematical concepts not yet very widely explored in curriculum experiments.

IV. ALGORITHMS IN ARITHMETIC

I conclude this paper with consideration of a pedagogically important and theoretically interesting example of a problem that needs deeper psychological analysis, namely, the problem of how the first algorithms in arithmetic are learned.

As an initial model for thinking about algorithms, I would like to propose the following. We have in mind a given collection of problems that we wish the child to be able to solve. To make our analysis definite at this point, let us consider a set of arithmetical problems. They might be in the form of $8-5=3$, $8+2=10$, $10-6=4$, $8-3=5$, etc. The machinery needed to solve these problems can be roughly divided into two parts. One part consists of direct storage in memory of certain elementary facts. Exactly what these elementary facts are will vary from stage to stage in the curriculum. Towards the beginning of arithmetic, it might consist of storage of the elementary addition facts: $1+1=2$, $1+2=3$, $2+1=3$, $1+0=1$, $2+0=2$, $3+0=3$, $0+3=3$, etc. The second part of the machinery consists of algorithms, or constructive rules, for transforming the elementary facts in memory into new elementary facts or, what is probably more important, transforming new stimulus presentations into one of these elementary facts stored in memory.

An immediate problem of psychological importance with respect to a given body of problems is how much should be stored in memory and

how much should be carried by the algorithmic rule. It is seldom the case that for a given set of problems we want all the answers stored directly in memory – it is certainly contrary to the usual spirit in teaching mathematics, but it is also unusual to want to store in memory only a minimal set of facts. For illustrative purposes, let me describe in some detail a way of teaching arithmetic that would consist of storing in memory a small number of facts and transferring the larger part of the load to the algorithmic rules. I emphasize that the example chosen is not one that is meant to have direct pedagogical applications. This system for computing sums is clearly not the sort of system we would wish to teach.

Let us suppose that our set of problems is just the following thirty

$$
\begin{array}{lll}
0+0=n & 0+n=0 & n+0=0 \\
0+1=n & 0+n=1 & n+1=1 \\
1+0=n & 1+n=1 & n+0=1 \\
1+1=n & 1+n=2 & n+1=2 \\
2+1=n & 2+n=3 & n+1=3 \\
1+2=n & 1+n=3 & n+2=3 \\
3+1=n & 3+n=4 & n+1=4 \\
1+3=n & 1+n=4 & n+3=4 \\
2+2=n & 2+n=4 & n+2=4 \\
4+1=n & 4+n=5 & n+1=5.
\end{array}
$$

We put the following four facts in memory

$$
\begin{array}{l}
1+1=2 \\
2+1=3 \\
3+1=4 \\
4+1=5.
\end{array}
$$

We have the following four rules of operation:

(1) Use the four facts in memory to replace equals by equals.
(2) Replace a term of the form $a+(b+c)$ by $(a+b)+c$, or vice versa.
(3) Replace a term of the form $a+b$ by $b+a$.
(4) Cancel an equation of the form $a+n=a+c$ to get $n=c$.

These four rules are then used to transform a problem, step by step, until we reach an expression of the form $n=c$. Thus,

$$
\begin{aligned}
2+2&=n && \text{Problem}\\
2+(1+1)&=n && \text{by (1)}\\
(2+1)+1&=n && \text{by (2)}\\
3+1&=n && \text{by (1)}\\
4&=n && \text{by (1)}
\end{aligned}
$$

or, similarly,

$$
\begin{aligned}
3+n&=5 && \text{Problem}\\
3+n&=4+1 && \text{by (1)}\\
3+n&=(3+1)+1 && \text{by (1)}\\
3+n&=3+(1+1) && \text{by (2)}\\
3+n&=3+2 && \text{by (1)}\\
n&=2 && \text{by (4).}
\end{aligned}
$$

There are several immediate criticisms to be made of this set-up, as I have described it. First, I have not been really explicit about parentheses in connection with rule (1). And I have not really made clear the role of the associative law, i.e., rule (2). More importantly, I have not written down a genuine algorithm for the set of problems. The four rules are four rules of proof, not an algorithm for solving any one of the thirty problems.

To convert the four rules into an algorithm, it is necessary to specify an order in which they are to be applied, and this order, to be efficient, should vary with the particular problem. Not only is it necessary to specify an order, but it also is necessary to show that the algorithm can be given to a machine and automatically used to solve any of the thirty problems.

To convert the present four rules into a genuine algorithm is somewhat tedious. Let me describe another simpler system that may be used to solve the same thirty problems.

We put in memory the following five definitions:

$$
\begin{aligned}
1&=/\\
2&=//\\
3&=///\\
4&=////\\
5&=/////.
\end{aligned}
$$

Our algorithm is then the following:

(1) Replace all Arabic numerals by their stroke definitions and delete all plus symbols.

(2) If there are strokes on both sides of the equal sign, cancel one-by-one starting from the left of each side until there remain no strokes on one side. Ignore n in cancelling.

(3) On the one side still having strokes, replace the strokes by an Arabic numeral, using the definitions in memory.

The solution in the form $n=c$ or $c=n$ will result.

Let us apply this algorithm to the two problems previously considered.

First problem:

$2+2=n$	Problem
$//\ \ //=n$	by (1)
$4=n$	by (3).

In this case no cancelling is required.

Second problem:

$3+n=5$	Problem
$///\ \ n=/////$	by (1)
$//\ \ n=////$	by (2)
$/\ \ n=///$	by (2)
$n=//$	by (2)
$n=2$	by (3).

It should be clear from these examples how the algorithm may be applied to solve the other twenty-eight problems in the original set, and moreover, how simply by adding new definitions in memory we may, without changing the algorithm, move on to similar problems involving larger numbers.

From a logical standpoint this algorithm is perhaps as simple as any to be found, and is very close in spirit to a direct characterization of the operation of counting. Consideration of its possible use by children takes us out of the domain of elementary mathematics – the theory of algorithms for simple mathematical systems – into the domain of psychology. Let me try to state some of the problems we encounter as we enter this domain.

(1) It seems highly unlikely that any children, without training,

actually use the algorithm just described. The perplexing question is: what algorithms do they in fact use? At the level at which this problem is often discussed, the obvious answer is that they use the algorithms taught in the classroom and presented in their textbooks. But even casual inspection of the curriculum shows the inadequacy of this response, for algorithms for the thirty problems listed above (or with the numerical variable 'n' replaced by a blank or box) are not explicitly taught, although some partial hints in terms of counting may be given. A typical curriculum instruction to teachers is to let the children find the answer "intuitively" by working with the numbers. Parenthetically, the use of the word "intuition" in its nominal, adjectival or adverbial form by a curriculum builder, reformer, planner or evaluator should be a signal to the psychologist that unexplained and ill-understood learning behavior is about to be mentioned, and, unfortunately, often described as if it were understood.

So the problem remains, how do children in the fourth, fifth or sixth month of the first grade, solve problems like those in our set of thirty?

(2) A proposal often heard is that children solve such problems by simple rote learning. This is a possible response when any single set of twenty or thirty simple problems is considered. It does not seem nearly as plausible when we look at the larger set of problems from which our thirty have been drawn. There are 55 ordered pairs of numbers summing to 9 or less ($0+0=$, $0+1=1$, $1+0=1$, etc.). There are then 165 problems of the same type as our thirty ($n+0=0$, $0+n=0$, $0+0=n$, etc.). And the number of problems is increased considerably further by adding triplets of the form $1+2+n=4$, $1+n+2=4$, etc. It is extremely doubtful that this large stock of problems is held in memory, available for direct access. The child solves them by applying some sort of algorithm. Some of the possibilities are the following.

(a) The child counts off the necessary number words, aloud or in silent speech. Thus, the solution to '$4+5=n$' is obtained by counting off five number names after 'four', namely 'five, six, seven, eight, nine'. The solution to '$4+n=9$' is obtained by counting off number names after 'four' until 'nine' is reached and then judging the cardinality of the set of number names counted off. Even without detailed analysis it is clear that the second kind of problem is harder than the first. The third kind of problem is still harder. The solution to '$n+5=9$' is obtained by counting

off enough number names such that five more take the child to 'nine'. It seems doubtful to me that the algorithm can be successfully applied in this form to the third kind of problem. Notice that no advantage has been taken of the commutativity of addition. Serious training on this property would enable the child to reduce problems of the third kind to those of the second kind. The relatively greater difficulty almost all first-grade children have with the third kind of problem, when the unknown is at the far left, indicates that if the algorithm just described is used, it is not augmented by the commutative law.

For a great many different reasons it seems improbable that the algorithms actually used require very many closely knit steps to obtain an answer. The counting algorithm just described is realistic for problems of the form '$4+5=n$' and not out of the question for problems of the form '$4+n=9$'. For problems of the form '$n+5=9$' the child may, without being explicitly conscious of it, make rough estimates of n and test the guess by counting. He remembers, say, that $5+5=10$, and 'nine' is close to 'ten', so he tries 3 or 4. Or, he may remember, that is, have in immediate storage, that $4+4=8$, and he uses this fact to guess 3, 4 or 5.

(b) In many ways the above discussion sells the counting algorithm short, because of the seeming difficulty of counting a set of number names like 'five, six, seven, eight' pronounced aloud or in silent speech, because the trace of 'five' may have departed before 'eight' is said. When the algorithm is externalized and applied in terms of physical objects (even the fingers) it seems much easier. I have seen something like the following used quite successfully in Ghana with harder problems than those we are now discussing.

The child has a counting set of pebbles on his desk. To solve the problem '$4+5=n$' he first counts out 4 pebbles from his pile. He stops, and then counts out five more. This counting is done by simultaneously saying the number names 'one, two, three, four' and pulling one pebble from the pile as he says each name. After counting out the set of four, and then counting out the set of five, he now counts the separated set of nine pebbles and gets the answer. He solves the problem '$4+n=9$', by first counting out a set of nine pebbles and then taking four away, that is, by counting off a set of four from the set of nine. (It is to be emphasized that each of these counting operations is a highly physical thing.) After taking away the set of four, he then counts the remaining set of five to obtain the

answer. Notice that the act of taking away four from the set of nine pebbles can be clearly and succinctly taught even though the subtraction symbol has not been introduced. As already remarked, lots of people have observed that for American children the '$n+5=9$' sort of problem is harder than the '$4+n=9$' sort. For the counting algorithms just described they would seem to be on an equal footing. I think, but do not have real evidence at hand, that the Ghanaian children have the same sort of relative difficulty. The explanation is most likely to be found in the decoding required to pass from the written problem to the physical execution of the algorithm. The detailed analysis of how the stimulus arrangement expressing the problem sets off the algorithm shall not be gone into here, but I may say in passing that this kind of example provides an excellent opportunity to analyze the behavioral semantics of the simplest sort of language. Briefly put, I interpret a problem format like '$4+n=9$' as a command in the imperative mood. The symbol '9' standing by itself to the right of the equals sign means for the pebble model "Count out a set of nine pebbles". The symbol '4' means "Count out a set of four pebbles from the set of nine". And, roughly speaking, the remaining phrase '$+n$' means "Count the remaining set of pebbles and record the answer". For this kind of semantic the classical notion of truth is replaced by that of a response, or class of responses, satisfying a command. What I have sketched here in the roughest sort of way can be made precise by using with only slight modification the standard methods and concepts of formal semantics.

From the standpoint of the usual way of characterizing algorithms, the pebble-counting algorithm is unusual, for the operations of the algorithm are performed on the pebbles and not on the number symbols themselves. In this case the number symbols have meaning and this meaning is used to give instructions for performing the algorithm. It would seem that it is this sort of algorithm many people now advocate in arithmetic in order to avoid development of great facility with algorithms defined wholly in terms of the number symbols and which may thus be learned without "understanding numbers".

In order to give a concrete sense of some of the complexities that arise in understanding how children learn and perform algorithms, I would like to review briefly two pertinent experiments.

In the first experiment children in the first, second and third grades

(ages 6, 7 and 8 years approximately) were asked to give the correct answers to the 63 problems of the form $1+2=n$, $1+n=3$ and $n+2=3$, with the sums ranging from 0 to 5. The problems were shown on a screen by a slide projector in the form $1+2=$⌊____⌋, $1+$⌊____⌋$=3$, etc., and the subjects responded by pushing one of six buttons marked 0, 1, 2, 3, 4, 5; the buttons were arranged linearly. A timer also measured the response latency from the appearance of a problem on the screen to the pushing of one of the six buttons. In a given daily session a subject was presented with each of the 63 problems for a total of 63 trials. One group of first graders had six sessions; the remaining subjects had three sessions. The only data we shall examine here are those resulting from summing over all grades, days and subjects. This summation yields a total of 280 responses for each of the 63 problems.

In line with the general discussion of possible algorithms, the following simple model is proposed for analyzing the rather complex data of this experiment. The fundamental operation, it is postulated, is that of counting. For problems of the type $a+b=m$, where a and b are given numbers and m is to be found, the time required is $(b+1)\alpha+\delta$.

Here δ is a constant of the sort familiar in reaction-time studies; α is the time it takes to count one step for problems of this type (hereafter called Type I); $b+1$ rather than b steps are called for, because '0' is the first possible answer, '1' the second, etc. In the case of Type II problems, whose form is $a+m=b$, the only change is to replace the timing parameter α by β. Thus the time required to solve $a+m=b$ is $(m+1)\ \beta+\delta$. Note that here m replaces b, because in all cases we think of counting up to the sum. For problems of Type III, that is, problems of the form $m+a=b$, we introduce a third parameter γ, and the time required to solve $m+a=b$ is $(m+1)\ \gamma+\delta$. Also in line with the earlier discussion it is natural to postulate that $\alpha<\beta<\gamma$. Concerning errors, it is also natural to postulate a parameter θ such that for the three types of problems the probability of an error on each counting step is $\theta\alpha$, $\theta\beta$ and $\theta\gamma$, respectively. Thus for n-step problems of Type I the probability of an error is $1-(1-\theta\alpha)^n$, which in first approximation is simply $n\theta\alpha$, because $0<\theta\alpha<1$. (It is assumed for simplicity that the probabilities of an error on the successive steps are statistically independent and that successive errors will not cancel each other out.)

The detailed analysis of this model will not be pursued here. The model

goes badly awry in a number of its detailed predictions, but several qualitative features are well confirmed without requiring statistical estimates of the five parameters α, β, γ, δ and θ. Here are some predictions and the supporting or disconfirming evidence.

(1) For the three problem types, the order of increasing difficulty both in response error and latency is I<II<III. The data are shown in Table III, with the distribution of the three types given for the first 21 problems

TABLE III

Error and latency distributions for three types of problems in rank-order blocks of size 21

	I		II		III	
First 21	11	11	5	6	5	4
Second 21	6	6	8	6	7	9
Third 21	4	4	8	9	9	8

with least errors, the second 21, and the third 21, and corresponding data for the first 21 problems in speed of response, the second 21 and the third 21. For each entry the error data are shown first and the latency data second.

The evidence that Type I problems are easiest is good, both in terms of errors and latencies because they are concentrated in the first 21 problems in both distributions, but the discrimination between Types II and III is subtle and does not strongly favor the hypothesis that $\beta<\gamma$, even though other pedagogical evidence does.

(2) For problems of a given type, the speed and accuracy is greater when the number of counting steps is less. Thus we may begin by looking at matching pairs, such as $3+\lfloor\_\_\_\_\rfloor=5$ and $2+\lfloor\_\_\_\_\rfloor=5$ to see if indeed the first is easier. To be more explicit, let each problem be defined by a triple (x, y, z) of numbers such that $x+y=z$. A matching pair then consists of two problems (x, y, z) and (y, x, z) of the same type, i.e., with the blank in the same spot. For example, $\lfloor\_\_\_\_\rfloor+1=5$ and $\lfloor\_\_\_\_\rfloor+4=5$ form a pair. There are 27 pairs with $x\neq y$. For simplicity, I restrict myself to the latency data. The prediction of the model is that the response will be faster for the member of each pair having the smaller number to find. Thus the response to $\lfloor\_\_\_\_\rfloor+4=5$, which should take

$2\gamma+\delta$ seconds, should be faster than the response to $\square+1=5$, which should take $5\gamma+\delta$ seconds. These predictions for the matching pairs are pretty well borne out by the data. The prediction is true for 20 of the 27 pairs. Four of those for which it is not are problems of the form $\square+0=a$; when these special cases of adding zero are eliminated, the results are even more favorable to the model.

(3) The problems with the smallest error rate and latency are consistent with the model, namely $0+0=\square$, $\square+0=0$ and $0+\square=0$. On the other hand, there are some striking anomalies, hard to explain from nearly any standpoint. On only four problems is the error rate greater than for $1+1=\square$, which was missed 26% of the time! Part of the explanation may be that the subjects responded very fast to this problem – it ranked fourth immediately after the three 'zero' problems just mentioned – and thus made many careless errors. The mean latency for $1+1=\square$ was 3.3 seconds; it was 2.6 seconds for $0+0=\square$ as the minimum of the set of 63 problems and 7.0 seconds for $4+\square=5$ as the maximum. Whatever the explanation, only a quite complicated model seems likely to fit this surprising error rate into the scheme of things. Other aspects of the response errors are not well explained by the model, but shall not be considered here.

The analysis presented has been necessarily very sketchy. A more detailed quantitative assessment will be made elsewhere of the family of models suggested by the simple five-parameter model examined here. The preliminary results seem to be encouraging enough to warrant such investigations in greater depth.

I want now to move on to a second experiment that has some interesting bearing on the complexity of understanding the learning of algorithms. Roughly speaking, the significance of the experiment I want to describe is related to the fact that we probably have been and will continue to be much misled by the mathematical structure of algorithms, so that we are deceived into thinking that young students learn the material very much in the way it is formulated from a mathematical standpoint. As in most areas of complex learning, what is actually going on is undoubtedly a good deal more subtle. The experiment is one performed with 9- and 10-year-old children who had already been given extensive instruction on the commutative, associative and distributive laws of arithmetic. They had had verbal instruction as to the significance

of these laws, and they had performed and executed presumably correctly, but without detailed check on the part of the teacher, a great many exercises applying the laws. Many of the exercises emphasized the fact that the commutative, associative and distributive laws are central to the justification of the more complex algorithms of elementary arithmetic – multiplication of two-digit numbers, the algorithm for long division, etc.

The experiment was conducted as part of the program of the Computer-Based Laboratory we have constructed in the last two years at Stanford; the experiment was performed adjacent to the school classroom in a very small room in which was located a teletype that was connected to the computer at Stanford. The school itself is approximately 20 km south of Stanford. The children participating in this experiment had used the teletype for at least a month for the purposes of review and drill in elementary mathematics, and were fully familiar with the instrument and the experimental setting.

We had noticed in earlier work that the students were having difficulty with the commutative, associative, and distributive laws and that they particularly had difficulties with exercises that called for a rapid shift from one law to another.

We decided to perform a fairly simple learning experiment on the mastery of this material. We broke up the types of problems into 48 categories; shortage of time prohibits me from giving a description of these 48 categories. They depend on which particular law is involved and where the blank occurs. The equations $5+3=3+\square$ and $5+3=\square+5$ would be two instances of two categories exemplifying the commutative law of addition. The subjects got 24 problems a day and they cycled through the entire 48 types every two days. The students were given 10 seconds to answer each problem. If an answer was not given in that time interval, the program returned control of the teletype to the computer, and the teletype printed out "Time is up" before repeating the problem again. After the second appearance and failure to respond correctly or within ten seconds, the correct answer was printed out, the problem was repeated for a final time, and the program moved on to a new problem.

The results for the first six days of the experiment are shown in Table IV.

For the first day the mean proportion of correct responses was 0.53, the

TABLE IV

Mean learning data for the first six days of the computer-based teletype experiment on the laws of arithmetic

Day	Prob. correct	Prob. wrong	Prob. time-out	Mean time in sec.
1	0.53	0.22	0.25	630
2	0.56	0.33	0.11	520
3	0.74	0.21	0.05	323
4	0.72	0.23	0.05	390
5	0.77	0.18	0.05	355
6	0.91	0.08	0.01	279

mean proportion of errors 0.22, the mean proportion of time-outs 0.25 and the mean completion time 630 seconds for the entire set of 24 problems. So the students began by being rather slow and by making lots of errors. The second day we see an increase, and an increase each day thereafter, except on the fourth day, until on the sixth day of the experiment the mean proportion correct is 0.91, the number of time-outs is very slight and there is a very considerable reduction in mean completion time from 630 seconds to 279 seconds. From the standpoint of learning we get very clear mean results. If I had the time I would show you the results for the the best student and the worst student in the class on the first day. Both of them showed considerable learning and one of the pleasing things about the experiment is that every student showed marked improvement in performance. Now one next question to ask is about how to analyse the difficulty of items. At the moment it appears that the best way is not in terms of the mathematical law involved, for example, the distributive law, but in terms of the kind of patterns required in the answer. These patterns can be defined fairly directly in a psychological rather than in a mathematical fashion. For instance, regardless of whether we are concerned with an associative, commutative or distributive law, if the student must fill in a single blank on the right of the equation, using a stimulus pattern or a numeral already occurring on the left, then the problem is relatively easy. What appears to be the case psychologically is that the students are perceiving the kind of pattern required without regard to the mathematical meaning of the law involved. Data for this sort of problem are labeled 'Type A' in Table V. Of next order of difficulty

TABLE V

Learning data on the three types of problems in the computer-based teletype experiment

Days	Type A			Type B			Type C		
	Prob. correct	Prob. wrong	Prob. time-out	Prob. correct	Prob. wrong	Prob. time-out	Prob. correct	Prob. wrong	Prob. time-out
1–2	0.63	0.21	0.16	0.37	0.35	0.28	0.06	0.80	0.14
3–4	0.89	0.09	0.02	0.62	0.36	0.02	0.26	0.53	0.21
5–6	0.93	0.06	0.01	0.75	0.24	0.01	0.23	0.55	0.22

are the problems requiring that two blanks be filled in on the left, but using numerals that occur on the right; for example,

$$(5 \times \boxed{\quad}) + (5 \times \boxed{\quad}) = 5 \times (6 + 3).$$

Data for this kind of problem are labeled 'Type B' in Table V.

Finally, most difficult are the problems that require the use of a number not shown on the right-hand side of the equation, which in the present experiment were essentially examples showing that neither subtraction nor division is commutative:

$$7 - \boxed{\quad} = 12 - 7.$$

Data for this sort of problem are labeled 'Type C' in Table V.

Admittedly a deeper sort of theory is required to explain the data of Table V than that exemplified in the simple five-parameter model discussed earlier. On the other hand, this experiment as well as the earlier one should make evident that a theory of how mathematics is learned and mathematical concepts are formed will not fall out of the consideration in any direct or simple way of the structure of mathematics itself as it is usually thought of by mathematicians. A new and rather subtle psychological theory of learning is clearly necessary. The present paper has merely tried to delineate what may be important aspects of such a new theory.

22. ON THE THEORY OF COGNITIVE PROCESSES*

I. THE PROBLEM

Within a number of disciplines ranging from psychology to computer science, the subject of cognitive processes is now being given a great deal of attention and discussion. Many of the problems and viewpoints expressed about cognitive processes have a very old ring in philosophy, and clearly have a very respectable ancestry in the history of philosophy. It is not too difficult to cite a number of definitions of what cognitive processes are taken to be. Even the man in the street may often have a fairly ready answer, namely, that talk about cognitive processes is just a fancy way of talking about thinking. This man-in-the-street answer is not a bad one. It makes quite clear that the central problem of a theory of cognitive processes must be to develop the theory of how we think. Philosophers and psychologists have had almost as much to say about thinking as they have about perception, and consequently, it might be assumed that we can rapidly move to consideration of current systematic theories of thinking, and consider their various strengths and weaknesses.

Unfortunately, however, this is too sanguine a description of the actual state of affairs. What has been written about thinking, either by philosophers or psychologists, scarcely qualifies as systematic theory in the way in which we think of mechanics, thermodynamics or quantum chemistry as being such developed theories. On the other hand, one current direction of interest does enable us to give a very clear and concrete description of the kind of thing any adequate theory of thinking or cognitive processes must be able to accomplish. It is simply this. Any adequate theory must, in principle, be sufficiently detailed and categorical to enable us to use it to construct a machine that can think and possibly perceive. In using a thinking machine as a central desideratum of any theory, I do not intend to enter into controversies about whether or not

* This paper has not been previously published. It was given as an Arnold Isenberg Memorial Lecture at Michigan State University, December 9, 1966.

machines can think. A good deal of what has been said in this controversy seems rather silly. In any case, I shall talk here about thinking machines; but for those who are unhappy with this phrase, I am willing to replace the general phrase by a list of problems that I expect the machine to be able to handle. This list of problems includes the learning of elementary mathematics, and the development of the abilities to play various competitive games ranging from chess to bridge, and to conduct a dialogue with a human being over a fairly wide range of subject matters. Another requirement is that a human being find it difficult to distinguish a dialogue conducted with the machine from a dialogue conducted with another human being. A machine that can do these things I am quite willing to baptize as a *thinking machine*. The extent to which a classical theory of cognitive processes provides a basis for the construction of such a machine is, in my own judgment, perhaps the most single useful criterion for evaluating the depth and significance of the theory.

Before we turn to any particular theories, a remark about the relation between thinking and perceiving is in order. For brevity, I shall refer to the *thinking-machine criterion*, but I see no objective way of separating thinking and perceiving, and consequently, I intend that perceiving be implied by the concept of thinking, at least in the present context.

II. SURVEY OF SOME THEORIES

Almost every major philosopher from Plato onward can be viewed as offering a theory of cognitive processes, and it is not difficult for anyone trained in philosophy to isolate major defects of any of the theories that have been historically important. Using the test proposed in this paper, it is quite clear that none of the major philosophers has provided a theory that is sufficiently deep and detailed to enable us to draw up the blueprints for a thinking machine.

Before examining this question in somewhat more detail, there is another way of putting the matter that will perhaps be more to the taste of those intrinsically opposed to machine talk. A very clear contrast can be drawn between the standards of evaluation traditionally used in dealing with epistemological theories and the standards that have been applied in this century in dealing with proposals about the foundations of mathematics. In the foundations of mathematics there is a recognized

body of hard fact, so to speak, that any approach to foundations must come to terms with. A serious philosophical approach to mathematics must offer a basis rich and exact enough to derive in non-subjective fashion the main results of classical analysis, including at the very least the elementary parts of the differential and integral calculus and the theory of differential equations. This criterion of being able to derive at least the major part of classical analysis introduces a note of realism and hard-headedness into discussions about the foundations of mathematics that is often lacking in epistemology. In evaluating different approaches to the foundations of mathematics, it is not required that they all yield precisely the same results in classical analysis. The account given by Zermelo-Fraenkel set theory is certainly different from that given by intuitionistic mathematics or recursive analysis, but in every case there is a large body of agreement about central results, and it is the responsibility of any proposed theory to give an account of these central results. The concept of a thinking machine is introduced to provide a similar criterion for evaluating theories of cognitive processes. If models – and here I mean semantical models – of a proposed theory cannot serve as the prototype of a thinking machine, and if it can be demonstrated that the models of the theory are not nearly adequate to account for some of the major cognitive phenomena mentioned already, then the theory should receive little attention as a fundamental account of cognitive processes.

In order to get a sense of what the thinking-machine criterion implies when it is used to evaluate a theory of cognitive processes, we may look at several examples of such theories. A very good place to begin is with Hume. The main features of his theory are to be found in Book I, Part I of *A Treatise of Human Nature* (references are to the Selby-Bigge edition).

Hume begins with his famous division of perceptions of the mind into two distinct kinds, namely, impressions and ideas. "The difference betwixt these consist in the degrees of force and liveliness in which they strike upon the mind, ..." He goes on to distinguish between simple and complex ideas, and insists upon the fundamental proposition that all simple ideas are initially derived from simple impressions. For our purposes, the next important distinction is made in terms of how ideas can reappear in the mind. The two ways in which they can reappear depend upon the two faculties of memory and imagination; and again, it

is the degree of vivacity that distinguishes the ideas of memory from those of imagination.

Next, we come to the famous account of how complex ideas are built up from simple ones. Here Hume introduces the three principles for connecting or associating ideas: resemblance, contiguity, and cause and effect. Hume then divides complex ideas into relations, modes and substances. For the purposes of building a thinking machine, what he has to say about relations is particularly pertinent and also terribly puzzling. He asserts without any very substantial argument that all philosophical relations, as he terms them, can be organized under seven general headings: resemblance, identity, space and time, quantity or number, quality, contrariety, and, again, cause and effect.

His familiar position on substance is well conveyed in this sentence. "The idea of a substance as well as that of a mode, is nothing but a collection of simple ideas, that are united by the imagination, and have a particular name assigned to them, by which we are able to recall, either to ourselves or others, that collection".

The final section of Part I is devoted to abstract ideas and the defense of Berkeley's thesis that there are no general ideas, but only particular ones. As Hume puts it, "All general ideas are nothing but particular ones, annexed to a certain term, which gives them a more extensive signification, and makes them recall upon occasion other individuals, which are similar to them". The doctrine of Berkeley and Hume that abstract ideas are particular in character is surely an essential step forward in constructing a thinking machine. The rationalistic doctrine of general ideas that they are attacking seems difficult to develop concretely and constructively.

Because the process of abstraction in some form is essential for any theory of cognitive processes, it is desirable to take a closer look at what Hume has to say about abstract ideas. A first major point in his argument that ideas must be particular and not general is that the mind cannot form a concept of quantity or quality without forming a precise notion of the degree of each. For example, according to Hume it is not possible to have a general idea of redness, as opposed to a particular idea of redness with a given hue, saturation, and so forth. Secondly, he emphasizes that abstract ideas must be individual in themselves, although "they may become general in their representation". What he emphasizes is that the image in the mind is only that of a particular object "though

the application of it in our reasoning be the same, as if it were universal". He continues, "This application of ideas beyond their nature proceeds from our collecting all their possible degrees of quantity and quality in such an imperfect manner as may serve the purposes of life... When we have found a resemblance among several objects, that often occur to us, we apply the same name to all of them, whatever differences we may observe in the degrees of their quantity and quality, and whatever other differences may appear among them. After we have acquired a custom of this kind, the hearing of that name revives the idea of one of these objects, and makes the imagination conceive it with all its particular circumstances and proportions" [p. 20].

All this is very sound and promising, and very modern in its ring. But how this custom or convention of associating many ideas under a general term takes place Hume does not really try to explain. That is, he does not really offer a mechanism for generating the abstractions. His most succinct description of what happens is in the following passage.

> For this is one of the most extraordinary circumstances in the present affair, that after the mind has produc'd an individual idea, upon which we reason, the attendant custom reviv'd by the general or abstract term, readily suggests any other individual, if by chance we form any reasoning, that agrees not with it. Thus shou'd we mention the word, triangle, and form the idea of a particular equilateral one to correspond to it, and shou'd we afterwards assert, *that the three angles of a triangle are equal to each other*, the other individuals of a scalenum and isoceles, which we overlook'd at first, immediately crowd in upon us, and make us perceive the falsehood of this proposition, tho' it be true with relation to that idea, which we had form'd [p. 21].

In a passage occurring on the next page Hume indicates that he thinks a deeper explanation is not possible. He says, "To explain the ultimate causes of our mental actions is impossible. 'Tis sufficient, if we can give any satisfactory account of them from experience and analogy".

Hume has other wise and important things to say about the process of abstraction, but the main tenets of his theory have been covered in the aspects discussed thus far. If without even asking questions of empirical correctness, we apply the thinking-machine criterion, I am sure it is clear to everyone that Hume's theory does not meet the criterion. He has not provided a sufficiently definite and deeply enough structured theory to enable us to construct a model, even in principle, which will be able to think, or, in fact, to solve even the simplest sort of problems. The real truth is that on the basis of the analysis Hume gives us we would not be

able to get off the ground in attempting to construct a thinking machine.

This point becomes even more obvious when we attempt to get a closer view of Hume's theory. At every turn we are unable to understand in any clear objective sense precisely what he intends to mean. We are not able to ground the discussion in any unequivocal empirical or physical concepts that could be used as a basis for beginning the construction of our thinking machine. Drawing an analogy from the foundations of mathematics, we might think, at first glance, that simple ideas should play the role of sets in classical foundations; but as we examine the matter more closely, we find that no structural assumptions about simple ideas, no enumerations of other properties, nor even any systematic listing of what are simple ideas, is given. Secondly, if we turn to the processes by which complex ideas are supposed to be constructed out of simple ideas, we find that only contiguity among the three principles of construction has anything like a very definite meaning. Again, when we turn to the seven general categories of relations, it is not clear how we build these relations into the potential structure of our machine. In all this discussion, a key role is played by the concept of resemblance; and it is precisely this concept that is perhaps the most difficult to come to grips with. To take a concrete example, suppose that our machine has a simple seeing eye consisting of a matrix of points, one thousand by one thousand with each point in the matrix being either light or dark, depending upon whether a simple line figure that we are attempting to recognize crosses a point. What does it mean to say that two figures resemble one another? No obvious criterion of an applicable operational sort comes to mind. It is possible, perhaps, to take the attitude that we cannot explicitly define what we mean by resemblance, but can state a large number of systematic properties. However, no such list of systematic properties is given, and matters are left in a totally indefinite state. No doubt this same vagueness about resemblance is at the heart of the difficulty already expressed about Hume's analysis of abstraction. The process by which individual ideas are gathered together under a general term depends upon resemblance; and without a clear account of resemblance we cannot begin to give an account of abstraction. What we are left with is a tantalizing beginning, one which makes important distinctions such as those between impressions and ideas, simple and complex ideas, memory and imagination, but

which is only a bare and crude sketch of what must be a very complicated mechanism.

It might be thought that contemporary philosophy or psychology can improve on Hume's theory in many different ways, and offer a very much more substantial theoretical analysis of cognitive processes. As I indicated at the beginning, I do not think this is the case, and it is important to say why by referring at least to typical representatives of contemporary theory.

In the case of philosophy, we have in this century a tradition of constructive theory about epistemological problems running from Whitehead and Russell to Carnap, Goodman and other philosophers. The focus of this body of work is not constant and does not necessarily always include the topic central to this paper, but there is a common core of problems and concerns that runs straight from Berkeley, Hume and Kant to Whitehead, Russell, Carnap and Goodman. In spite of the quite different aims, for example, of Hume and Whitehead, a significant overlap may be found in Hume's discussion of abstract ideas and Whitehead's development of his method of extensive abstraction. But for the purposes of the present paper, Whitehead is a bad example, because his concern for what we might term the psychology of thinking and perception is minimal. He is not really interested in giving an account of realistic mechanisms of thinking and perceiving. Rather, he attempts to construct the main lines of contemporary physical science from some basic assumptions and ideas about perception that can scarcely be regarded as a faithful account of how the human organism operates, or how a thinking machine could possibly be made to operate. (In my own judgment, the sober task of having to construct a thinking and perceiving machine might have had a most salutary effect upon Whitehead's ruminations – even the earlier ones embodied in *The Principles of Natural Knowledge* and *The Concept of Nature*.)

Probably the most ambitious recent effort in constructive epistemology is Goodman (1951). The clarity and precision of Goodman's analysis is incomparably greater than that of Hume's; but much is lost in the process, for it is undeniable that major aspects and characteristics of human knowing and thinking that are treated by Hume are omitted by Goodman. Three examples are the synthesis of complex ideas out of simple ideas, the structural character of memory, and the simulation of the real world in imagination. Goodman does not consider the more

active parts of the human apparatus for thinking and perceiving, but concentrates instead on relatively simple aspects of perception of the sort one associates with Locke's *tabula rasa.* In making this remark about Goodman's work, I do not mean to say that he claims to have given a complete account of thinking and perceiving, but has not done so. He quite correctly and modestly appraises what he has accomplished, and certainly one of the virtues of what he has done is that it may be evaluated as a specific set of results, in a way that is not possible for much of the philosophical analysis in this area. Just to give a sense of how far we are from an adequate solution, even within the domain covered by Goodman (1951, Parts II and III), we may note the considerable body of psychological evidence that an adequate account of the perception of size and shape must enter at least into the kinematics of eye and body movement, and also the complexities of binocular vision. In fact, once motion or binocular vision is considered, there seems to be an excellent basis for challenging the entire concept of the visual field that is the basis of much of Goodman's analysis of visual perception. Yet, it is all too easy to criticize Goodman – he does it himself in terms of outlining what he has not yet accomplished; it is not nearly as easy to proceed to the constructive task of improving on his results. I do think that a mistaken aspect of Goodman's approach is to work within a nominalistic framework and not to use all possible devices of modern mathematics and logic as tools of analysis in trying to get a grip on the main problems. Once the intricacies of motion are introduced, the whole ritual of working in a nominalistic fashion seems futile. The full apparatus of mathematical analysis is required even to analyze quite simple problems – for example, the perceived invariance of a static world as seen by a moving Cyclops.

By turning from philosophy to psychology, we might hope to find in recent years a larger body of constructive work and a closer approximation to the set of blueprints we need for our thinking machine. But first there is one point that needs to be cleared up. It is very much to be emphasized that I do not have in mind blueprints that would tell, for example, how DNA molecules are used to code information from the outside world and to synthesize concepts. Rather, the blueprints should be at the level of logical function and logical organization. They need not include a physiological specification of how human organisms work at the molec-

ular level. Ultimately, of course, it is a desirable and essential scientific goal to achieve this molecular understanding, but there is every reason to think that we are a very great distance from it. We still have much important work to do in solving the logical problem of how a thinking machine is best organized to perform characteristically human tasks in cognition. Neither psychologists nor philosophers are prepared to offer a physiological analysis of the actual structural mechanisms used in the human brain and nervous system; and physiologists, on the other hand, are certainly not prepared to make the inference from what they know about the behavior of neurons or larger parts of the nervous system to major functional characteristics of the system. Moreover, the task of analyzing the logical character of human thinking and perceiving is not only in a classical philosophical tradition, but also has current relevance for the attempts actually to build thinking machines by using digital computers.

If we begin with the mathematical models of learning processes which have been most thoroughly explored in the past decade, it is easy enough to show that they are scarcely adequate to provide structural mechanisms that will enable us to draw the blueprints of our thinking machine. For both the linear model and the small-element stimulus-sampling Markov models, I have tried to make out the case in some detail elsewhere (Suppes, 1964) and will not repeat the arguments. In the last several years, more complex models of learning have begun to appear. It will perhaps be useful to consider as a typical example one that introduces mechanisms of memory as well as mechanisms of conditioning in order to give a deeper running account of learning processes. A well worked-out theory is to be found in the recent report by Atkinson and Shiffrin (1968). They propose a two-process model for memory with the following features. The memory system has two central components, a short-term memory buffer and a long-term store. Experiments for which the model was designed are those in which a series of items is presented to the subject for subsequent recall. Familiar examples would be digit-span studies, paired-associates learning, or second-language experiments on vocabulary learning. Each stimulus-item presented to the subject is postulated to enter the short-term buffer which has the characteristics of a pushdown store. The term *pushdown* comes from the fact that when an item enters the store it enters at the top and works its way down to the first position

as new items are entered. The short-term buffer is postulated to have a fixed finite size. Once the buffer is filled, each time a new item is entered, an old one is displaced. But this displacement need not consist simply of the earliest item being displaced by the latest item. More complicated probabilistic arrangements are postulated in order to give a more realistic account of actual experimental data. A simple one-parameter assumption is to postulate that the oldest item in slot one is dropped with probability δ; if that item is not dropped, then the item in position 2 is dropped with probability δ, and so forth. We thus obtain an approximately geometrical distribution for the probability of any item being dropped.

Concerning the long-term store, it is postulated by Atkinson and Shiffrin (1968) that while items are held in the short-term buffer, information about them is transferred to the long-term store. This information will not in all cases be sufficient to allow recall of the item, and even when the information is sufficient, the subject may not be able to recall the item because his search of the long-term store is unsuccessful. A probabilistic mechanism for search of the long-term store is introduced; it has the property that the greater the number of items in long-term store, the smaller the probability that any particular one will be retrieved. The assumptions that have been described qualitatively and rather loosely here are worked out in detail by Atkinson and Shiffrin (1968), and applied successfully to the prediction of data from several experiments. On the other hand, it is also clear that the model of Atkinson and Shiffrin (1968) is still too simple to account for any of the major features that Hume was striving to deal with. There are no mechanisms for making comparisons or judging resemblances. The information about a stimulus item that is put in long-term store is measured in unanalyzed fashion by a single numerical parameter. We could scarcely use this parameter in any direct way to aid us in drawing up the blueprints of our thinking machine. In fact, the properties of the memory already used in current computers are at least as complex as those postulated by Atkinson and Shiffrin (1968), but the complicated problems of cross-referencing and structuring to provide judgments of resemblance and synthesis of new concepts or complex ideas are scarcely touched. The direction of work exemplified by the paper of Atkinson and Shiffrin and related papers by others is promising and will lead to useful results; but it is easy to underestimate

the distance that still lies between the simplicity of these models and the complexity implied by our thinking-machine criterion.

When we survey the problems of cognition from Hume onward and examine at the same time processes by which we are able to build up mathematics on a conceptual basis, the overwhelming importance of processes of abstraction becomes apparent. As the discussions of Berkeley and Hume indicate, the importance of abstraction has been fully appreciated for a long time. It is perhaps unfortunate that it has not received more attention in psychology. The psychological studies of discrimination, generalization and transfer have not as yet provided a basis upon which to build a theory of abstraction. Indeed, they have scarcely yet been adequate to build a theory of the processes they initially consider. It is from certain standpoints a conceptual miracle that the whole foundations of mathematics can be constructed on the empty set just by the appropriate cascading of sets. That we can extract mathematics by a process of abstraction from entities no more complicated than the empty set is a real intellectual surprise.

There has been one direction of research in psychology aimed at constructing theories that mirror to some extent the processes of abstraction reflected in the foundations of mathematics. I refer to theories of mediation; it will be useful to take a cursory look at what they have accomplished.

The paradigm of mediation is of the following sort. The subject learns to associate item B to item A. He then learns to associate item C to item B. He is now presented item A; the prediction of mediation theory is that he will learn more rapidly to associate C to A on the basis of B as a mediator. By picking arbitrary items A, B and C, we attempt to study the pure process of mediation apart from any cognitive structuring of the stimuli. Apart from specific experimental work, general supporting evidence for the process of mediation is characterized in terms of the obvious facilitating effect of verbalizing concepts and ideas. On the surface, the positive effects of mediation in organizing thought seem to be evident, and are closely related to some of the things Hume has to say about abstract ideas. A typical problem for precisely stated models of mediation is to predict the relative efficacy of the eight different paradigms of mediation that can be based upon the precise location of the mediating stimulus B and the stimuli A and C. For example, we can begin with the

stimulus *B*, elicit stimulus *A* as a response in the first training, then present stimulus *C* and elicit stimulus *B* as a response. Finally, we test as before the extent to which the mediation of *B* affects positively the elicitation of *C* when *A* is presented. It is easy to compute that there are eight possible such paradigms given that the test is always in the form *A-C*, and that *B* is always the mediating item. Experimental studies of the eight paradigms have been reported by Peterson *et al.* (1964), and others. A mathematically formulated model of the sort needed to make predictions of relative facilitation is given in Crothers and Suppes (1967); their model is closely allied to the one of Atkinson and Shiffrin already described, although the models were developed independently.

For some purposes, the stripping away of all cognitive structure in studies of mediation seems appropriate in asking what are the pure effects of the arrangement of stimuli in obtaining mediational effects. On the other hand, from the standpoint of the present paper, the stripping away of all structure means that the mediational models that are developed, and the experimental results that are obtained, have little bearing on the more complex processes required to outline the structure of a thinking machine. The central concept of mediation theory, which in many cases we would think of as verbalization of concepts, certainly seems in the right direction. It also seems to be a way of catching in psychological terms the important and powerful methods of abstraction exemplified in axiomatic set theory, as applied to building up the structure of mathematics. Unfortunately, mediation theory has not yet been successful in postulating the additional aspects of structure central to erecting a substantial theory. The fundamental problems faced by Hume in his discussion of abstract ideas have not been solved in current attempts at mediation theory. The method of subsuming particular ideas under general terms, which Hume found extraordinary and essentially unexplainable, is essentially as much a mystery for current mediation theories. Once again, what seem to be missing are the concrete principles of resemblance and concept synthesis.

Many cognitive psychologists would judge that the examples of theory I have selected from psychology or philosophy all reflect the stimulus-response tradition of which Hume is one of the first creators. There do exist cognitive alternatives in psychology, running from gestalt psychology to the theories of Piaget and Bruner. In many respects, the theoretical stance and experimental work of these cognitive psychologists have served

as an excellent corrective to excessive concentration by stimulus-response psychologists on a few paradigm experiments and a few overly simplified issues. On the other hand, if we look to the cognitive psychologists for the blueprints of our thinking machine, we will find that it is about as hopeless as looking for architectural details of a house from a sketch of an impressionist like Renoir. This lack of concrete detail and well worked-out theory is characteristic.

There is, however, one new direction that is considerably more promising and should be looked at in a different light. This is the approach to cognitive problems through information-processing languages developed by Allen Newell and Herbert Simon. Although in principle the use of information-processing languages and programs is independent of the existence of digital computers, in practice this theoretical framework has only come to the fore as it has been applied to the programming of computers. Specific tests have been made of the extent to which the programs written do imitate, or to use the favorite current word, *simulate*, human behavior in specified circumstances. Among the circumstances that have been studied rather thoroughly are relatively simple experiments on paired-associates learning or concept identification.

It is not possible here to examine in detail the accomplishments and limitations of what has been done thus far with information-processing programs. It is fair to say that the kind of problems framed initially for our thinking machine have not yet been satisfactorily solved by this rather direct approach to the construction of a thinking machine. Without doubt, part of the problem has been that too many aspects of human thinking are not caught in the structure of information-processing languages. Their structure is relatively close to ordinary English, or at least does not provide a framework for what must be essential aspects of any thinking device with the approximate power of the human mind.

A convergence of the programming or simulation approach and the mathematical-models approach seems likely to occur in the future. There is some reason to think we are on the edge of rapid developments. Any theory that is successful would seem to need functional and structural concentration in at least four areas: problems of concept formation and learning; problems of perception, particularly the geometry of perception; problems of information processing in the Bayesian sense; and problems of organization for accessing and cross-referencing long-term memory.

Rather than attempt to survey all of these problem areas, I would like to conclude with some relatively detailed remarks about two special problems: the logic of belief as reflected in the organization of belief structures; and the problems of mechanisms for changing belief.

III. LOGIC OF BELIEF

In illustrating how difficult it is to understand how belief structures are organized in the human memory, it will be possible to work from some simple examples. Think of writing down memories of all your personal experiences beginning with those of earliest childhood. As a rough approximation, let us suppose that this is done in a very thorough fashion, and you are able to fill something about the size of a complete set of *Encyclopedia Britannica.* I do not mean to suggest this would be as much information as you could provide, but it will do to illustrate the problems of indexing and accessing. We now ask some simple questions of you, on the one hand, and of the written record, on the other. Have you ever been to Calcutta? Have you been in an airplane at an altitude greater than 50000? Have you been married at least twice? Do you have two sisters? Are your grandparents now living? Did you do well in high-school geometry? Did you like the teacher of English you had as a senior in high school? Can you speak Chinese? What I find fantastic about these questions is that any individual in the audience can answer them almost instantaneously, whereas almost any procedure for organizing the written record will make it laborious to find the answers to at least some of the questions, even with a large computer available. One kind of problem is cross-indexing of subject matter in what is apparently a wide variety of ways. The second problem is the central one of trying to become clear about the logical or semantical form in which beliefs are stored. Certainly it seems clear that all of them are not stored in terms of a verbal symbolic representation. Just about all of us can introspectively testify to beliefs or memories that are represented by visual or auditory images. Even if we restrict ourselves to beliefs that are stored in terms of a symbolic representation, it is by no means clear how this storage takes place. For example, can a case really be made out that symbolic beliefs are stored in the form of kernel sentences?

In this connection, we encounter a problem of rationality that has been

little discussed either in philosophy or psychology. We are all aware that beliefs or knowledge we have in detail at a given time will gradually decay and disappear if they are not reviewed and used on some sort of intermittent schedule. If we grant for simplicity of the present discussion that review is the main device for maintaining knowledge stored in memory, we can see that the reflective man needs to decide what is most rational for him to remember, and what degree of detail is appropriate. Experience left to take its own course will dictate an answer; but just as in the case of other analyses of rational behavior, it would be interesting to have some clear concepts of rational maintenance of belief to compare to what occurs in actual practice. The problem arises in a direct way in using computers, because storage devices are finite in size; thus we have the option of simply throwing away large bodies of material, or putting it in relatively inaccessible storage. Here the problem of rationality can be given a quite clear formulation in some general decision framework. What I want to emphasize is that a similar sort of theory can and should be developed for the maintenance of human knowledge in the mind of a given individual.

IV. MECHANISMS FOR CHANGING BELIEFS

Traditionally, there are three important ways of changing beliefs. One is to learn a new fact; the second is to make a new inference from known facts; and the third is to discover a new concept and possibly a new law. The scholastic ring of my description of these three methods is testimony enough to how far removed they are from the details of what actually goes on when belief structures are changed in the mind of a given person. Two central mechanisms that have not received sufficient notice are the mechanism of information selection and the assignment of a measure of significance to the information that is selected. The peripheral nervous system of an organism is under constant bombardment by a variety of forms of energy impinging upon receptors. In addition, the organism is continually making conscious or unconscious decisions about how to scan and probe the surrounding environment. So very little of the potential information impinging on the organism is actually selected for attention that very strong selective principles must be at work. There have been few attempts at characterization of a rational way of handling these

mechanisms of attention, although they have been studied in other settings. For example, they have been discussed in connection with the design of search procedures and the organization of artificial perceiving devices. It is often not only rational, but necessary to respond to many immediate stimuli simply in order to avoid immediate harm. One parody of the philosopher is as the absentminded thinker strolling about in a state of concentration that ignores all immediate stimulus input; but this is a parody. In our modern urban society, a philosopher who walked about oblivious of his surroundings would be considered foolish rather than wise, and, in any case, would not be long for this world. On the other hand, it continues to be a part of both the popular and serious conception of rationality that deliberate and considered response to many stimuli and situations is appropriate.

The most thoroughly worked-out theory for analyzing the structure of beliefs and the application of beliefs to behavior is the theory of expected utility, as developed by Ramsey, de Finetti, Savage, and others. It will be worth examining in conclusion how adequately this theory provides a mechanism for changing beliefs. Even if we ignore the fundamental problems of attention and selection, there are at least three main defects in the theory.

Recall that beliefs are expressed in this theory in terms of an algebra of possible events and a probability measure P on this algebra.

My first point is that the probability measure P, effectively expressing my beliefs at time t, cannot be used to express what I actually observe immediately after time t; for P is already "used up", so to speak, in expressing the a priori probability of each possible event that might occur. Thus it cannot be used to express the unconditional occurrence of that which in fact did happen at time t. Pragmatically, the situation is clear. If an event A occurs and is noticed, the individual then changes his belief pattern from the measure P to the conditional measure P_A. What has not been adequately commented upon in discussions of these matters is that the probability measure P held at time t cannot be used to express what actually happened immediately after t, but only to express, at the most, how P would change *if* so and so did happen.

My second point is that the theory is not adequate to the many cases in which changes in belief may be expressed simply as changes in probability, but not explicitly in terms of changes in conditional probability,

because the changes in probability are not completely analyzable in terms of the explicitly noticed occurrence of events. For example, the probability that I assign to the event of rain tomorrow will change from morning to afternoon, even though I am not able to express in explicit form the evidence I have used in coming to this change.

My third point is that the introduction of a new concept will ordinarily require a change in the algebra of events and not just in the probability measure *P*. Scientific examples of this kind of change are easily given.

In my view, any interesting concept of human rationality must be tied to an adequate theory of human powers and limitations of cognition. To move beyond the relatively simple scheme offered by Ramsey and de Finetti, and meet the three criticisms just stated, we must develop a deeper and more detailed theory of cognitive processes than has yet been formulated.

23. STIMULUS-RESPONSE THEORY OF FINITE AUTOMATA*

I. INTRODUCTION

Ever since the appearance of Chomsky's famous review (1959) of Skinner's *Verbal Behavior* (1957), linguists have conducted an effective and active campaign against the empirical or conceptual adequacy of any learning theory whose basic concepts are those of stimulus and response, and whose basic processes are stimulus conditioning and stimulus sampling.

Because variants of stimulus-response theory had dominated much of experimental psychology in the two decades prior to the middle fifties, there is no doubt that the attack of the linguists has had a salutary effect in disturbing the theoretical complacency of many psychologists. Indeed, it has posed for all psychologists interested in systematic theory a number of difficult and embarrassing questions about language learning and language behavior in general. However, in the flush of their initial victories, many linguists have made extravagant claims and drawn sweeping, but unsupported conclusions about the inadequacy of stimulus-response theories to handle any central aspects of language behavior. I say 'extravagant' and 'unsupported' for this reason. The claims and conclusions are supported neither by careful mathematical argument to show that in principle a conceptual inadequacy is to be found in all standard stimulus-response theories, nor by systematic presentation of empirical evidence to show that the basic assumptions of these theories are empirically false. To cite two recent books of some importance, neither theorems nor data are to be found in Chomsky (1965) or Katz and Postal (1964), but rather one can find many useful examples of linguistic analysis, many interesting and insightful remarks about language behavior, and many incompletely worked out arguments about theories of language learning.

The central aim of the present paper and its projected successors is to prove in detail that stimulus-response theory, or at least a mathematically

* Reprinted from *Journal of Mathematical Psychology* **6** (1969), 327–355.

precise version, can indeed give an account of the learning of many phrase-structure grammars. I hope that there will be no misunderstanding about the claims I am making. The mathematical definitions and theorems given here are entirely subservient to the conceptual task of showing that the basic ideas of stimulus-response theory are rich enough to generate in a natural way the learning of many phrase-structure grammars. I am not claiming that the mathematical constructions in this paper correspond in any exact way to children's actual learning of their first language or to the learning of a second language at a later stage. A number of fundamental empirical questions are generated by the formal developments in this paper, but none of the relevant investigations have yet been carried out. Some suggestions for experiments are mentioned below. I have been concerned to show that linguists are quite mistaken in their claims that even *in principle*, apart from any questions of empirical evidence, it is not possible for conditioning theory to give an account of any essential parts of language learning. The main results in this paper, and its sequel dealing with general context-free languages, show that this linguistic claim is false. The specific constructions given here show that linguistic objections to the processes of stimulus conditioning and sampling as being unable in principle to explain any central aspects of learning a grammar must be reformulated in less sweeping generality.

The mathematical formulation and proof of the main results presented here require the development of a certain amount of formal machinery. In order not to obscure the main ideas, it seems desirable to describe in a preliminary and intuitive fashion the character of the results.

The central idea is quite simple – it is to show how by applying accepted principles of conditioning an organism may theoretically be taught by an appropriate reinforcement schedule to respond as a finite automaton. An automaton is defined as a device with a finite number of internal states. When it is presented with one of a finite number of letters from an alphabet, as a function of this letter of the alphabet and its current internal state, it moves to another one of its internal states. (A more precise mathematical formulation is given below.) In order to show that an organism obeying general laws of stimulus conditioning and sampling can be conditioned to become an automaton, it is necessary first of all to interpret within the usual run of psychological concepts, the notion of a letter of an alphabet and the notion of an internal state. In my own

thinking about these matters, I was first misled by the perhaps natural attempt to identify the internal state of the automaton with the state of conditioning of the organism. This idea, however, turned out to be clearly wrong. In the first place, the various possible states of conditioning of the organism correspond to various possible automata that the organism can be conditioned to become. Roughly speaking, to each state of conditioning there corresponds a different automaton. Probably the next most natural idea is to look at a given conditioning state and use the conditioning of individual stimuli to represent the internal states of the automaton. In very restricted cases this correspondence will work, but in general it will not, for reasons that will become clear later. The correspondence that turns out to work is the following: the internal states of the automaton are identified with the responses of the organism. There is no doubt that this "surface" behavioral identification will make many linguists concerned with deep structures (and other deep, abstract ideas) uneasy, but fortunately it is an identification already suggested in the literature of automata theory by E. F. Moore and others. The suggestion was originally made to simplify the formal characterization of automata by postulating a one-one relation between internal states of the machine and outputs of the machine. From a formal standpoint this means that the two separate concepts of internal state and output can be welded into the single concept of internal state and, for our purposes, the internal states can be identified with responses of the organism.

The correspondence to be made between letters of the alphabet that the automaton will accept and the appropriate objects within stimulus-response theory is fairly obvious. The letters of the alphabet correspond in a natural way to sets of stimulus elements presented on a given trial to an organism. So again, roughly speaking, the correspondence in this case is between the alphabet and selected stimuli. It may seem like a happy accident, but the correspondences between inputs to the automata and stimuli presented to the organism, and between internal states of the machine and responses of the organism, are conceptually very natural.

Because of the conceptual importance of the issues that have been raised by linguists for the future development of psychological theory, perhaps above all because language behavior is the most characteristically human aspect of our behavior patterns, it is important to be as clear as

possible about the claims that can be made for a stimulus-response theory whose basic concepts seem so simple and to many so woefully inadequate to explain complex behavior, including language behavior. I cannot refrain from mentioning two examples that present very useful analogies. First is the reduction of all standard mathematics to the concept of set and the simple relation of an element being a member of a set. From a naive standpoint, it seems unbelievable that the complexities of higher mathematics can be reduced to a relation as simple as that of set membership. But this is indubitably the case, and we know in detail how the reduction can be made. This is not to suggest, for instance, that in thinking about a mathematical problem or even in formulating and verifying it explicitly, a mathematician operates simply in terms of endlessly complicated statements about set membership. By appropriate explicit definition we introduce many additional concepts, the ones actually used in discourse. The fact remains, however, that the reduction to the single relationship of set membership can be made and in fact has been carried out in detail. The second example, which is close to our present inquiry, is the status of simple machine languages for computers. Again, from the naive standpoint it seems incredible that modern computers can do the things they can in terms either of information processing or numerical computing when their basic language consists essentially just of finite sequences of 1's and 0's; but the more complex computer languages that have been introduced are not at all for the convenience of the machines but for the convenience of human users. It is perfectly clear how any more complex language, like ALGOL, can be reduced by a compiler or other device to a simple machine language. The same attitude, it seems to me, is appropriate toward stimulus-response theory. We cannot hope to deal directly in stimulus-response connections with complex human behavior. We can hope, as in the two cases just mentioned, to construct a satisfactory systematic theory in terms of which a chain of explicit definitions of new and ever more complex concepts can be introduced. It is these new and explicitly defined concepts that will be related directly to the more complex forms of behavior. The basic idea of stimulus-response association or connection is close enough in character to the concept of set membership or to the basic idea of automata to make me confident that new and better versions of stimulus-response theory may be expected in the future and that the scientific potentiality of

theories stated essentially in this framework has by no means been exhausted.

Before turning to specific mathematical developments, it will be useful to make explicit how the developments in this paper may be used to show that many of the common conceptions of conditioning, and particularly the claims that conditioning refers only to simple reflexes like those of salivation or eye blinking, are mistaken. The mistake is to confuse particular restricted applications of the fundamental theory with the range of the theory itself. Experiments on classical conditioning do indeed represent a narrow range of experiments from a broader conceptual standpoint. It is important to realize, however, that *experiments* on classical conditioning do not define the range and limits of conditioning theory itself. The main aim of the present paper is to show how any finite automaton, no matter how complicated, may be constructed purely within stimulus-response theory. But from the standpoint of automata, classical conditioning represents a particularly trivial example of an automaton. Classical conditioning may be represented by an automaton having a one-letter alphabet and a single internal state. The next simplest case corresponds to the structure of classical discrimination experiments. Here there is more than a single letter to the alphabet, but the transition table of the automaton depends in no way on the internal state of the automaton. In the case of discrimination, we may again think of the responses as corresponding to the internal states of the automaton. In this sense there is more than one internal state, contrary to the case of classical conditioning, but what is fundamental is that the transition table of the automaton does not depend on the internal states but only on the external stimuli presented according to a schedule fixed by the experimenter. It is of the utmost importance to realize that this restriction, as in the case of classical conditioning experiments, is not a restriction that is in any sense inherent in conditioning theory itself. It merely represents concentration on a certain restricted class of experiments.

Leaving the technical details for later, it is still possible to give a very clear example of conditioning that goes beyond the classical cases and yet represents perhaps the simplest non-trivial automaton. By non-trivial I mean: there is more than one letter in the alphabet; there is more than one internal state; and the transition table of the automaton is a function of both the external stimulus and the current internal state. As an example,

we may take a rat being run in a maze. The reinforcement schedule for the rat is set up so as to make the rat become a two-state automaton. We will use as the external alphabet of the automaton a two-letter alphabet consisting of a black or a white card. Each choice point of the maze will consist of either a left turn or a right turn. At each choice point either a black card or a white card will be present. The following table describes both the reinforcement schedule and the transition table of the automaton.

	L	*R*
LB	1	0
LW	0	1
RB	0	1
RW	1	0

Thus the first row shows that when the previous response has been left (L) and a black stimulus card (*B*) is presented at the choice point, with probability one the animal is reinforced to turn left. The second row indicates that when the previous response is left and a white stimulus card is presented at the choice point, the animal is reinforced 100% of the time to turn right, and so forth, for the other two possibilities. From a formal standpoint this is a simple schedule of reinforcement, but already the double aspect of contingency on both the previous response and the displayed stimulus card makes the schedule more complicated in many respects than the schedules of reinforcement that are usually run with rats. I have not been able to get a uniform prediction from my experimental colleagues as to whether it will be possible to teach rats to learn this schedule. (Most of them are confident pigeons can be trained to respond like non-trivial two-state automata.[1]) One thing to note about this schedule is that it is recursive in the sense that if the animal is properly trained according to the schedule, the length of the maze will be of no importance. He will always make a response that depends only upon his previous response and the stimulus card present at the choice point.

There is no pretense that this simple two-state automaton is in any sense adequate to serious language learning. I am not proposing, for example, that there is much chance of teaching even a simple one-sided linear grammar to rats. I am proposing to psychologists, however, that already automata of a small number of states present immediate experimental challenges in terms of what can be done with animals of each

species. For example, what is the most complicated automaton a monkey may be trained to imitate? In this case, there seems some possibility of approaching at least reasonably complex one-sided linear grammars (using the theorem that any one-sided linear grammar is definable by a finite-state automaton). In the case of the lower species, it will be necessary to exploit to the fullest the kind of stimuli to which the organisms are most sensitive and responsive in order to maximize the complexity of the automata they can imitate.

If a generally agreed upon definition of complexity for finite automata can be reached, it will be possible to use this measure to gauge the relative level of organizational complexity that can be achieved by a given species, at least in terms of an external schedule of conditioning and reinforcement. I do want to emphasize that the measures appropriate to experiments with animals are almost totally different from the measures that have been discussed in the recent literature of automata as complexity measures for computations. What is needed in the case of animals is the simple and orderly arrangement on a complexity scale of automata that have a relatively small number of states and that accept a relatively small alphabet of stimuli. The number of distinct training conditions is not a bad measure and can be used as a first approximation. Thus in the case of classical conditioning, this number is one. In the case of discrimination between black and white stimuli, the number is two. In the case of the two-state automaton described for the maze experiment, this number is four, but there are some problems with this measure. It is not clear that we would regard as more complex than this two-state automaton, an organism that masters a discrimination experiment consisting of six different responses to six different discriminating stimuli. Consequently, what I have said here about complexity is pre-systematic. I do think the development of an appropriate scale of complexity can be of theoretical interest, especially in cross-species comparison of intellectual power.

The remainder of this paper is devoted to the technical development of the general ideas already discussed. Section II is concerned with standard notions of finite and probabilistic automata. Readers already familiar with this literature should skip this section and go on to the treatment of stimulus-response theory in Section III. It has been necessary to give a rigorous axiomatization of stimulus-response theory in order to formulate

the representation theorem for finite automata in mathematically precise form. However, the underlying ideas of stimulus-response theory as formulated in Section III will be familiar to all experimental psychologists. In Section IV the most important result of the paper is proved, namely, that any finite automaton can be represented at asymptote by an appropriate model of stimulus-response theory. In Section V some extensions of these results to probabilistic automata are sketched, and an example from arithmetic is worked out in detail.

The relationship between stimulus-response theory and grammars is established in Section IV by known theorems relating automata to grammars. The results in the present paper are certainly restricted regarding the full generality of context-free languages. Weakening these restrictions will be the focus of a subsequent paper.

Some results on tote hierarchies and *plans* in the sense of Miller *et al.* (1960) are also given in Section IV. The representation of tote hierarchies by stimulus-response models follows directly from the main theorem of that section.

II. AUTOMATA

The account of automata given here is formally self-contained, but not really self-explanatory in the deeper sense of discussing and interpreting in adequate detail the systematic definitions and theorems. I have followed closely the development in the well-known article of Rabin and Scott (1959), and for probabilistic automata, the article of Rabin (1963).

DEFINITION 1: *A structure* $\mathfrak{A}=\langle A, \Sigma, M, s_0, F\rangle$ *is a finite (deterministic) automaton if and only if*

(i) A *is a finite, nonempty set (the set of states of* $\mathfrak{A}$*),*

(ii) Σ *is a finite, nonempty set (the alphabet),*

(iii) M *is a function from the Cartesian product* $A \times \Sigma$ *to* A *(*M *defines the transition table of* $\mathfrak{A}$*),*

(iv) s_0 *is in* A *(*s_0 *is the initial state of* $\mathfrak{A}$*),*

(v) F *is a subset of* A *(*F *is the set of final states of* $\mathfrak{A}$*).*

In view of the generality of this definition it is apparent that there are a great variety of automata, but as we shall see, this generality is easily matched by the generality of the models of stimulus-response theory.

In notation now nearly standardized, Σ^* is the set of finite sequences of elements of Σ, including the empty sequence Λ. The elements of Σ^* are

ordinarily called *tapes*. If $\sigma_1, \ldots, \sigma_k$ are in Σ, then $x = \sigma_1 \cdots \sigma_k$ is in Σ^*. (As we shall see, these tapes correspond in a natural way to finite sequences of sets of stimulus elements.) The function M can be extended to a function from $A \times \Sigma^*$ to A by the following recursive definition for s in A, x in Σ^*, and σ in Σ.

$$M(s, \Lambda) = s$$
$$M(s, x\sigma) = M(M(s, x), \sigma).$$

DEFINITION 2: *A tape x of Σ^* is accepted by $\mathfrak{A}$ if and only if $M(s_0, x)$ is in F. A tape x that is accepted by $\mathfrak{A}$ is a sentence of $\mathfrak{A}$.*

We shall also refer to tapes as *strings* of the alphabet Σ.

DEFINITION 3: *The language L generated by $\mathfrak{A}$ is the set of all sentences of $\mathfrak{A}$, i.e., the set of all tapes accepted by $\mathfrak{A}$.*

Regular languages are sometimes defined just as those languages generated by some finite automata. An independent, set-theoretical characterization is also possible. The basic result follows from Kleene's (1956) fundamental analysis of the kind of events definable by McCulloch-Pitts nets. Several equivalent formulations are given in the article by Rabin and Scott. From a linguistic standpoint probably the most useful characterization is that to be found in Chomsky (1963, pp. 368–371). Regular languages are generated by one-sided linear grammars. Such grammars have a finite number of rewrite rules, which in the case of right-linear rules, are of the form

$$A \to xB.$$

Whichever of several equivalent formulations is used, the fundamental theorem, originally due to Kleene, but closely related to the theorem of Myhill given by Rabin and Scott, is this.

THEOREM ON REGULAR LANGUAGES: *Any regular language is generated by some finite automaton, and every finite automaton generates a regular language.*

For the main theorem of this article, we need the concepts of isomorphism and equivalence of finite automata. The definition of isomorphism is just the natural set-theoretical one for structures like automata.

DEFINITION 4: *Let $\mathfrak{A} = \langle A, \Sigma, M, s_0, F \rangle$ and $\mathfrak{A}' = \langle A', \Sigma', M', s_0', F' \rangle$ be finite automata. Then $\mathfrak{A}$ and $\mathfrak{A}'$ are isomorphic if and only if there exists a function f such that*

(i) *f is one-one,*
(ii) *Domain of f is $A \cup \Sigma$ and range of f is $A' \cup \Sigma'$,*
(iii) *For every a in $A \cup \Sigma$*

$$a \in A \quad \text{if and only if} \quad f(a) \in A',$$

(iv) *For every s in A and σ in Σ*

$$f(M(s, \sigma)) = M'(f(s), f(\sigma)),$$

(v) $f(s_0) = s_0'$,
(vi) *For every s in A*

$$s \in F \quad \text{if and only if} \quad f(s) \in F'.$$

It is apparent that conditions (i)–(iii) of the definition imply that for every a in $A \cup \Sigma$

$$a \in \Sigma \quad \text{if and only if} \quad f(a) \in \Sigma',$$

and consequently, this condition on Σ need not be stated. From the standpoint of the general algebraic or set-theoretical concept of isomorphism, it would have been more natural to define an automaton in terms of a basic set $B = A \cup \Sigma$, and then require that A and Σ are both subsets of B. Rabin and Scott avoid the problem by not making Σ a part of the automaton. They define the concept of an automaton $\mathfrak{A} = \langle A, M, s_0, F \rangle$ with respect to an alphabet Σ, but for the purposes of this paper it is also desirable to include the alphabet Σ in the definition of $\mathfrak{A}$ in order to make explicit the natural place of the alphabet in the stimulus-response models, and above all, to provide a simple setup for going from one alphabet Σ to another Σ'. In any case, exactly how these matters are handled is not of central importance here.

DEFINITION 5: *Two automata are equivalent if and only if they accept exactly the same set of tapes.*

This is the standard definition of equivalence in the literature. As it stands, it means that the definition of equivalence is neither stronger nor weaker than the definition of isomorphism, because, on the one hand, equivalent automata are clearly not necessarily isomorphic, and, on the other hand, isomorphic automata with different alphabets are not equivalent. It would seem natural to weaken the notion of equivalence to include two automata that generate distinct but isomorphic languages, or

sets of tapes, but this point will bear on matters here only tangentially.

A finite automaton is *connected* if for every state s there is a tape x such that $M(s_0, x)=s$. It is easy to show that every automaton is equivalent to a connected automaton, and the representation theorem of Section IV is restricted to connected automata. It is apparent that from a functional standpoint, states that cannot be reached by any tape are of no interest, and consequently, restriction to connected automata does not represent any real loss of generality. The difficulty of representing automata with unconnected states by stimulus-response models is that we have no way to condition the organism with respect to these states, at least in terms of the approach developed here.

It is also straightforward to establish a representation for probabilistic automata within stimulus-response theory, and, as will become apparent in Section V, there are some interesting differences in the way we may represent deterministic and probabilistic automata within stimulus-response theory.

DEFINITION 6: *A structure* $\mathfrak{A}=\langle A, \Sigma, p, s_0, F\rangle$ *is a (finite) probabilistic automaton if and only if*

(i) A *is a finite, nonempty set,*

(ii) Σ *is a finite, nonempty set,*

(iii) p *is a function on* $A\times\Sigma$ *such that for each* s *in* A *and* σ *in* Σ, $p_{s,\sigma}$ *is a probability density over* A, *i.e.,*

(a) *for each* s' *in* A, $p_{s,\sigma}(s')\geqslant 0$,

(b) $\sum_{s'\in A} p_{s,\sigma}(s')=1$,

(iv) s_0 *is in* A,

(v) F *is a subset of* A.

The only change in generalizing from Definition 1 to Definition 6 is found in (iii), although it is natural to replace (iv) by an initial probability density. It is apparent how Definition 4 must be modified to characterize the isomorphism of probabilistic automata, and so the explicit definition will not be given.

III. STIMULUS-RESPONSE THEORY

The formalization of stimulus-response theory given here follows closely the treatment in Estes and Suppes (1959b) and Suppes and Atkinson (1960). Some minor changes have been made to facilitate the treatment of finite

automata, but it is to be strongly emphasized that none of the basic ideas or assumptions has required modification.

The theory is based on six primitive concepts, each of which has a direct psychological interpretation. The first one is the set S of stimuli, which we shall assume is not empty, but which we will not restrict to being either finite or infinite on all occasions. The second primitive concept is the set R of responses and the third primitive concept the set E of possible reinforcements. As in the case of the set of stimuli, we need not assume that either R or E is finite, but in the present applications to the theory of finite automata we shall make this restrictive assumption. (For a proper treatment of phonology it will clearly be necessary to make R, and probably E as well, infinite with at the very least a strong topological if not metric structure.)

The fourth primitive concept is that of a measure μ on the set of stimuli. In case the set S is finite this measure is often the number of elements in S. For the general theory we shall assume that the measure of S itself is always finite, i.e., $\mu(S) < \infty$.

The fifth primitive concept is the sample space X. Each element x of the sample space represents a possible experiment, that is, an infinite sequence of trials. In the present theory, each trial may be described by an ordered quintuple $\langle C, T, s, r, e \rangle$, where C is the conditioning function, T is the subset of stimuli presented to the organism on the given trial, s is the sampled subset of T, r is the response made on the trial, and e is the reinforcement occurring on that trial. It is not possible to make all the comments here that are required for a full interpretation and understanding of the theory. For those wanting a more detailed description, the two references already given will prove useful. A very comprehensive set of papers on stimulus-sampling theory has been put together in the collection edited by Neimark and Estes (1967). The present version of stimulus-response theory should in many respects be called stimulus-sampling theory, but I have held to the more general stimulus-response terminology to emphasize the juxtaposition of the general ideas of behavioral psychology on the one hand and linguistic theory on the other. In addition, in the theoretical applications to be made here the specific sampling aspects of stimulus-response theory are not as central as in the analysis of experimental data.

Because of the importance to be attached later to the set T of stimuli

presented on each trial, its interpretation in classical learning theory should be explicitly mentioned. In the case of simple learning, for example, in classical conditioning, the set T is the same on all trials and we would ordinarily identify the sets T and S. In the case of discrimination learning, the set T varies from trial to trial, and the application we are making to automata theory falls generally under the discrimination case. The conditioning function C is defined over the set R of responses and C_r is the subset of S conditioned or connected to response r on the given trial. How the conditioning function changes from trial to trial is made clear by the axioms.

From the quintuple description of a given trial it is clear that certain assumptions about the behavior that occurs on a trial have already been made. In particular it is apparent that we are assuming that only one sample of stimuli is drawn on a given trial, that exactly one response occurs on a trial and that exactly one reinforcement occurs on a trial. These assumptions have been built into the set-theoretical description of the sample space X and will not be an explicit part of our axioms.

Lying behind the formality of the ordered quintuples representing each trial is the intuitively conceived temporal ordering of events on any trial, which may be represented by the following diagram:

State of conditioning at beginning of trial	→	presentation of stimuli	→	sampling of stimuli	→	response	→	reinforcement	→	state of conditioning at beginning of new trial.
C_n	→	T_n	→	s_n	→	r_n	→	e_n	→	C_{n+1}

The sixth and final primitive concept is the probability measure P on the appropriate Borel field of cylinder sets of X. The exact description of this Borel field is rather complicated when the set of stimuli is not finite, but the construction is standard, and we shall assume the reader can fill in details familiar from general probability theory. It is to be emphasized that all probabilities must be defined in terms of the measure P.

We also need certain notation to take us back and forth between elements or subsets of the sets of stimuli, responses, and reinforcements to events of the sample space X. First, r_n is the event of response r on trial n, that is, the set of all possible experimental realizations or elements of X

having r as a response on the nth trial. Similarly, $e_{r,n}$ is the event of response r's being reinforced on trial n. The event $e_{0,n}$ is the event of no reinforcement on trial n. In like fashion, C_n is the event of conditioning function C occurring on trial n, T_n is the event of presentation set T occurring on trial n, and so forth. Additional notation that does not follow these conventions will be explicitly noted.

We also need a notation for sets defined by events occurring up to a given trial. Reference to such sets is required in expressing that central aspects of stimulus conditioning and sampling are independent of the pattern of past events. If I say that Y_n is an n-cylinder set, I mean that the definition of Y_n does not depend on any event occurring after trial n. However, an even finer breakdown is required that takes account of the postulated sequence $C_n \rightarrow T_n \rightarrow s_n \rightarrow r_n \rightarrow e_n$ on a given trial, so in saying that Y_n is a C_n-cylinder set what is meant is that its definition does not depend on any event occurring after C_n on trial n, i.e., its definition could depend on T_{n-1} or C_n, for example, but not on T_n or s_n. As an abbreviated notation, I shall write $Y(C_n)$ for this set and similarly for other cylinder sets. The notation Y_n without additional qualification shall always refer to an n-cylinder set.

Also, to avoid an overly cumbersome notation, event notation of the sort already indicated will be used, e.g., $e_{r,n}$, for reinforcement of response r on trial n, but also the notation $\sigma \in C_{r,n}$ for the event of stimulus σ's being conditioned to response r on trial n.

To simplify the formal statement of the axioms it shall be assumed without repeated explicit statement that any given events on which probabilities are conditioned have positive probability. Thus, for example, the tacit hypothesis of Axiom S2 is that $P(T_m) > 0$ and $P(T_n) > 0$.

The axioms naturally fall into three classes. Stimuli must be sampled in order to be conditioned, and they must be conditioned in order for systematic response patterns to develop. Thus, there are naturally three kinds of axioms: sampling axioms; conditioning axioms; and response axioms. A verbal formulation of each axiom is given together with its formal statement. From the standpoint of formulations of the theory already in the literature, perhaps the most unusual feature of the present axioms is not to require that the set S of stimuli be finite. It should also be emphasized that for any one specific kind of detailed application additional specializing assumptions are needed. Some indication of these will be

given in the particular application to automata theory, but it would take us too far afield to explore these specializing assumptions in any detail and with any faithfulness to the range of assumptions needed for different experimental applications.

DEFINITION 7: *A structure* $\mathscr{S}=\langle S, R, E, \mu, X, P\rangle$ *is a stimulus-response model if and only if the following axioms are satisfied:*

Sampling Axioms

S1. $P(\mu(s_n)>0)=1$.

(On every trial a set of stimuli of positive measure is sampled with probability 1.*)*

S2. $P(s_m \mid T_m)=P(s_n \mid T_n)$.

(If the same presentation set occurs on two different trials, then the probability of a given sample is independent of the trial number.)

S3. *If* $s\cup s'\subseteq T$ *and* $\mu(s)=\mu(s')$ *then* $P(s_n \mid T_n)=P(s'_n \mid T_n)$.

(Samples of equal measure that are subsets of the presentation set have an equal probability of being sampled on a given trial.)

S4. $P(s_n \mid T_n, Y_n(C_n))=P(s_n \mid T_n)$.

(The probability of a particular sample on trial n, given the presentation set of stimuli, is independent of any preceding pattern $Y_n(C_n)$ *of events.)*

Conditioning Axioms

C1. *If* $r, r'\in R$, $r\neq r'$ *and* $C_r\cap C_{r'}\neq 0$, *then* $P(C_n)=0$.

(On every trial with probability 1 *each stimulus element is conditioned to at most one response.)*

C2. *There exists a* $c>0$ *such that for every* σ, C, r, n, s, e_r, *and* Y_n

$$P(\sigma\in C_{r,n+1} \mid \sigma\notin C_{r,n}, \sigma\in s_n, e_{r,n}, Y_n)=c.$$

(The probability is c of any sampled stimulus element's becoming conditioned to the reinforced response if it is not already so conditioned, and this probability is independent of the particular response, trial number, or any preceding pattern Y_n *of events.)*

C3. $P(C_{n+1} \mid C_n, e_{0,n})=1$.

(With probability 1, *the conditioning of all stimulus elements remains the same if no response is reinforced.)*

C4. $P(\sigma\in C_{r,n+1} \mid \sigma\in C_{r,n}, \sigma\notin s_n, Y_n)=1$.

(With probability 1, *the conditioning of unsampled stimuli does not change.)*

Response Axioms

R1. *If* $\bigcup_{r \in R} C_r \cap s \neq 0$ *then*

$$P(r_n \mid C_n, s_n, Y(s_n)) = \frac{\mu(s \cap C_r)}{\mu(s \cap \bigcup C_r)}.$$

(If at least one sampled stimulus is conditioned to some response, then the probability of any response is the ratio of the measure of sampled stimuli conditioned to this response to the measure of all the sampled conditioned stimuli, and this probability is independent of any preceding pattern $Y(s_n)$ *of events.)*

R2. *If* $\bigcup_{r \in R} C_r \cap s = 0$ *then there is a number* ρ_r *such that*

$$P(r_n \mid C_n, s_n, Y(s_n)) = \rho_r.$$

(If no sampled stimulus is conditioned to any response, then the probability of any response r is a constant guessing probability ρ_r *that is independent of n and any preceding pattern* $Y(s_n)$ *of events.)*

A general discussion of these axioms and their implications for a wide range of psychological experiments may be found in the references already cited. The techniques of analysis used in the next section of this paper are extensively exploited and applied to a number of experiments in Suppes and Atkinson (1960).

IV. REPRESENTATION OF FINITE AUTOMATA

A useful beginning for the analysis of how we may represent finite automata by stimulus-response models is to see what is wrong with the most direct approach possible. The difficulties that turn up may be illustrated by the simple example of a two-letter alphabet (i.e., two stimuli σ_1 and σ_2, as well as the "start-up" stimulus σ_0) and a two-state automaton (i.e., two responses r_1 and r_2). Consideration of this example, already mentioned in the introductory section, will be useful for several points of later discussion.

By virtue of Axiom S1, the single presented stimulus must be sampled on each trial, and we assume that for every n,

$$0 < P(\sigma_0 \in s_n), P(\sigma_1 \in s_n), P(\sigma_2 \in s_n) < 1.$$

Suppose, further, the transition table of the machine is:

	r_1	r_2
$r_1\sigma_1$	1	0
$r_1\sigma_2$	0	1
$r_2\sigma_1$	0	1
$r_2\sigma_2$	1	0

which requires knowledge of both r_i and σ_j to predict what response should be next. The natural and obvious reinforcement schedule for imitating this machine is:

$$P(e_{1,n} \mid \sigma_{1,n}, r_{1,n-1}) = 1$$
$$P(e_{2,n} \mid \sigma_{1,n}, r_{2,n-1}) = 1$$
$$P(e_{2,n} \mid \sigma_{2,n}, r_{1,n-1}) = 1$$
$$P(e_{1,n} \mid \sigma_{2,n}, r_{2,n-1}) = 1,$$

where $\sigma_{i,n}$ is the event of stimulus σ_i's being sampled on trial n. But for this reinforcement schedule the conditioning of each of the two stimuli continues to fluctuate from trial to trial, as may be illustrated by the following sequence. For simplification and without loss of generality, we may assume that the conditioning parameter c is 1, and we need indicate no sampling, because as already mentioned, the single stimulus element in each presentation set will be sampled with probability 1. We may represent the states of conditioning (granted that each stimulus is conditioned to either r_1 or r_2) by subsets of $S=\{\sigma_0, \sigma_1, \sigma_2\}$. Thus, if $\{\sigma_1, \sigma_2\}$ represents the conditioning function, this means both elements σ_1 and σ_2 are conditioned to r_1; $\{\sigma_1\}$ means that only σ_1 is conditioned to r_1, and so forth. Consider then the following sequence from trial n to $n+2$:

$$\langle\{\sigma_2\}, \sigma_2 \in s, r_1, e_2\rangle \rightarrow \langle 0, \sigma_2 \in s, r_2, e_2\rangle \rightarrow \langle 0, \sigma_2 \in s, r_2, e_1\rangle.$$

The response on trial $n+1$ satisfies the machine table, but already on $n+2$ it does not, for $r_{2,n+1}\sigma_{2,n+2}$ should be followed by $r_{1,n+2}$. It is easy to show that this difficulty is fundamental and arises for any of the four possible conditioning states. (In working out these difficulties explicitly, the reader should assume that each stimulus is conditioned to either r_1 or r_2, which will be true for n much larger than 1 and $c=1$.)

What is needed is a quite different definition of the states of the Markov chain of the stimulus-response model. (For proof of a general Markov-chain theorem for stimulus-response theory, see Estes and Suppes, 1959b.) Naively, it is natural to take as the states of the Markov chain the possible states of conditioning of the stimuli in S, but this is wrong on two counts in the present situation. First, we must condition the *patterns* of responses and presentation sets, so we take as the set of stimuli for the model, $R \times S$, i.e., the Cartesian product of the set R of responses and the set S of stimuli. What the organism must be conditioned to respond to on trial n is the pattern consisting of the preceding response given on trial $n-1$ and the presentation set occurring on trial n.

It is still not sufficient to define the states of the Markov chain in terms of the states of conditioning of the elements in $R \times S$, because for reasons that are given explicitly and illustrated by many examples in Estes and Suppes (1959b) and Suppes and Atkinson (1960), it is also necessary to include in the definition of state the response r_i that actually occurred on the preceding trial. The difficulty that arises if $r_{i,n-1}$ is not included in the definition of state may be brought out by attempting to draw the tree in the case of the two-state automaton already considered. Suppose just the pattern $r_1\sigma_1$ is conditioned, and the other four patterns, σ_0, $r_1\sigma_2$,

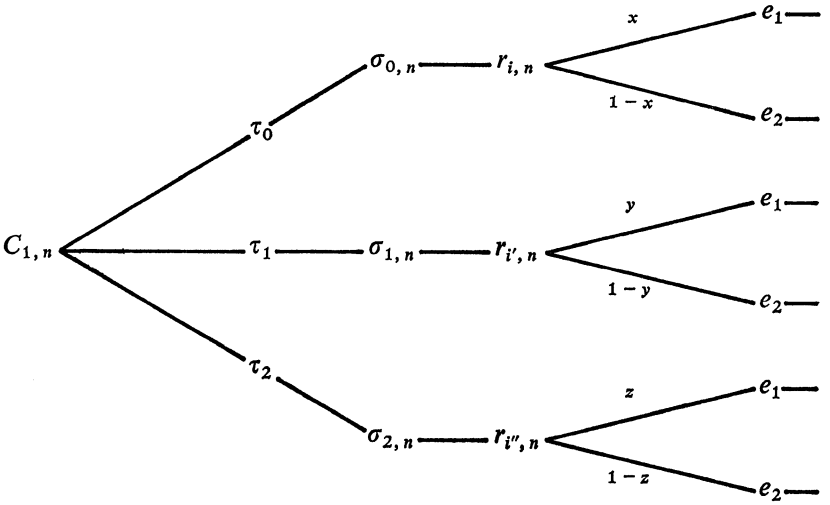


Fig. 1.

$r_2\sigma_1$, and $r_2\sigma_2$, are not. Let us represent this conditioning state by C_1, and let τ_j be the noncontingent probability of σ_j, $0 \leqslant j \leqslant 2$, on every trial with every $\tau_j > 0$. Then the tree is shown in Figure 1.

The tree is incomplete, because without knowing what response actually occurred on trial $n-1$ we cannot complete the branches (e.g., specify the responses), and for a similar reason we cannot determine the probabilities x, y and z. Moreover, we cannot remedy the situation by including among the branches the possible responses on trial $n-1$, for to determine their probabilities we would need to look at trial $n-2$, and this regression would not terminate until we reached trial 1.

So we include in the definition of state the response on trial $n-1$. On the other hand, it is not necessary in the case of deterministic finite automata to permit among the states all possible conditioning of the patterns in $R \times S$. We shall permit only two possibilities – the pattern is unconditioned or it is conditioned to the appropriate response because conditioning to the wrong response occurs with probability zero. Thus with p internal states or responses and m letters in Σ, there are $(m+1)p$ patterns, each of which is in one of two states, conditioned or unconditioned, and there are p possible preceding responses, so the number of states in the Markov chain is $p2^{(m+1)p}$. Actually, it is convenient to reduce this number further by treating σ_0 as a single pattern regardless of what preceding response it is paired with. The number of states is then $p2^{mp+1}$. Thus, for the simplest 2-state, 2-alphabet automaton, the number of states is 64. We may denote the states by ordered $mp+2$-tuples

$$\langle r_j, i_0, i_{0,1}, \ldots, i_{0,m}, \ldots, i_{p-1,m} \rangle,$$

where i_{kl} is 0 or 1 depending on whether the pattern $r_k\sigma_l$ is unconditioned or conditioned with $0 \leqslant k \leqslant p-1$ and $1 \leqslant l \leqslant m$; r_j is the response on the preceding trial, and i_0 is the state of conditioning of σ_0. What we want to prove is that starting in the purely unconditioned set of states $\langle r_j, 0, 0, \ldots, 0 \rangle$, with probability 1 the system will always ultimately be in a state that is a member of the set of fully conditioned states $\langle r_j, 1, 1, \ldots, 1 \rangle$. The proof of this is the main part of the proof of the basic representation theorem.

Before turning to the theorem we need to define explicitly the concept of a stimulus-response model's asymptotically becoming an automaton. As has already been suggested, an important feature of this definition is

this. The basic set S of stimuli corresponding to the alphabet Σ of the automaton is not the basic set of stimuli of the stimulus-response model, but rather, this basic set is the Cartesian product $R \times S$, where R is the set of responses. Moreover, the definition has been framed in such a way as to permit only a single element of S to be presented and sampled on each trial; this, however, is an inessential restriction used here in the interest of conceptual and notational simplicity. Without this restriction the basic set would be not $R \times S$, but $R \times \mathscr{P}(S)$, where $\mathscr{P}(S)$ is the power set of S, i.e., the set of all subsets of S, and then each letter of the alphabet Σ would be a subset of S rather than a single element of S. What is essential is to have $R \times S$ rather than S as the basic set of stimuli to which the axioms of Definition 6 apply.

For example, the pair (r_i, σ_j) must be sampled *and* conditioned as a pattern, and the axioms are formulated to require that what is sampled and conditioned be a subset of the presentation set T on a given trial. In this connection to simplify notation I shall often write $T_n = (r_{i,n-1}, \sigma_{j,n})$ rather than

$$T = \{(r_i, \sigma_j)\},$$

but the meaning is clear. T_n is the presentation set consisting of the single pattern (or element) made up of response r_i on trial $n-1$ and stimulus element σ_j on trial n, and from Axiom S1 we know that the pattern is sampled because it is the only one presented.

From a psychological standpoint something needs to be said about part of the presentation set being the previous response. In the first place, and perhaps most importantly, this is not an ad hoc idea adopted just for the purposes of this paper. It has already been used in a number of experimental studies unconnected with automata theory. Several worked-out examples are to be found in various chapters of Suppes and Atkinson (1960).

Secondly, and more importantly, the use of $R \times S$ is formally convenient, but is not at all necessary. The classical S-R tradition of analysis suggests a formally equivalent, but psychologically more realistic approach. Each response r produces a stimulus σ_r, or more generally, a set of stimuli. Assuming again, for formal simplicity just one stimulus element σ_r, rather than a set of stimuli, we may replace R by the set of

stimuli S_R, with the purely contingent presentation schedule

$$P(\sigma_{r,n} \mid r_{n-1}) = 1,$$

and in the model we now consider the Cartesian product $S_R \times S$ rather than $R \times S$. Within this framework the important point about the presentation set on each trial is that one component is purely subject-controlled and the other purely experimenter-controlled – if we use familiar experimental distinctions. The explicit use of S_R rather than R promises to be important in training animals to perform like automata, because the external introduction of σ_r reduces directly and significantly the memory load on the animal.[2] The importance of S_R for models of children's language learning is less clear.

DEFINITION 8: *Let* $\mathscr{S} = \langle R \times S, R, E, \mu, X, P \rangle$ *be a stimulus-response model where*

$$R = \{r_0, \ldots, r_{p-1}\}$$
$$S = \{\sigma_0, \ldots, \sigma_m\}$$
$$E = \{e_0, \ldots, e_{p-1}\},$$

and $\mu(S')$ *is the cardinality of* S' *for* $S' \subseteq S$. *Then* $\mathscr{S}$ *asymptotically becomes the automaton* $\mathfrak{A}(\mathscr{S}) = \langle R, S - \{\sigma_0\}, M, r_0, F \rangle$ *if and only if*

(i) *as* $n \to \infty$ *the probability is* 1 *that the presentation set* T_n *is* $(r_{i,n-1}, \sigma_{j,n})$ *for some* i *and* j,

(ii) $M(r_i, \sigma_j) = r_k$ *if and only if* $\lim_{n\to\infty} P(r_{k,n} \mid T_n = (r_{i,n-1}, \sigma_{j,n})) = 1$ *for* $0 \leqslant i \leqslant p-1$ *and* $1 \leqslant j \leqslant m$,

(iii) $\lim_{n\to\infty} P(r_{0,n} \mid T_n = (r_{i,n-1}, \sigma_{0,n})) = 1$ *for* $0 \leqslant i \leqslant p-1$,

(iv) $F \subseteq R$.

A minor but clarifying point about this definition is that the stimulus σ_0 is not part of the alphabet of the automaton $\mathfrak{A}(\mathscr{S})$, because a stimulus is needed to put the automaton in the initial state r_0, and from the standpoint of the theory being worked out here, this requires a stimulus to which the organism will give response r_0.[3] That stimulus is σ_0. The definition also requires that asymptotically the stimulus-response model $\mathscr{S}$ is nothing but the automaton $\mathfrak{A}(\mathscr{S})$. It should be clear that a much weaker and more general definition is possible. The automaton $\mathfrak{A}(\mathscr{S})$ could merely be embedded asymptotically in $\mathscr{S}$ and be only a part of the activities of $\mathscr{S}$. The simplest way to achieve this generalization is to make

the alphabet of the automaton only a proper subset of $S-\{\sigma_0\}$ and correspondingly for the responses that make up the internal states of the automaton; they need be only a proper subset of the full set R of responses. This generalization will not be pursued here, although something of the sort will be necessary to give an adequate stimulus-response account of the semantical aspects of language.

REPRESENTATION THEOREM FOR FINITE AUTOMATA: *Given any connected finite automaton, there is a stimulus-response model that asymptotically becomes isomorphic to it. Moreover, the stimulus-response model may have all responses initially unconditioned.*

Proof: Let $\mathfrak{A}=\langle A, \Sigma, M, s_0, F\rangle$ be any connected finite automaton. As indicated already, we represent the set A of internal states by the set R of responses; we shall use the natural correspondence $s_i - r_i$, for $0\leqslant i\leqslant p-1$, where p is the number of states. We represent the alphabet Σ by the set of stimuli $\sigma_1, \ldots, \sigma_m$, and, for reasons already made explicit, we augment this set of stimuli by σ_0, to obtain

$$S = \{\sigma_0, \sigma_1, \ldots, \sigma_m\}.$$

For subsequent reference let f be the function defined on $A\cup\Sigma$ that establishes the natural one-one correspondence between A and R, and between Σ and $S-\{s_0\}$. (To avoid some trivial technical points I shall assume that A and Σ are disjoint.)

We take as the set of reinforcements

$$E = \{e_0, e_1, \ldots, e_{p-1}\},$$

and the measure $\mu(S')$ is the cardinality of S' for $S'\subseteq S$, so that as in Definition 8, we are considering a stimulus-response model $\mathscr{S}=\langle R\times S, R, E, \mu, X, P\rangle$. In order to show that $\mathscr{S}$ asymptotically becomes an automaton, we impose five additional restrictions on $\mathscr{S}$.

They are these.

First, in the case of reinforcement e_0 the schedule is this:

$$(1) \qquad P(e_{0,n} \mid \sigma_{0,n}) = 1,$$

i.e., if $\sigma_{0,n}$ is part of the presentation set on trial n, then with probability 1 response r_0 is reinforced – note that the reinforcing event $e_{0,n}$ is independent of the actual occurrence of the event $r_{0,n}$.

Second, the remaining reinforcement schedule is defined by the transition table M of the automaton $\mathfrak{A}$. Explicitly, for $j, k\neq 0$ and for all

i and n

(2) $P(e_{k,n} \mid \sigma_{j,n} r_{i,n-1}) = 1$ if and only if $M(f^{-1}(r_i), f^{-1}(\sigma_j)) = f^{-1}(r_k)$.

Third, essential to the proof is the additional assumption beyond (1) and (2) that the stimuli $\sigma_0, \ldots, \sigma_m$ each have a positive, noncontingent probability of occurrence on each trial (a model with a weaker assumption could be constructed but it is not significant to weaken this requirement). Explicitly, we then assume that for any cylinder set $Y(C_n)$ such that $P(Y(C_n))>0$

(3) $P(\sigma_{i,n}) = P(\sigma_{i,n} \mid Y(C_n)) \geqslant \tau_i > 0$

for $0 \leqslant i \leqslant m$ and for all trials n.

Fourth, we assume that the probability ρ_i of response r_i occurring when no conditioned stimuli is sampled is also strictly positive, i.e., for every response r_i

(4) $\rho_i > 0$,

which strengthens Axiom R2.

Fifth, for each integer k, $0 \leqslant k \leqslant mp+1$, we define the set Q_k as the set of states that have exactly k patterns conditioned, and $Q_{k,n}$ is the event of being in a state that is a member of Q_k on trial n. We assume that at the beginning of trial 1, no patterns are conditioned, i.e.,

(5) $P(Q_{0,1}) = 1$.

It is easy to prove that given the sets R, S, E and the cardinality measure μ, there are many different stimulus-response models satisfying restrictions (1)–(5), but for the proof of the theorem it is not necessary to select some distinguished member of the class of models because the argument that follows shows that all the members of the class asymptotically become isomorphic to $\mathfrak{A}$.

The main thing we want to prove is that as $n \to \infty$

(6) $P(Q_{mp+1,n}) = 1$.

We first note that if $j<k$ the probability of a transition from Q_k to Q_j is zero, i.e.,

(7) $P(Q_{j,n} \mid Q_{k,n-1}) = 0$,

moreover,

$$P(Q_{j,n} \mid Q_{k,n-1}) = 0, \tag{8}$$

even if $j>k$ unless $j=k+1$. In other words, in a single trial, at most one pattern can become conditioned.

To show that asymptotically (6) holds, it will suffice to show that there is an $\varepsilon>0$ such that on each trial n for $0\leqslant k\leqslant mp<n$ if $P(Q_{k,n})>0$,

$$P(Q_{k+1,n+1} \mid Q_{k,n}) \geqslant \varepsilon. \tag{9}$$

To establish (9) we need to show that there is a probability of at least ε of a stimulus pattern that is unconditioned at the beginning of trial n becoming conditioning on that trial. The argument given will be a uniform one that holds for any unconditioned pattern. Let $r^*\sigma^*$ be such a pattern on trial n.

Now it is well known that for a connected automaton, for every internal state s, there is a tape x such that

$$M(s_0, x) = s \tag{10}$$

and the length of x is not greater than the number of internal states. In terms of stimulus-response theory, x is a finite sequence of length not greater than p of stimulus elements. Thus we may take $x=\sigma_{i_1},\ldots,\sigma_{i_p}$ with $\sigma_{i_p}=\sigma^*$. We know by virtue of (3) that

$$\min_{0\leqslant i\leqslant m} \tau_i = \tau > 0. \tag{11}$$

The required sequence of responses $r_{i_1},\ldots,r_{i_p-1}$ will occur either from prior conditioning or if any response is not conditioned to the appropriate pattern, with guessing probability ρ_i. By virtue of (4)

$$\min_{0\leqslant i\leqslant p-1} \rho_i = \rho > 0. \tag{12}$$

To show that the pattern $r^*\sigma^*$ has a positive probability ε, of being conditioning on trial n, we need only take n large enough for the tape x to be "run", say, $n>p+1$, and consider the joint probability

$$P^* = P(\sigma_n^*, r_{n-1}^*, \sigma_{i_{p-1},n-1}, r_{i_{p-2},n-2}, \ldots, \sigma_{i_j,\,n-i_p}, r_{0,n-i_p-1}, \sigma_{0,n-i_p-1}). \tag{13}$$

The basic axioms and the assumptions (1)–(5) determine a lower bound on

P^* independent of n. First we note that for each of the stimulus elements $\sigma_0, \sigma_{i_j}, \ldots \sigma^*$, by virtue of (3) and (11)

$$P(\sigma_n^* \mid \ldots) \geqslant \tau, \ldots, P(\sigma_{0,n-i_p-1}) \geqslant \tau .$$

Similarly, from (4) and (12), as well as the response axioms, we know that for each of the responses $r_0, r_{i_1}, \ldots, r^*$

$$P(r_{n-1}^* \mid \ldots) \geqslant \rho, \ldots, P(r_{0,n-i_p-1} \mid \sigma_{0,n-i_p-1}) \geqslant \rho .$$

Thus we know that

$$P^* \geqslant \rho^p \tau^{p+1},$$

and given the occurrence of the event $\sigma_n^* r_{n-1}^*$, the probability of conditioning is c, whence we may take

$$\varepsilon = c\rho^p \tau^{p+1} > 0,$$

which establishes (9) and completes the proof.

Given the theorem just proved there are several significant corollaries whose proofs are almost immediate. The first combines the representation theorem for regular languages with that for finite automata to yield:

COROLLARY ON REGULAR LANGUAGES: *Any regular language is generated by some stimulus-response model at asymptote.*

Once probabilistic considerations are made a fundamental part of the scene, we can in several different ways go beyond the restriction of stimulus-response generated languages to regular languages, but I shall not explore these matters here.

I suspect that many psychologists or philosophers who are willing to accept the sense given here to the reduction of finite automata and regular languages to stimulus-response models will be less happy with the claim that one well-defined sense of the concepts of *intention*, *plan* and *purpose* can be similarly reduced. However, without any substantial new analysis on my part this can be done by taking advantage of an analysis already made by Miller and Chomsky (1963). The story goes like this. In 1960, Miller, Galanter, and Pribram published a provocative book entitled *Plans and the Structure of Behavior*. In this book they severely criticized stimulus-response theories for being able to account for so little of the significant behavior of men and the higher animals. They especially objected to the conditioned reflex as a suitable concept for

building up an adequate scientific psychology. It is my impression that a number of cognitively oriented psychologists have felt that the critique of *S-R* theory in this book is devastating.

As I indicated in the introductory section, I would agree that conditioned reflex *experiments* are indeed far too simple to form an adequate scientific basis for analyzing more complex behavior. This is as hopeless as would be the attempt to derive the theory of differential equations, let us say, from the elementary algebra of sets. Yet the more general theory of sets does encompass in a strict mathematical sense the theory of differential equations.

The same relation may be shown to hold between stimulus-response theory and the theory of plans, insofar as the latter theory has been systematically formulated by Miller and Chomsky.[4] The theory of plans is formulated in terms of tote units ('tote' is an acronym for the cycle test-operate-test-exit). A plan is then defined as a tote hierarchy, which is just a form of oriented graph, and every finite oriented graph may be represented as a finite automaton. So we have the result:

COROLLARY: *Any tote hierarchy in the sense of Miller and Chomsky is isomorphic to some stimulus-response model at asymptote.*

V. REPRESENTATION OF PROBABILISTIC AUTOMATA

From the standpoint of the kind of learning models and experiments characteristic of the general area of what has come to be termed *probability learning*, there is near at hand a straightforward approach to probabilistic automata. It is worth illustrating this approach, but it is perhaps even more desirable to discuss it with some explicitness in order to show why it is not fully satisfactory, indeed for most purposes considerably less satisfactory than a less direct approach that follows from the representation of deterministic finite automata already discussed.

The direct approach is dominated by two features: a probabilistic reinforcement schedule and the conditioning of input stimuli rather than response-stimulus patterns. The main simplification that results from these features is that the number of states of conditioning and consequently the number of states in the associated Markov chain is reduced. A two-letter, two-state probabilistic automaton, for example, requires 36 states in the associated Markov chain, rather than 64, as in the

deterministic case. We have, as before, the three stimuli σ_0, σ_1, and σ_2 and their conditioning possibilities, 2 for σ_0 as before, but now, in the probabilistic case, 3 for σ_1 and σ_2, and we also need as part of the state, not for purposes of conditioning, but in order to make the response-contingent reinforcement definite, the previous response, which is always either r_1 or r_2. Thus we have $2 \cdot 3 \cdot 3 \cdot 2 = 36$. Construction of the trees to compute transition probabilities for the Markov chain follows closely the logic outlined in the previous section. We may define the probabilistic reinforcement schedule by two equations, the first of which is deterministic and plays exactly the same role as previously:

$$P(e_{1,n} \mid \sigma_{0,n}) = 1$$

and

$$P(e_{1,n} \mid \sigma_{i,n} r_{j,n-1}) = \pi_{ij}$$

for $1 \leqslant i, j \leqslant 2$.

The fundamental weakness of this setup is that the asymptotic transition table representing the probabilistic automaton only holds in the mean. Even at asymptote the transition values fluctuate from trial to trial depending upon the actual previous reinforcement, not the probabilities π_{ij}. Moreover, the transition table is no longer the transition table of a Markov process. Knowledge of earlier responses and reinforcements will lead to a different transition table, whenever the number of stimuli representing a letter of the input alphabet is greater than one. These matters are well known in the large theoretical literature of probability learning and will not be developed further here.

For most purposes of application it seems natural to think of probabilistic automata as a generalization of deterministic automata intended to handle the problem of errors. A similar consideration of errors after a concept or skill has been learned is common in learning theory. Here is a simple example. In the standard version of all-or-none learning, the organism is in either the unconditioned state (U) or the conditioned state (C). The transition matrix for these states is

	C	U
C	1	0
U	c	$1-c$

and, consistent with the axioms of Section III,

(1) $P(\text{Correct response} \mid C) = 1$

(2) $P(\text{Correct response} \mid U) = \rho$

and

$$P(U_1) = 1,$$

i.e., the probability of being in state U on trial 1 is 1. Now by changing (2) to

$$P(\text{Correct response} \mid C) = 1 - \varepsilon,$$

for $\varepsilon > 0$, we get a model that predicts errors after conditioning has occurred.

Without changing the axioms of Section III we can incorporate such probabilistic-error considerations into the derivation of a representation theorem for probabilistic automata. One straightforward procedure is to postulate that the pattern sampled on each trial actually consists of N elements, and that in addition M background stimuli common to all trials are sampled, or available for sampling. By specializing further the sampling axioms S1–S4 and by adjusting the parameters M and N, we can obtain any desired probability ε of an error.

Because it seems desirable to develop the formal results with intended application to detailed learning data, I shall not state and prove a representation theorem for probabilistic automata here, but restrict myself to considering one example of applying a probabilistic automaton model to asymptotic performance data. The formal machinery for analyzing learning data will be developed in a subsequent paper.

The example I consider is drawn from arithmetic. For more than three years we have been collecting extensive data on the arithmetic performance of elementary-school students, in the context of various projects on computer-assisted instruction in elementary mathematics. Prior to consideration of automaton models, the main tools of analysis have been linear regression models. The dependent variables in these models have been the mean probability of a correct response to an item and the mean success latency. The independent variables have been structural features of items, i.e., arithmetic problems, that may be objectively identified independently of any analysis of response data. Detailed results for such

models are to be found in Suppes *et al.* (1967, 1968). The main conceptual weakness of the regression models is that they do not provide an explicit temporal analysis of the steps being taken by a student in solving a problem. They can identify the main variables but not connect these variables in a dynamically meaningful way. In contrast, analysis of the temporal process of problem solution is a natural and integral part of an automaton model.

An example that is typical of the skills and concepts encountered in arithmetic is column addition of two integers. For simplicity I shall consider only problems for which the two given numbers and their sum all have the same number of digits. It will be useful to begin by defining a deterministic automaton that will perform the desired addition by outputting one digit at a time reading from right to left, just as the students are required to do at computer-based teletype terminals. For this purpose it is convenient to modify in inessential ways the earlier definition of an automaton. An automaton will now be defined as a structure $\mathfrak{A}=\langle A, \Sigma_I, \Sigma_O, M, Q, s_0\rangle$ where A, Σ_I and Σ_O are non-empty finite sets, with A being the set of internal states as before, Σ_I the input alphabet, and Σ_O the output alphabet. Also as before, M is the transition function mapping $A\times\Sigma_I$ into A, and s_0 is the initial state. The function Q is the output function mapping $A\times\Sigma_I$ into Σ_O.

For column addition of two integers in standard base-ten representation, an appropriate automaton is the following:

$$
\begin{aligned}
A &= \{0, 1\},\\
\Sigma_I &= \{(m, n): 0 \leqslant m, n \leqslant 9\}\\
\Sigma_O &= \{0, 1, \ldots, 9\}\\
M(k, (m, n)) &= \begin{cases} 0 & \text{if} \quad m+n+k \leqslant 9.\\ 1 & \text{if} \quad m+n+k > 9, \end{cases} \quad \text{for} \quad k = 0, 1.\\
Q(k, (m, n)) &= (k+m+n) \bmod 10.\\
s_0 &= 0.
\end{aligned}
$$

Thus the automaton operates by adding first the ones' column, storing as internal state 0 if there is no carry, 1 if there is a carry, outputting the sum of the ones' column modulus 10, and then moving on to the input of the two tens' column digits, etc. The initial internal state s_0 is 0 because at the beginning of the problem there is no "carry".

For the analysis of student data it is necessary to move from a deterministic to a probabilistic automaton. The number of possible parameters that can be introduced is uninterestingly large. Each transition $M(k, (m, n))$ may be replaced by a probabilistic transition $1-\varepsilon_{k,m,n}$ and $\varepsilon_{k,m,n}$, and each output $Q(k(m, n))$, by ten probabilities for a total of 2200 parameters. Using the sort of linear regression model described above we have found that a fairly good account of student performance data can be obtained by considering two structural variables, C_i, the number of carries in problem item i, and D_i, the number of digits or columns. Let p_i be the mean probability of a correct response on item i and let

$$z_i = \log \frac{1 - p_i}{p_i}.$$

The regression model is then characterized by the equation

$$(1) \qquad z_i = \alpha_0 + \alpha_1 C_i + \alpha_2 D_i,$$

and the coefficients α_0, α_1, and α_2 are estimated from the data.

A similar three-parameter automaton model is structurally very natural. First, two parameters, ε and η, are introduced according to whether there is a "carry" to the next column.

$$P(M(k, (m, n)) = 0 \mid k + m + n \leqslant 9) = 1 - \varepsilon$$

and

$$P(M(k, (m, n)) = 1 \mid k + m + n > 9) = 1 - \eta.$$

In other words, if there is no "carry", the probability of a correct transition is $1-\varepsilon$ and if there is a "carry" the probability of such a transition is $1-\eta$. The third parameter, γ, is simply the probability of an output error. Conversely, the probability of a correct output is:

$$P(Q(k, (m, n)) = (k + m + n) \bmod 10) = 1 - \gamma.$$

Consider now problem i with C_i carries and D_i digits. If we ignore the probability of two errors leading to a correct response – e.g., a transition error followed by an output error – then the probability of a correct answer is just:

$$(2) \qquad P(\text{Correct Answer to Problem } i) = (1 - \gamma)^{D_i} (1 - \eta)^{C_i} (1 - \varepsilon)^{D_i - C_i - 1}.$$

As already indicated it is important to realize that this equation is an approximation of the "true" probability. However, to compute the exact probability it is necessary to make a definite assumption about how the probability γ of an output error is distributed among the 9 possible wrong responses. A simple and intuitively appealing one-parameter model is the one that arranges the 10 digits on a circle in natural order with 9 next to 0, and then makes the probability of an error j steps to the right or left of the correct response δ^j. For example, if 5 is the correct digit, then the probability of responding 4 is δ, of 3 is δ^2, of 2 is δ^3, of 1 is δ^4, of 0 is δ^5, of 6 is δ, of 7 is δ^2, etc. Thus in terms of the original model

$$\gamma = 2(\delta + \delta^2 + \delta^3 + \delta^4) + \delta^5 .$$

Consider now the problem

$$\begin{array}{r} 47 \\ +\,15. \\ \hline \end{array}$$

Then, where d_i = the ith digit response,

$$P(d_1 = 2) = (1 - \gamma)$$
$$P(d_2 = 6) = (1 - \gamma)(1 - \eta) + \eta\delta .$$

Here the additional term is $\eta\delta$, because if the state entered is 0 rather than 1 when the pair (7, 5) is input, the only way of obtaining a correct answer is for 6 to be given as the sum of $0+4+1$, which has a probability δ. Thus the probability of a correct response to this problem is $(1-\gamma)[(1-\gamma)(1-\eta)+\eta\delta]$. Hereafter we shall ignore the $\eta\delta$ (or $\varepsilon\delta$) terms.

Returning to Equation (2) we may get a direct comparison with the linear regression model defined by Equation (1), if we take the logarithm of both sides to obtain:

$$\log p_i = D_i \log(1 - \gamma) + C_i \log(1 - \eta) + (D_i - C_i - 1)\log(1 - \varepsilon), \tag{3}$$

and estimate $\log(1-\gamma)$, $\log(1-\eta)$, and $\log(1-\varepsilon)$ by regression with the additive constant set equal to zero. We also may use some other approach to estimation such as minimum χ^2 or maximum likelihood. An analytic solution of the standard maximum-likelihood equations is very messy indeed, but the maximum of the likelihood function can be found numerically.

The automaton model naturally suggests a more detailed analysis of the data. Unlike the regression model, the automaton provides an immediate analysis of the digit-by-digit responses. Ignoring the $\varepsilon\delta$-type terms, we can in fact find the general maximum-likelihood estimates of γ, ε, and η when the response data are given in this more explicit form.

Let there be n digit responses in a block of problems. For $1 \leqslant i \leqslant n$ let $\mathbf{x}_i$ be the random variable that assumes the value 1 if the ith response is correct and 0 otherwise. It is then easy to see that

$$P(\mathbf{x}_i = 1) = \begin{cases} (1-\gamma) & \text{if } i \text{ is a ones'-column digit} \\ (1-\gamma)(1-\varepsilon) & \text{if it is not a ones' column and there is no carry to the } i\text{th digit} \\ (1-\gamma)(1-\eta) & \text{if there is a carry to the } i\text{th digit,} \end{cases}$$

granted that $\varepsilon\delta$-type terms are ignored. Similarly for the same three alternatives

$$P(\mathbf{x}_i = 0) = \begin{cases} \gamma \\ 1-(1-\gamma)(1-\varepsilon) \\ 1-(1-\gamma)(1-\eta). \end{cases}$$

So for a string of actual digit responses $x_1, \ldots, x_n$ we can write the likelihood function as:

$$(4) \qquad L(x_1, \ldots, x_n) = (1-\gamma)^a \gamma^b (1-\varepsilon)^c (1-\eta)^d \times [1-(1-\gamma)(1-\varepsilon)]^e [1-(1-\gamma)(1-\eta)]^f,$$

where a = number of correct responses, b = number of incorrect responses in the ones' column, c = number of correct responses not in the ones' column when the internal state is 0, d = number of correct responses when the internal state is 1, e = number of incorrect responses not in the ones' column when the internal state is 0, and f = number of incorrect responses when the internal state is 1. In the model statistical independence of responses is assured by the correction procedure. It is more convenient to estimate $\gamma' = 1-\gamma$, $\varepsilon' = 1-\varepsilon$, and $\eta' = 1-\eta$. Making this change, taking the log of both sides of (4) and differentiating with respect to each of the variables, we obtain three equations that determine the maximum-

likelihood estimates of γ', ε', and η':

$$\frac{\partial L}{\partial \gamma'} = \frac{a}{\gamma'} - \frac{b}{1-\gamma'} - \frac{e\varepsilon'}{1-\gamma'\varepsilon'} - \frac{f\eta'}{1-\gamma'\eta'} = 0,$$

$$\frac{\partial L}{\partial \varepsilon'} = \frac{c}{\varepsilon'} - \frac{e\gamma'}{1-\gamma'\varepsilon'} = 0,$$

$$\frac{\partial L}{\partial \eta'} = \frac{d}{\eta'} - \frac{f\gamma'}{1-\gamma'\eta'} = 0.$$

Solving these equations, we obtain as estimates:

$$\hat{\gamma}' = \frac{a-c-d}{a+b-c-d},$$

$$\hat{\varepsilon}' = \frac{c(a+b-c-d)}{(c+e)(a-c-d)},$$

$$\hat{\eta}' = \frac{d(a+b-c-d)}{(d+f)(a-c-d)}.$$

The most interesting feature of these estimates is that $\hat{\gamma}'$ is just the ratio of correct responses to total responses in the ones' column. The two equations that yield estimates of ε' and η' are especially transparent if they are rewritten:

$$(1-\gamma)(1-\varepsilon) = \gamma'\varepsilon' = c/(c+e),$$
$$(1-\gamma)(1-\eta) = \gamma'\eta' = d/(d+f).$$

Additional analysis of this example will not be pursued here. I do want to note that the internal states 0 and 1 are easily externalized as oral responses and most teachers do indeed require such externalization at the beginning.

To many readers the sort of probabilistic automaton just analyzed will not seem to be the sort of device required to account for language behavior. Certainly the automata that are adequate to analyze arithmetic are simpler in structure than what is needed even for the language output of a two-year-old child. On the other hand, I have already begun constructing probabilistic automata that will generate the language output of young

children and the preliminary results are far from discouraging. Probabilistic automata and their associated probabilistic grammars seem likely to be the right devices for such language analysis.

NOTES

[1] Indeed, Phoebe C. E. Diebold and Ebbey Bruce Ebbesen have already successfully trained two pigeons. Sequences of several hundred responses are easily obtained, and the error rate is surprisingly low – well under 1 %.

[2] Exactly such a procedure has proved very successful in the pigeon experiments with Diebold and Ebbesen mentioned in note 1.

[3] Two other points about the definition are the following. First, in this definition and throughout the rest of the article e_0 is the reinforcement of response r_0, and not the null reinforcement, as in Section III. This conflict arises from the different notational conventions in mathematical learning theory and automata theory. Second, strictly speaking I should write $\mathfrak{A}_F(\mathscr{S})$ because F is not uniquely determined by $\mathscr{S}$.

[4] After this was written Gordon Bower brought to my attention the article by Millenson (1967) that develops this point informally.

REFERENCES

Adams, E. W.: 1959, 'The Foundations of Rigid Body Mechanics and the Derivation of its Laws from those of Particle Mechanics', in L. Henkin, P. Suppes, and A. Tarski (eds.), *The Axiomatic Method*, North-Holland Publ. Co., Amsterdam, pp. 250–265.

Allais, M.: 1952, *Traité d'économie pure*, 2nd. ed., Imprimerie Nationale, Paris.

Arrow, K. J.: 1951, *Social Choice and Individual Values*, Wiley, New York.

Atkinson, R. C. and R. M. Shiffrin: 1968, 'Human Memory: A Proposed System and its Control Processes', in K. W. Spence and J. T. Spence (eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, Academic Press, New York, pp. 89–195.

Atkinson, R. C. and P. Suppes: 1958, 'An Analysis of Two-Person Game Situations in Terms of Statistical Learning Theory', *Journal of Experimental Psychology* **55**, 369–378.

Baker, G. A.: 1958, 'Formulation of Quantum Mechanics based on the Quasi-Probability Distribution induced on Phase Space', *Physical Review* **109**, 2198–2206.

Banach, S.: 1932, *Théorie des opérations linéaires*, Subwencji Funcuszu Kultury Narodowej, Warsaw.

Barker, S.: 1957, *Induction and Hypothesis*, Cornell University Press, Ithaca, N.Y.

Birkhoff, G.: 1948, *Lattice Theory*, Rev. ed., American Mathematical Society, New York.

Birkhoff, G. and S. MacLane: 1941, *A Survey of Modern Algebra*, Macmillan, New York.

Birkhoff, G. and J. von Neumann: 1936, 'The Logic of Quantum Mechanics', *Annals of Mathematics* **37**, 823–843.

Blackwell, D. and M. A. Girshick: 1954, *Theory of Games and Statistical Decisions*, Wiley, New York.

Bower, G. H.: 1961, 'Application of a Model to Paired-Associate Learning', *Psychometrica* **26**, 255–280.

Braithwaite, R. B.: 1955, *Theory of Games as a Tool for the Moral Philosopher*, Cambridge University Press, Cambridge.

Bruner, J. S., J. J. Goodnow, and G. A. Austin: 1956, *A Study of Thinking*, Wiley, New York.

Burke, C. J. and W. K. Estes: 1957, 'A Component Model for Stimulus Variables in Discrimination Learning', *Psychometrika* **22**, 133–145.

Burke, C. J., W. K. Estes, and S. Hellyer: 1954, 'Rate of Verbal Conditioning in Relation to Stimulus Variability', *Journal of Experimental Psychology* **48**, 153–161.

Bush, R. R. and W. K. Estes (eds.): 1959, *Studies in Mathematical Learning Theory*, Stanford University Press, Stanford, Calif.

Campbell, N. R.: 1920, *Physics the Elements*, Cambridge University Press, Cambridge.

Campbell, N. R.: 1928, *An Account of the Principles of Measurement and Calculation*, Longmans, Green, London & New York.

Carnap, R.: 1936, 1937, 'Testability and Meaning', *Philosophy of Science* **3**, 419–471; and **4**, 1–40.

Carnap, R.: 1950, *Logical Foundations of Probability*, University of Chicago Press, Chicago.

Chernoff, H.: 1954, 'Rational Selection of Decision Functions', *Econometrica* **22**, 422–443.

Chisholm, R. M.: 1957, *Perceiving: A Philosophical Study*, Cornell University Press, New York.

Chomsky, N.: 1959, Review of B. F. Skinner, *Verbal behavior*, *Language* **35**, 26–58.

Chomsky, N.: 1963, 'Formal Properties of Grammars', in R. D. Luce, R. R. Bush, and E. Galanter (eds.), *Handbook of Mathematical Psychology*, Vol. 2, Wiley, New York, pp. 125–155.

Chomsky, N.: 1965, *Aspects of the Theory of Syntax*, Massachusetts Institute of Technology Press, Cambridge, Mass.

Coombs, C. H.: 1950, 'Psychological Scaling without a Unit of Measurement', *Psychological Review* **57**, 145–158.

Coombs, C. H. and D. C. Beardslee: 1954, 'On Decision-Making under Uncertainty', in R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), *Decision Processes*, Wiley, New York, pp. 255–286.

Copi, I.: 1954, *Symbolic Logic*, Macmillan, New York.

Crothers, E. and P. Suppes: 1967, *Experiments in Second-Language Learning*, Academic Press, New York.

Davidson, D. and J. Marschak: 1959, 'Experimental Tests of a Stochastic Decision Theory', in C. W. Churchman and P. Ratoosh (eds.), *Measurement: Definition and Theories*, Wiley, New York, pp. 233–269.

Davidson, D. and P. Suppes: 1955, Finitistic Rational Choice Structures. Report No. 3, Stanford University, Stanford Value Theory Project, February.

Davidson, D. and P. Suppes: 1956, 'A Finitistic Axiomatization of Subjective Probability and Utility', *Econometrica* **24**, 264–275.

Davidson, D., J. C. C. McKinsey, and P. Suppes: 1954, Outlines of a Formal Theory of Value, I. Report No. 1, Stanford University, Stanford Value Theory Project, February 10. Published in *Philosophy of Science* **22** (1955), 140–160.

Davidson, D., S. Siegel, and P. Suppes: 1955, Some Experiments and Related Theory on the Measurement of Utility and Subjective Probability. Report No. 4, Stanford University, Stanford Value Theory Project, August. Published as Chapter 2 of *Decision-Making: An Experimental Report*, Stanford University Press, Stanford, Calif., 1957, pp. 19–81.

Davidson, D., P. Suppes, and S. Siegel: 1957, *Decision Making: An Experimental Approach*, Stanford University Press, Stanford, Calif.

Debreu, G.: 1958, 'Stochastic Choice and Cardinal Utility', *Econometrica* **26**, 440–444.

Debreu, G.: 1959, *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, Wiley, New York.

de Finetti, B.: 1937, 'La prévision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré* **7**, 1–68. English translation in H. E. Kyburg, Jr., and H. E. Smokler (eds.), *Studies in Subjective Probability*, Wiley, New York, 1964, pp. 93–158.

Dirac, P. A. M.: 1945, 'On the Analogy between Classical and Quantum Mechanics', *Reviews of Modern Physics* **17**, 195–199.

Doob, J. L.: 1960, 'Some Problems concerning the Consistency of Mathematical

Models', in R. E. Machol (ed.), *Information and Decision Processes*, McGraw-Hill, New York, pp. 27–33.
Edwards, W.: 1954, 'The Theory of Decision Making', *Psychological Bulletin* **51**, 380–417.
Estes, W. K.: 1950, 'Toward a Statistical Theory of Learning', *Psychological Review* **57**, 94–107.
Estes, W. K.: 1954, 'Individual Behavior in Uncertain Situations: An Interpretation in Terms of Statistical Association Theory', in R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), *Decision Processes*, Wiley, New York, pp. 127–137.
Estes, W. K.: 1957, 'Of Models and Men', *American Psychologist* **12**, 609–617.
Estes, W. K.: 1959, 'The Statistical Approach to Learning Theory', in S. Koch (ed.), *Psychology: A Study of a Science*', Vol. 2, McGraw-Hill, New York, pp. 380–491.
Estes, W. K.: 1960, 'Learning Theory and the New Mental Chemistry', *Psychological Review* **67**, 207–223.
Estes, W. K. and C. J. Burke: 1953, 'A Theory of Stimulus Variability in Learning', *Psychological Review* **60**, 276–286.
Estes, W. K. and P. Suppes: 1959a, 'Foundations of Linear Models', in R. R. Bush and W. K. Estes (eds.), *Studies in Mathematical Learning Theory*, Stanford University Press, Stanford, Calif., pp. 137–179.
Estes, W. K. and P. Suppes: 1959b, Foundations of Statistical Learning Theory, II: The Stimulus Sampling Model. Technical Report No. 26, Stanford University, Institute for Mathematical Studies in the Social Sciences, October 22.
Feynman, R. P.: 1948, 'Space-Time Approach to Non-Relativistic Quantum Mechanics', *Reviews of Modern Physics* **20**, 367–387.
Fine, A.: 1968, 'Logic, Probability, and Quantum Theory', *Philosophy of Science* **35**, 101–111.
Fisher, I.: 1927, 'A Statistical Method for Measuring "Marginal Utility" and Testing the Justice of a Progressive Income Tax', in J. H. Hollander (ed.), *Economic Essays Contributed in Honor of John Bates Clark*, Macmillan, New York, pp. 157–193.
Friedman, M. and L. J. Savage: 1952, 'The Expected Utility Hypothesis and the Measurability of Utility', *Journal of Political Economy* **60**, 463–474.
Frisch, R.: 1932, 'New Methods of Measuring Marginal Utility', in *Beiträge zur ökonomischen Theorie*, J. C. B. Mohr (Paul Siebeck), Tübingen, pp. 1–142.
Frisch, R.: 1937, 'General Choice-Field Theory', in *Report of Third Annual Research Conference on Economics and Statistics*, Cowles Commission for Research in Economics, pp. 64–69.
Gaifman, H.: 1964, 'Concerning Measures in First-Order Calculi', *Israel Journal of Mathematics* **2**, 1–18.
Goodman, N.: 1951, *The Structure of Appearance*, Harvard University Press, Cambridge, Mass.
Guilford, J. P.: 1936, *Psychometric Methods*, McGraw-Hill, New York.
Hailperin, T.: 1954, 'Remarks on Identity and Description in First-Order Axiom Systems', *Journal of Symbolic Logic* **19**, 14–20.
Hanes, R. M.: 1949, 'The Construction of Subjective Brightness Scales from Fractionation Data: A Validation', *Journal of Experimental Psychology* **39**, 719–728.
Hare, R. M.: 1952, *The Language of Morals*, Oxford University Press, Oxford.
Hempel, C. G.: 1965, *Aspects of Scientific Explanation*, The Free Press, New York.
Hermes, H.: 1938, *Eine Axiomatisierung der allgemeinen Mechanik* (Forschungen zur

Logik und zur Grundlegung der exakten Wissenschaften, Neue Folge, Heft 3), S. Hirzel, Leipzig.

Hill, S. A.: 1961, A Study of the Logical Abilities of Children. Unpublished doctoral dissertation, Stanford University.

Hoelder, O.: 1901, 'Die Axiome der Quantitaet und die Lehre vom Mass', *Berichte über die Verhandlungen der Königlichen Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53**, 1–64.

Hull, C. L. and K. W. Spence: 1938, 'Correction vs. Non-Correction Method of Trial-and-Error Learning in Rats', *Journal of Comparative Psychology* **25**, 127–145.

Jeffrey, R. C.: 1965, *The Logic of Decision*, McGraw-Hill, New York.

Kant, I.: 1949a, *Critique of Pure Reason* (translated by M. Muller), 2nd ed., revised, Macmillan, New York.

Kant, I.: 1949b, *Critique of Practical Reason and Other Writings in Moral Philosophy* (translated and edited by L. W. Beck), University of Chicago Press, Chicago.

Katz, J. J. and P. M. Postal: 1964, *An Integrated Theory of Linguistic Descriptions*, Massachusetts Institute of Technology Press, Cambridge, Mass.

Kemeny, J. G.: 1951, Review of R. Carnap, *Logical Foundations of Probability*. *Journal of Symbolic Logic* **16**, 205–207.

Kemeny, J. G.: 1953, 'A Logical Measure Function', *Journal of Symbolic Logic* **18**, 289–308.

Khinchin, A. I.: 1949, *Statistical Mechanics*, Dover, New York.

Kleene, S. C.: 1956, 'Representation of Events in Nerve Nets and Finite Automata', in C. E. Shannon and J. McCarthy (eds.), *Automata Studies*, Princeton University Press, Princeton, N. J., pp. 3–41.

Kochen, S. and E. P. Specker: 1965, 'Logical Structures arising in Quantum Theory', in J. W. Addison, L. Henkin, and A. Tarski (eds.), *Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, North-Holland Publ. Co., Amsterdam, pp. 177–189.

Koopman, B. O.: 1957, 'Quantum Theory and the Foundations of Probability', in *Proceedings of Symposia in Applied Mathematics*, Vol. 7: *Applied Probability*, McGraw-Hill, New York, pp. 97–102.

Krantz, D. H.: 1967, 'Extensive Measurement in Semiorders', *Philosophy of Science* **34**, 348–362.

Lakatos, I. (ed.): 1968, *The Problem of Inductive Logic*, North-Holland Publ. Co., Amsterdam.

Landau, L. D. and E. M. Lifshitz: 1958, *Quantum Mechanics: Non-Relativistic Theory*, Pergamon Press, London.

Lange, O.: 1934, 'The Determinateness of the Utility Function', *Review of Economic Studies* **1**, 218–225.

Lange, O. and F. M. Taylor: 1938, *On the Economic Theory of Socialism* (edited by B. Lippincott), University of Minnesota Press, Minneapolis.

Levi, Isaac: 1967, Review of J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic. British Journal of the Philosophy of Science* **19**, 73–83.

Lindsay, R. B. and H. Margenau: 1936, *Foundations of Physics*, Wiley, New York.

Luce, R. D.: 1956, 'Semiorders and a Theory of Utility Discrimination', *Econometrica* **24**, 178–191.

Luce, R. D.: 1959, *Individual Choice Behavior*, Wiley, New York.

Luce, R. D. and A. A. J. Marley: 1969, 'Extensive Measurement when Concatenation is Restricted and Maximal Elements may Exist', in S. Morgenbesser, P. Suppes, and

M. G. White (eds.), *Philosophy, Science and Method: Essays in Honor of Ernest Nagel*, St. Martin's Press, New York, in press.

Luce, R. D. and H. Raiffa: 1957, *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.

Luce, R. D. and P. Suppes: 1965, 'Preference, Utility, and Subjective Probability', in R. D. Luce, R. R. Bush, and E. H. Galanter (eds.), *Handbook of Mathematical Psychology*, Wiley, New York, pp. 249–410.

Mackey, G. W.: 1957, 'Quantum Mechanics and Hilbert Space', *The American Mathematical Monthly* **64**, 45–57.

Mackey, G. W.: 1963, *Mathematical Foundations of Quantum Mechanics*, W. A. Benjamin, Inc., New York.

Margenau, H.: 1963, 'Measurements and Quantum States, Part II', *Philosophy of Science* **30**, 138–157.

Margenau, H. and R. N. Hill: 1961, 'Correlation between Measurements in Quantum Theory', *Progress of Theoretical Physics* **26**, 722–738.

Marschak, J.: 1950, 'Rational Behavior, Uncertain Prospects, and Measurable Utility', *Econometrica* **18**, 111–141.

McKinsey, J. C. C.: 1935, 'On the Independence of Undefined Ideas', *Bulletin of the American Mathematical Society* **41**, 291–297.

McKinsey, J. C. C. and P. Suppes: 1955, 'On the Notion of Invariance in Classical Mechanics', *The British Journal for the Philosophy of Science* **5**, 290–302.

Millenson, J. R.: 1967, 'An Isomorphism between Stimulus-Response Notation and Information Processing Flow Diagrams', *The Psychological Record* **17**, 305–319.

Miller, G. A. and N. Chomsky: 1963, 'Finitary Models of Language Users', in R. D. Luce, R. R. Bush, and E. Galanter (eds.), *Handbook of Mathematical Psychology*, Vol. 2, Wiley, New York, pp. 419–492.

Miller, G. A., E. Galanter, and K. H. Pribram: 1960, *Plans and the Structure of Behavior*, Holt, New York.

Milnor, J.: 1954, 'Games Against Nature', in R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), *Decision Processes*, Wiley, New York, pp. 49–59.

Mosteller, F. and P. Nogee: 1951, 'An Experimental Measurement of Utility', *Journal of Political Economy* **59**, 371–404.

Moyal, J. E.: 1949, 'Quantum Mechanics as a Statistical Theory', *Proceedings of the Cambridge Philosophical Society* **45**, 99–124.

Murphy, J. V. and R. E. Miller: 1955, 'The Effect of Spatial Contiguity of Cue and Reward in the Object-Quality Learning of Rhesus Monkeys', *Journal of Comparative and Physiological Psychology* **48**, 221–229.

Murphy, J. V. and R. E. Miller: 1959, 'Spatial Contiguity of Cue, Reward and Response in Discrimination Learning by Children', *Journal of Experimental Psychology* **58**, 485–489.

Nagel, E.: 1931, 'Measurements', *Erkenntnis* **2**, 313–333.

Nash, J. F.: 1950, 'The Bargaining Problem', *Econometrica* **18**, 155–162.

Nash, J. F.: 1951, 'Non-Cooperative Games', *Annals of Mathematics* **54**, 286–295.

Neimark, E. and W. K. Estes (eds.): 1967, *Stimulus Sampling Theory*, Holden-Day, San Francisco.

Newell, A. and H. A. Simon: 1956, 'The Logic Theory Machine', *IRE Transactions on Information Theory*, IT-2 **2**, 61–79.

Newell, A., J. C. Shaw, and H. A. Simon: 1957, 'Empirical Explorations of the Logic Theory Machine', in *Proceedings of the 1957 Western Joint Computer Conference,*

Los Angeles, February, 1957, Institute of Radio Engineers, New York, pp. 218–240.

Newton, I.: 1769, *Universal Arithmetick* (translated by Mr. Raphson, revised and corrected by Mr. Cunn), J. Senex, London.

Noll, W.: 1959, 'The Foundations of Classical Mechanics in the Light of Recent Advances in Continuum Mechanics', in L. Henkin, P. Suppes, and A. Tarski (eds.), *The Axiomatic Method*, North-Holland Publ. Co., Amsterdam, pp. 266–281.

Noll, W.: 1964, 'Euclidean Geometry and Minkowskian Chronometry', *The American Mathematical Monthly* **71**, 129–143.

Padoa, A.: 1901, 'Essai d'une théorie algébrique des nombres entiers, précédé d'une introduction logique à une théorie déductive quelconque', *Bibliothèque du Congrès International de Philosophie 1900, Paris* **3**, 309–365.

Papandreou, A. G.: 1957, *A Test of a Stochastic Theory of Choice*, University of California Press, Berkeley-Los Angeles.

Peters, R. S.: 1958, *The Concept of Motivation*, Routledge & Kegan Paul, London.

Peterson, M. J., F. B. Colavita, D. B. Sheahan, and K. D. Blattner: 1964, 'Verbal Mediating Chains and Response Availability as a Function of the Acquisition Paradigm', *Journal of Verbal Learning and Verbal Behavior* **3**, 11–18.

Quine, W. V.: 1950, *Methods of Logic*, Holt, New York.

Rabin, M. O.: 1963, 'Probabilistic Automata', *Information and Control* **6**, 230–245. Reprinted in E. F. Moore (ed.), *Sequential Machines*, Addison-Wesley, Reading, Mass., 1964, pp. 98–114.

Rabin, M. O. and D. Scott: 1959, 'Finite Automata and their Decision Problems', *IBM Journal of Research and Development* **3**, 114–125. Reprinted in E. F. Moore (ed.), *Sequential Machines*, Addison-Wesley, Reading, Mass., 1964, pp. 63–91.

Ramsey, F. P.: 1931, *The Foundations of Mathematics and Other Logical Essays*, Harcourt Brace, New York.

Reichenbach, H.: 1944, *Philosophic Foundations of Quantum Mechanics*, University of California Press, Berkeley, Calif.

Restle, F.: 1961, 'Statistical Methods for a Theory of Cue Learning', *Psychometrika* **26**, 291–306.

Restle, F.: 1964, 'Sources of Difficulty in Learning Paired-Associates', in R. C. Atkinson (ed.), *Studies in Mathematical Psychology*, Stanford University Press, Stanford, Calif., pp. 116–172.

Robb, A. A.: 1936, *Geometry of Space and Time*, Cambridge University Press, Cambridge.

Rubin, H.: 1949a, 'An Axiomatic Approach to Integration', *Bulletin of the American Mathematical Society* **55**, 1064. (Abstract)

Rubin, H.: 1949b, 'Measures and Axiomatically Defined Integrals', *Bulletin of the American Mathematical Society* **55**, 1064. (Abstract)

Rubin, H.: 1954, 'Postulates for Rational Behavior under Uncertainty'. Unpublished manuscript.

Rubin, H.: 1959, 'On the Foundations of Quantum Mechanics', in L. Henkin, P. Suppes, and A. Tarski (eds.), *The Axiomatic Method with Special Reference to Geometry and Physics*, North-Holland Publ. Co., Amsterdam, pp. 333–340.

Rubin, H. and P. Suppes: 1954, 'Transformations of Systems of Relativistic Particle Mechanics', *Pacific Journal of Mathematics* **4**, 563–601.

Rubin, H. and P. Suppes: 1955, 'A Note on Two-Place Predicates and Fitting Sequences of Measure Functions', *Journal of Symbolic Logic* **20**, 121–122.

Russell, B.: 1903, *Principles of Mathematics*, G. Allen & Unwin, London.
Samuelson, P. A.: 1947, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, Mass.
Samuelson, P. A.: 1952, 'Probability, Utility, and the Independence Axiom', *Econometrica* **20**, 670–678.
Savage, L. J.: 1954, *The Foundations of Statistics*', Wiley, New York.
Schiff, L. I.: 1949, *Quantum Mechanics*, McGraw-Hill, New York.
Scott, D.: 1964, 'Measurement Structures and Linear Inequalities', *Journal of Mathematical Psychology* **1**, 233–247.
Scott, D. and P. Krauss: 1966, 'Assigning Probabilities to Logical Formulas', in J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic*, North-Holland Publ. Co., Amsterdam.
Scott, D. and P. Suppes: 1958, 'Foundational Aspects of Theories of Measurement', *Journal of Symbolic Logic* **28**, 113–128. Reprinted as Article 4 in this volume.
Scriven, M.: 1965, 'An Essential Unpredictability in Human Behavior', in E. Nagel and B. B. Wohlman (eds.), *Scientific Psychology*, Basic Books, New York, pp. 411–425.
Shannon, C. E. and W. Weaver: 1949, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill.
Sierpinski, W.: 1934, *Hypothèse du continu*, Warsaw and Lwów.
Simon, H. A.: 1957, *Models of Man*, Wiley, New York.
Skinner, B. F.: 1957, *Verbal Behavior*, Appleton, New York.
Sneed, J. D.: 1966, 'Von Neumann's Argument for the Projection Postulate', *Philosophy of Science* **33**, 22–39.
Stevens, S. S.: 1936, 'A Scale for the Measurement of a Psychological Magnitude; Loudness', *Psychological Review* **43**, 405–416.
Stevens, S. S. and J. Volkmann: 1940, 'The Relation of Pitch to Frequency: A Revised Scale', *American Journal of Psychology* **53**, 329–353.
Stoll, E.: 1962, Geometric Concept Formation in Kindergarten Children. Unpublished doctoral dissertation, Stanford University.
Suppes, P.: 1951, 'A Set of Independent Axioms for Extensive Quantities', *Portugaliae Mathematica* **10**, 163–172. Reprinted as Article 3 in this volume.
Suppes, P.: 1956a, A Set of Axioms for Paired Comparisons. Stanford University, Center for Behavioral Sciences.
Suppes, P.: 1956b, 'The Role of Subjective Probability and Utility in Decision-Making', *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5**, 61–73. Reprinted as Article 6 in this volume.
Suppes, P.: 1957, *Introduction to Logic*, Van Nostrand, Princeton, N.J.
Suppes, P.: 1959a, 'Axioms for Relativistic Kinematics with or without Parity', in L. Henkin, P. Suppes, and A. Tarski (eds.), *The Axiomatic Method*, North-Holland Publ. Co., Amsterdam, pp. 291–307. Reprinted as Article 12 in this volume.
Suppes, P.: 1959b, 'A Linear Learning Model for a Continuum of Responses', in R. R. Bush and W. K. Estes (eds.), *Studies in Mathematical Learning Theory*, Stanford University Press, Stanford, Calif., pp. 400–414.
Suppes, P.: 1960a, 'Some Open Problems in the Foundations of Subjective Probability', in R. E. Machol (ed.), *Information and Decision Processes*, Wiley, New York, pp. 162–169.
Suppes, P.: 1960b, 'A Comparison of the Meaning and Uses of Models in Mathematics

and the Empirical Sciences', *Synthese* **12**, 287–301. Reprinted as Article 1 in this volume.

Suppes, P.: 1961a, 'Behavioristic Foundations of Utility', *Econometrica* **29**, 1–17. Reprinted as Article 9 in this volume.

Suppes, P.: 1961b, 'Probability Concepts in Quantum Mechanics', *Philosophy of Science* **28**, 378–389. Reprinted as Article 13 in this volume.

Suppes, P.: 1964, 'The Kinematics and Dynamics of Concept Formation', *Proceedings for the 1964 International Congress for Logic, Methodology and Philosophy of Science*, North-Holland Publ. Co., Amsterdam, pp. 405–414.

Suppes, P.: 1965a, Towards a Behavioral Foundation of Mathematical Proofs. Technical Report No. 44, Stanford University, Institute for Mathematical Studies in the Social Sciences, January 2, 1962. Published in K. Ajdukiewicz (ed.), *The Foundations of Statements and Decisions: Proceedings of the International Colloquium on Methodology of Science, September 18–23, 1961*, PWN-Polish Scientific Publishers, Warszawa, pp. 327–341. Reprinted as Article 20 in this volume.

Suppes, P.: 1965b, 'Logics Appropriate to Empirical Theories', in J. W. Addison, L. Henkin, and A. Tarski (eds.), *Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, North-Holland Publ. Co., Amsterdam, pp. 364–395.

Suppes, P.: 1966, 'Concept Formation and Bayesian Decisions', in J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic*, North-Holland Publ. Co., Amsterdam, pp. 21–48.

Suppes, P. and R. Atkinson: 1959, Markov Learning Models for Multiperson Situations, I: The Theory. Technical Report No. 21, Stanford University, Applied Mathematics and Statistics Laboratory, February 20. Published in *Markov learning models for multiperson situations*, Stanford University Press, Stanford, Calif., pp. 1–46.

Suppes, P. and R. C. Atkinson: 1960, *Markov Learning Models for Multiperson Interactions*, Stanford University Press, Stanford, Calif.

Suppes, P. and R. Ginsberg: 1962a, 'Application of a Stimulus Sampling Model to Children's Concept Formation with and without an Overt Correction Response', *Journal of Experimental Psychology* **63**, 330–336.

Suppes, P. and R. Ginsberg: 1962b, 'Experimental Studies of Mathematical Concept Formation in Young Children', *Science Education* **46**, 230–240.

Suppes, P. and R. Ginsberg: 1963, 'A Fundamental Property of All-or-None Models, Binomial Distribution of Responses Prior to Conditioning, with Application to Concept Formation in Children'. Technical Report No. 39. Stanford University, Institute for Mathematical Studies in the Social Sciences, September 20, 1961. Published in *Psychological Review* **70**, 139–161.

Suppes, P. and S. Hill: 1964, *First Course in Mathematical Logic*, Blaisdell, New York.

Suppes, P. and M. Winet: 1955, 'An Axiomatization of Utility based on the Notion of Utility Differences', *Journal of Management Science* **1**, 259–270. Reprinted as Article 8 in this volume.

Suppes, P. and J. Zinnes: 1963, 'Basic Measurement Theory', in R. D. Luce, R. R. Bush, and E. H. Galanter (eds.), *Handbook of Mathematical Psychology*, Vol. 1, Wiley, New York, pp. 3–76.

Suppes, P., L. Hyman, and M. Jerman: 1967, 'Linear Structural Models for Response and Latency Performance in Arithmetic on Computer-Controlled Terminals', in J. P. Hill (ed.), *Minnesota Symposia on Child Psychology*, University of Minnesota Press, Minneapolis, pp. 160–200.

Suppes, P., M. Jerman, and D. Brian: 1968, *Computer-Assisted Instruction: Stanford's 1965–66 Arithmetic Program*, Academic Press, New York.

Tait, W. W.: 1959, 'A Counterexample to a Conjecture of Scott and Suppes', *Journal of Symbolic Logic* **24**, 15–16.

Tarski, A.: 1951, *A Decision Method for Elementary Algebra and Geometry*, 2nd ed., University of California Press, Berkeley, Calif.

Tarski, A.: 1953, 'A General Method in Proofs of Undecidability', in A. Tarski, A. Mostowski, and R. M. Robinson (eds.), *Undecidable Theories*, North-Holland Publ. Co., Amsterdam, pp. 3–35.

Tarski, A.: 1954, 1955, 'Contributions to the Theory of Models, I, II, III', *Indagationes Mathematicae* **16**, 572–581, 582–588; and **17**, 56–64.

Tarski, A.: 1930, 'Über einige fundamentale Begriffe der Metamathematik', *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie* **23**, 22–29. Reprinted in A. Tarski, *Logic, Semantics, Metamathematics*, Oxford University Press, Oxford, 1956.

Varadarajan, V. S.: 1962, 'Probability in Physics and a Theorem on Simultaneous Observability', *Communications on Pure and Applied Mathematics* **15**, 189–217.

Vaught, R.: 1954, 'Remarks on Universal Classes of Relational Systems', *Indagationes Mathematicae* **16**, 589–591.

von Neumann, J. and O. Morgenstern: 1947, *Theory of Games and Economic Behavior*, 2nd ed., Princeton University Press, Princeton, N. J.

Weyl, H.: 1922, *Space-Time-Matter*, 4th ed., Methuen, London.

Weyl, H.: 1931, *The Theory of Groups and Quantum Mechanics* (translated by H. P. Robertson), Dover, New York.

Wiener, N.: 1919–1920, 'A New Theory of Measurement', *Proceedings of the London Mathematical Society* **19**, 181–205.

Wigner, E.: 1932, 'On the Quantum Correction for Thermodynamic Equilibrium', *Physical Review* **40**, 749–759.

Woodger, J. H.: 1957, *The Axiomatic Method in Biology*, Cambridge University Press, Cambridge.

Yosida, K. and E. Hewitt: 1952, 'Finitely Additive Measures', *Transactions of the American Mathematical Society* **72**, 46–66.

Zeeman, E. C.: 1964, 'Causality Implies the Lorentz Group', *Journal of Mathematical Physics* **4**, 490–493.

INDEX OF NAMES

INDEX OF SUBJECTS

SYNTHESE LIBRARY

Monographs on Epistemology, Logic, Methodology,
Philosophy of Science, Sociology of Science and of Knowledge, and on the
Mathematical Methods of Social and Behavioral Sciences

Editors:

‡D. DAVIDSON and J. HINTIKKA: (eds.), *Words and Objections: Essays on the Work of W. V. Quine*. 1969, VIII+366 pp. Dfl. 48.—

‡J. W. DAVIS, D. J. HOCKNEY, and W. K. WILSON (eds.), *Philosophical Logic*. 1969, VIII + 277 pp. Dfl. 45.—

‡ROBERT S. COHEN and MARX W. WARTOFSKY (eds.), *Boston Studies in the Philosophy of Science*. Volume V: *Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968*. 1969, VIII + 482 pp. Dfl. 58.—

‡ROBERT S. COHEN and MARX W. WARTOFSKY (eds.), *Boston Studies in the Philosophy of Science*. Volume IV: *Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968*. 1969, VIII + 537 pp. Dfl. 69.—

‡NICHOLAS RESCHER, *Topics in Philosophical Logic*. 1968, XIV + 347 pp. Dfl 62.—

‡GÜNTHER PATZIG, *Aristotle's Theory of the Syllogism. A Logical-Philological Study of Book A of the Prior Analytics*. 1968, XVII + 215 pp. Dfl. 45.—

‡C. D. BROAD, *Induction, Probability, and Causation. Selected Papers*. 1968, XI + 296 pp. Dfl. 48.—

‡ROBERT S. COHEN and MARX W. WARTOFSKY (eds.), *Boston Studies in the Philosophy of Science*. Volume III: *Proceedings of the Boston Colloquium for the Philosophy of Science 1964/1966*. 1967, XLIX + 489 pp. Dfl. 65.—

‡GUIDO KÜNG, *Ontology and the Logistic Analysis of Language. An Enquiry into the Contemporary Views on Universals*. 1967, XI + 210 pp. Dfl. 34.—

p.t.o.

*EVERT W. BETH and JEAN PIAGET, *Mathematical Epistemology and Psychology*. 1966, XXII + 326 pp. Dfl. 54.—

*EVERT W. BETH, *Mathematical Thought. An Introduction to the Philosophy of Mathematics*. 1965, XII + 208 pp. Dfl. 30.—

‡PAUL LORENZEN, *Formal Logic*. 1965, VIII + 123 pp. Dfl. 18.75

‡GEORGES GURVITCH, *The Spectrum of Social Time*. 1964, XXVI + 152 pp. Dfl. 20.—

‡A. A. ZINOV'EV, *Philosophical Problems of Many-Valued Logic*. 1963, XIV + 155 pp. Dfl. 23.—

‡MARX W. WARTOFSKY (ed.), *Boston Studies in the Philosophy of Science*. Volume I: *Proceedings of the Boston Colloquium for the Philosophy of Science, 1961–1962*. 1963, VIII + 212 pp. Dfl. 22.50

‡B. H. KAZEMIER and D. VUYSJE (eds.), *Logic and Language. Studies dedicated to Professor Rudolf Carnap on the Occasion of his Seventieth Birthday*. 1962, VI + 246 pp. Dfl. 24.50

*EVERT W. BETH, *Formal Methods. An Introduction to Symbolic Logic and to the Study of Effective Operations in Arithmetic and Logic*. 1962, XIV + 170 pp. Dfl. 23.50

*HANS FREUDENTHAL (ed.), *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences. Proceedings of a Colloquium held at Utrecht, The Netherlands, January 1960*. 1961, VI + 194 pp. Dfl. 21.—

‡P. L. R. GUIRAUD, *Problèmes et méthodes de la statistique linguistique*. 1960, VI + 146 pp. Dfl. 15.75

* J. M. BOCHEŃSKI, *A Precis of Mathematical Logic*. 1959, X + 100 pp. Dfl. 15.75

Sole Distributors in the U.S.A. and Canada:

*GORDON & BREACH, INC., 150 Fifth Avenue, New York, N.Y. 10011
‡HUMANITIES PRESS, INC., 303 Park Avenue South, New York, N.Y. 10010