STRUCTURAL LEARNING

II. Issues and Approaches
An interdisciplinary series concerned with the following questions: (1) How can competence (knowledge) be characterized in a way which has behavioral relevance? (2) How can human knowledge be operationalized? (3) What are the mechanisms which govern learning? Performance? Motivation? Perception? Memory? (4) How can learning be optimized?

Cognitive, developmental, educational, and mathematical psychologists, philosophers (epistemology), computer scientists (simulation and artificial intelligence, theory of computation), generative linguists, mathematics educators, and those in educational technology will find the series particularly appealing.

1. SCANDURA • Structural Learning: I. Theory and Research
2. SCANDURA (ED) • Structural Learning: II. Issues and Approaches (Contributors: ARBIB • CORCORAN • DOMOTOR • GREENO • LOVELL • NEWELL • ROSENBLOOM • SHAW • SIMON • SUPPES • WITTROCK • WITZ)
PREFACE

The emerging (inter)discipline of structural learning is in part an indirect outgrowth of an invitational conference on structural learning held at the University of Pennsylvania on April 20-22, 1968. At that time a small group of scholars from a number of related fields was brought together in an attempt to forge new ground in a previously virgin territory. We were not interested in reporting finished research but, rather, we were unsatisfied with the kind of research being done on complex human learning and wanted to see what we could do about improving it. More particularly, we wanted to identify some of the basic problems involved and, if possible, come up with new ways of attacking these problems.

As it turned out, the conference was a much greater success than most of us ever dreamed it would be. The interaction—including not less than half-a-hundred spirited arguments—was lively indeed. This writer finds it impossible to capture in words the all-concentrative atmosphere. The clashing red of Corcoran's flushing face and thinning hair as he challenged my formulation of rule-governed behavior or as he reacted to Rosenbloom's reasoned criticism of his discussion of suppositional theories of proof can only be fully appreciated by those who were present. Wittrock's passioned defense of S-R mediation theories and the look in Greeno's eye when Rosenbloom described his mathematical golf were equally memorable. Perhaps the most classic comment to come out of the conference, however, were the two words spoken by R. Duncan Luce. In response to Corcoran's comment that it would take "fifty years to accomplish" something I said, Luce commented in skeptical overtones, "At least!"

Although the chapters included in this book were prepared after the conference, it is safe to say that all were influenced by the ideas expressed there. For some, the conference helped to shape in fundamental ways the very direction of their future research. Greeno's chapter in Volume II, for example, is a far more vivid account of what he was trying to say at the conference. Corcoran, too, readily acknowledges the value of some of the criticism leveled at his proposals. If it were not for the conference, it is unlikely that I would have even attempted to analyze mathematical structures, let alone attempt to devise a comprehensive theory of structural learning. Without the excellent criticism I received from more than one conferee, I might still be hung up in the unrewarding and seemingly unending cycle of criticizing S-R mediation theory.

Since then, five increasingly successful meetings on structural
learning have been held and a sixth is being planned as this book goes to press. A number of the contributions to these volumes were initiated at these meetings.

In the first volume in this series, *Structural Learning I: Theory and Research*, I described what I see as the rudiments of three deterministic partial theories of structural learning. The first involves competence, partial theories which deal only with the problem of how to account for the various kinds of behavior of which people are typically capable. Special attention is given to mathematical competence. Nothing is said about learning or performance. The second partial theory is concerned with motivation, learning, and performance under idealized conditions, and is obtained from the first partial theory by imposing further structure on it. This theory says nothing about memory or the limited capacity of human subjects to process information. It applies only where the subject has all of the necessary information readily available, and his capacity to process information is essentially unlimited (as when subjects have access to paper and pencil and/or other memory aids). The final theory is obtained from the second by making additional assumptions, which bring memory and finite information processing into the picture. The theory is still partial, however, since no attempt is made to deal with certain ultra-short-term behavioral phenomena which appear to depend directly on particular physiological characteristics (e.g., the short period of time images are presumably retained on the retina of the eye).

*Structural Learning II: Issues and Approaches* is an edited volume and reflects the major approaches currently being taken in structural learning. Chapter 1 deals with the basic question of whether competence (knowledge) should be characterized in terms of rules (automata), on the one hand, or associations on the other. The basic question is posed by me, and a formal resolution demonstrating the possible reduction of automata to associations is then presented by Suppes in a classic paper. (I extend this argument, showing the relationship between rules and associations.) In a series of reactions to this paper, Arbib points out certain limitations in Suppes' argument and Suppes replies. In the final section, I have argued that the reduction of rule learning to S-R conditioning is only meaningful if an adequate theory based on rules can be devised, and conversely, if a theory based on rules can be devised, then there is little point in reducing the theory to S-R conditioning.

The bulk of Chapter 2 is devoted to a series of earlier experiments on rule learning by my associates and me. Reacting to this research, Wittrock suggests the need for extension of the "model" and poses some empirical results of his own as a challenge for interpretation in terms of the model. In the final section, this challenge is accepted; it is also noted that the reported research was more a result of having available a suitable language within which to formulate empirical questions than a behavior theory per se.

The two contributions by Greeno in Chapter 3 deal with graph theoretical models. First, Greeno develops an alternative language for representing the more molar characteristics of structured knowledge. What have been termed rules in Volume I are represented in Greeno's formulation as arcs which map sets of points into points. The framework he introduces is an ingenious attempt to represent interrelationships among rules. The second part of the chapter is devoted to a report of two preliminary experiments which were directly motivated by Greeno's graph-theoretical formulation.
Piagetian models constitute the subject of Chapter 4. In the first article, Lovell reviews six basic ideas characteristic of the Piagetian Theory which he feels have particular relevance to mathematics education. Within this perspective, he reviews some of the evidence obtained from research, and finally deals with some implications for teaching. The final article by Witz is based on a more formal representation of Piagetian notions and describes an important structural change observed in one of his four-year-old subjects. This work appears to signal a significant new approach to Piagetian research.

Chapter 5 deals with attempts to simulate human behavior with a computer. The first article summarizes and brings up to date over 15 years of pioneering research in the area by Simon and Newell. In order that the reader not be carried away with enthusiasm, Shaw in the second article identifies a number of fundamental limitations of the entire simulation enterprise. In particular, he argues that the problems involved in understanding any program complex enough to simulate complex human behavior are apt to be as difficult as understanding the human behavior itself.

Chapter 6 ranges over a wide variety of competence models, with particular reference to logic and mathematics. In the first article, Rosenbloom describes some mathematical tasks for use in psychological research. These tasks, which Rosenbloom calls "mathematical golf", are suitable for use with young children and adults alike. Furthermore, they have all of the essential characteristics of formal mathematical systems and so bear directly on problems of mathematics learning. The next three articles by Corcoran deal with discourse grammars and the structure of logical reasoning. In part one, Corcoran describes certain parallels between mathematical reasoning and the structure of language. His arguments in part two concerning the nature of a "correct" theory of proof (or, equivalently, a theory of proof with behavioral relevance) have been to some extent incorporated in Volume I. In part three, Corcoran argues forcefully in favor of what he calls suppositional theories of proof (natural deductive systems) as opposed to linear theories.

In the next article, I have proposed a theory of mathematical knowledge, based on the idea that rules in a characterizing rule set may act on other rules in the set to produce new rules. Allowing rules to interact in this way accounts for what may be called "creative behavior". The final article by Domotor argues the case for the mathematical modeling of psychological phenomena. To advance his case, he proposes a rather sophisticated model for illustration. It should be noted that Domotor, more than the other authors in this section, has been heavily influenced by the Stanford emphasis on formalization.

In Chapter 7, I propose a new theory of structural learning, together with some empirical results. The first article describes the theory informally, in its preliminary formulation, together with some preliminary data which support it. Definitional foundations for competence theories and the idealized theory of behavior are presented in the second article. A basic tenet of this theory is that competence theories provide a measuring device of sorts against which to measure human knowledge. Behavior, then, amounts to putting knowledge to use and acquiring new knowledge, which takes place according to a specified set of mechanisms.

The original conferees are indebted to the Graduate School of Education, University of Pennsylvania, for providing partial support and
facilities for the original conference on mathematics and structural learning. Morris Viteles, former dean, and David Goddard, former provost, both had a hand in making funds available.

I would like to thank Diane Triman, John Durnin, and Wallace Wulfeck for their help in proving and correcting the final manuscript. Sincere thanks are also due Mary Tye for her usual excellent typing of the manuscript.

The book is dedicated to those who would use it as a basis for further research on structural learning.

Joseph M. Scandura
Philadelphia
CONTENTS

PREFACE vii

Chapter 1. BASIC UNIT IN STRUCTURAL LEARNING: ASSOCIATION OR AUTOMATON (RULE)? 1

A SET-FUNCTION LANGUAGE 3
Joseph M. Scandura

STIMULUS-RESPONSE THEORY OF FINITE AUTOMATA 11
Patrick Suppes

MEMORY LIMITATIONS OF STIMULUS-RESPONSE MODELS 19
Michael A. Arbib

STIMULUS-RESPONSE THEORY OF AUTOMATA AND TOTE HIERARCHIES: A REPLY TO ARBIB: 25
Patrick Suppes

A REPLY TO SUPPES' REPLY 29
Michael Arbib

S-R THEORY OR AUTOMATA? (A FINAL WORD) 31
Joseph M. Scandura

Chapter 2. NEW DIRECTIONS FOR RESEARCH ON RULE LEARNING 37

EMPIRICAL RESEARCH 39
Joseph M. Scandura

COMMENTS ON SCANDURA'S APPROACH TO RULE-GOVERNED BEHAVIOR 65
Merlin C. Wittrock

A REPLY TO WITTROCK 69
Joseph M. Scandura

Chapter 3. GRAPH THEORETIC MODELS 71

PSYCHOLOGICAL REPRESENTATION OF STRUCTURED KNOWLEDGE 73
James C. Greeno
Chapter 4. **PIagetian Models** 105

*Intellectual Growth and Understanding Mathematics* 107
  Kenneth Lovell

*Activity Structures in Four-Year Olds* 123
  Klaus G. Witz

Chapter 5. **Simulation Models** 129

*Human Problem Solving: The State of the Theory in 1970* 131
  Herbert A. Simon and Allen Newell

*CogNITION, SIMulAtION, And the ProBlEm of ComplexItY* 153
  Robert E. Shaw

Chapter 6. **Competence Models in Mathematics** 169

*Structural Models for Use in Psychological Research* 171
  Paul C. Rosenbloom

*Mathematical Reasoning and the Structure of Language* 179
  John Corcoran

*The Nature of a Correct Theory of Proof and Its Value* 195
  John Corcoran

*Two Theories of Proof* 207
  John Corcoran

*A Theory of Mathematical Knowledge: Can Rules Account for Creative Behavior* 225
  Joseph M. Scandura

*Formulating Mathematical Models of Psychological Phenomena* 237
  Zoltan Domotor

Chapter 7. **A Theory of Structural Learning** 249

*Deterministic Theorizing in Structural Learning: Three Levels of Empiricism* 251
  Joseph M. Scandura

*Structural Learning: Definitional Foundations for Competence Theories and the Idealized Theory of Behavior* 275
  Joseph M. Scandura
BASIC UNIT IN STRUCTURAL LEARNING: ASSOCIATION OR AUTOMATON (RULE)?
A SET-FUNCTION LANGUAGE

JOSEPH M. SCANDURA

The search for a suitable scientific language in psychology has had a long history. Unfortunately, as with theories, there is no a priori basis for deciding between alternatives. Which will prove most useful can be determined only after a period of use. Nonetheless, certain characteristics appear desirable. One of these is precision. The primary requirement, however, is that the language accurately represent the important characteristics of the phenomena in question. Without such fidelity the language can have no real value—it is the sine qua non.

In order to construct a precise descriptive language, which adequately reflects meaningful learning, a basic behavior unit must be selected. The history of science has shown that the hypothesis-generating and predictive value of any theory or scientific language is determined in large part by the appropriateness of its basic building blocks.

Many theorists have been primarily concerned with extending S-R formulations to account for complex phenomena (e.g., Berlyne, 1965; Kendler and Kendler, 1962; Maltzman, 1955; Goss, 1961; Osgood, 1953; Staats and Staats, 1963). Although it has been repeatedly emphasized that the S-R approach is simply a way of working, of baring essentials, the neo-associationist implicitly believes that the association provides the most precise and efficient unit with which to describe behavior.

Other theorists and highly reputable writers (e.g., Ausubel, 1963; Bartlett, 1932a, 1958; Dienes, 1963; Gagné, 1962, 1965; Mandler, 1962, 1965; Miller, Galanter and Pribram, 1960; Piaget, as described in Flavell, 1963, Polya, 1962, 1965; Newell, Shaw and Simon, 1958) feel that the S-R language does not capture the essence of meaningful learning. Typically, they find the idea of an association or network of associations to be incapable of reflecting all that a human does when confronted with a problem situation. Constructs are needed to enable us to think (Mandler, 1962, 1965).


... Unfortunately, there has been no sufficiently precise language available for formulating their ideas.

The choice, to date, has been between a precise, but seemingly in- appropriate S-R language, and presumably more relevant cognitive formula- tions which leave much to be desired insofar as scientific cohesiveness and rigor are concerned.

No serious investigator in the area today really questions that meaningful behavior involves the ability to 'perform successfully on an entire class of specific tasks, rather than simply on one member of the class' (Gagné, 1962, 355). For example, most psychologists, cognitive and noncognitive alike, believe that knowing how to add means that the learner is able to give the correct response to any addition problem, not just one. The issue seems to be whether it is more feasible to view the underlying knowledge as consisting of networks of associations or whether it would be better to adopt a new basic behavior unit. Thus Gagné (1964, 1965) has shown how many higher forms of learning depend on simpler forms, such as the association. The unfortunate fact of the matter, however, is that no one has been able to devise a completely satisfactory way to represent rules in terms of associations (cf. Scandura, 1966d, 1967d). For example, Gagné's (1964, 1965) representations of the rule do not use the S-R language but involve such constructs as chains of concepts and, more recently (1966a) 'action' concepts as well. As Tracy Kendler (1964) suggested in reacting to Gagné's (1964) original paper on the subject, new properties may emerge at the rule level. It would seem that in the study of meaningful learning it may well be more desirable to adopt as basic a higher form of learning.

... Over the past several years, a precise formulation of the notion of a rule has evolved. Since this formulation involves sets and functions, and since these characterizing notions have been used by the author and some of his students in formulating research, the label Set-Function Language (SFL) has been used. The SFL retains many basic tenets of cog- nitive formulations, but like all scientific languages, is free of spe- cific theoretical assumptions. In addition, the SFL is based on extreme- ly basic, and highly general, notions (sets and functions), so that it deals only with essential aspects of the constructs and empirical phenomena involved.

---

Excerpted from J. M. Scandura, "Role of Rules in Behavior: Toward an Operational Definition of 'What (Rule) is Learned'." Psychological Review, 1970, 77, 516-533, with permission of the publisher.
During the summer of 1962, Greeno and Scandura (1966d) found in a verbal concept learning situation that transfer occurred on the first presentation of a new item or not at all. Specifically, they had their Ss learn common responses (nonsense syllables) to each stimulus exemplar (nouns) of varying concepts. After each S-R pair had been learned, a transfer list was presented containing one new instance of each concept from the first list together with a paired control. The Ss either gave the correct responses to new concept exemplars on the first learning trial, or they learned the items at the same rate as their controls. The data were consistent with the hypotheses of all-or-none transfer.

It later occurred to Scandura that Ss might also transfer on an all-or-none basis to new instances of rules in which the stimuli may be paired with different responses. In this case, one new instance of a rule could be used as a test to determine whether the rule is learned, thereby making it possible to predict the responses to other (new) stimuli associated with the rule.

To test this point, a number of pilot studies were conducted during 1963 (Scandura, 1966d, 1967d, 1969a); in one experiment (Scandura, 1969a), a total of 15 (highly educated) Ss overlearned the list shown in Figure 1. Prior to learning the list, both the Ss and the experimenter agreed on the relevant dimensions and values--size (large-small), color (black-white), and shape (circle-triangle). The Ss were told to learn the pairs as efficiently as they could, since this might make it possible for them to respond appropriately to the transfer stimuli. After learning, the Test 1 stimuli were presented and the Ss were instructed to respond on the basis of what they had just learned. Positive reinforcement was given no matter what the response. Then, the Test 2 stimuli were presented in the same manner. The results were clear-cut. All but three of these Ss gave the responses "black" and "large," respectively, to the Test 1
stimuli (see Figure 1) and also responded with "white" and "small" to the Test 2 stimuli.

On what basis could this happen? It was surely not a simple case of stimulus generalization; the responses did not depend solely on common stimulus properties. The first Test 1 stimulus, for example, is as much like the fourth learning stimulus as the first. Perhaps the simplest interpretation of the obtained results is that most of the Ss discovered the two underlying principles during List 1 learning and later applied them to the test stimuli. These principles might be stated, "If (the stimulus is a) triangle, then (the response is the name of the) color" and "if circle, then size." In effect, whenever an S responded to the first test stimulus in accordance with one of these principles, he almost invariably responded in the same way to the second. Since this study was conducted, a relatively large amount of relevant data has been collected with essentially the same results (Roughead & Scandura, 1968; Scandura, 1967b, 1969b; Scandura & Durnin, 1968; Scandura, Woodward, & Lee, 1967).

The second observation was that each of Gagne's (1965) eight types of learning could be represented by a set of ordered stimulus-response pairs (Scandura, 1966d, 1967d, 1968) in which each stimulus was paired with a unique response. That is, each type conformed precisely to the set-theoretic definition of the mathematical notion of a function. To see this, first recall Gagne's eight types of learning: (1) signal learning—the establishment of a conditioned response, which is general, diffuse, and emotional, and not under voluntary control, to some signal; (2) S-R learning—making very precise movements, under voluntary control to very specific stimuli; (3) chaining—connecting together in a sequence two (or more) previously learned S-R pairs; (4) verbal association—a subvariety of chaining in which verbal stimuli and responses are involved; (5) multiple discrimination—learning a set of distinct chains which are free of interference; (6) concept learning—learning to respond to stimuli in terms of abstracted properties like color, shape, and number; (7) principle (rule) learning—acquiring the idea involved in such propositions as "If A, then B" where A and B are concepts—that is, a chain or relationship between concepts, internal representations (of concepts) rather than observables being linked; (8) problem solving—combining old principles so as to form new ones.

The first four types clearly involve a single stimulus and a single response. (Chaining and verbal associations, of course, may involve intermediary steps.) Multiple discrimination simply refers to a set of discrete S-R pairings (possibly with intermediate steps), each of which may act independently of the others and, hence, must be represented as a separate entity. Knowing a concept, however, may involve any number of different stimuli (exemplars), and each of these stimuli is paired with a common (unique) response. In addition, rules involve multiple responses. The stimuli and responses, however, are not paired in an arbitrary way; each stimulus has a unique response attached to it (see Figure 1, for an example).

In effect a rule can be denoted by a function whose domain is a set of stimuli and whose range is a set of responses. The concept and the association become special cases. A concept can be represented by a function in which each stimulus is paired with a common response, while an association can be viewed as a function whose defining set consists of a single S-R pair.
A Set-Function Language

What Gagné (1965) called problem solving involves a higher level of analysis. In particular, "combining old principles so as to form new ones" requires (higher order) rules which act on other rules. More generally, higher order rules may involve any number of combinations (sets) of old rules and any number of new ones, paired so that there is a unique new rule attached to each set of old ones. (Details are deferred to the section on higher order rules.)

Was this only a more formal way of expressing what psychologists have said all along--that responses are "functionally" dependent on stimuli? Scandura could not help but feel that there was a deeper significance. Still, defining rules, concepts, and associations in terms of their denotive sets left me with the unsatisfactory feeling of not knowing what they really were; or, to put it differently, how to characterize the knowledge underlying the observables.

2. A CHARACTERIZATION OF THE RULE CONSTRUCT

A function can be defined as a set of ordered pairs or as an ordered triple. The denotation of a rule, (i.e., class of S-R behaviors which can be generated by a rule) seems best characterized by the former type of definition, but the rule construct itself conforms more closely to the latter type of definition involving a set of inputs, a set of outputs, and a connecting operation.

Consider, for example, the task of summing arithmetic series (e.g., $1 + 3 + 5 + 7 + 9$). In this case, any one of an equivalence class of overt stimuli (like the sign, "$1 + 3 + 5 + 7 + 9$") may represent the same number series (i.e., $1 + 3 + 5 + 7 + 9$). Each such equivalence class serves as an effective (functionally distinct) stimulus. Effective responses (sums) may similarly be thought of as equivalence classes of overt responses (e.g., "25"). The denotation of the rule, then, consists of the set of ordered pairs whose first elements are equivalence classes of representations of number series, and whose second elements are equivalence classes of representations of their respective sums.

Underlying rules are, however, probably more naturally thought of not as acting on effective stimuli (responses) themselves but on properties of the entities denoted by these effective stimuli. Thus, for example, the property of having "a common difference of two between adjacent terms" refers to the number series, $1 + 3 + 5$, and not to its name, "$1 + 3 + 5$." Note that a distinction is being made between the entity (e.g., number series) and the equivalence class of representations

---

*By an equivalence class of overt stimuli (responses) or an effective stimulus is meant a class of overt stimuli, each of which has the same set of defining properties. The term "effective" is used to emphasize that we are talking about the stimuli and responses "effectively" operating in the situation rather than the overt stimuli and responses themselves. Thus, for example, the stimuli "5" and "five" would, for most purposes, count as the same effective stimulus since they both represent the same number. The stimuli "5" and "6," on the other hand, would correspond to different effective stimuli. In previous papers, Scandura (1966d,1967d) used the term "functionally distinct."

The distinction between an entity and the sign used to represent it will also play a role in the present analysis.
of that entity. However, since there is a one-to-one relation between equivalence classes of overt stimuli (the signs) and the abstract entities denoted, we can ignore the distinction. These properties, in turn, determine (via the rule) other properties (of the responses). One rule for summing arithmetic series, for example, may be represented by the expression, \([(A + L)/2]N\), where \(A\) refers to the first term, \(L\) to the last term, and \(N\) to the number of terms of the arithmetic series in question. The critical inputs associated with this rule are triples of values of the dimensions, \(A\), \(L\), and \(N\) (e.g., \(A = 1, L = 7, N = 4\)). These triples may be viewed as (composite) properties of the entities denoted by the stimuli. We may refer to these critical properties as response determining (D) properties. The set of outputs consists of response properties (numbers) derived from the properties in D. These properties (numbers) determine equivalence classes of number names (e.g., the number property, 16, which is the sum of the series, \(1 + 3 + 5 + 7\), defines the equivalence class of all signs of the form "16"). (Notice, however, that these number properties may also be viewed as properties of the series themselves. In this role, the number properties are called sums, which just happen to be properties of arithmetic series which can be derived from other presumably more easily determined properties, like the first term and the number of terms.)

In effect, a rule may be defined as an ordered triple \((D, O, R)\) where \(D\) refers to the determining properties of the stimuli, and \(O\) to the combining operation or transformation by which the derived properties (of the responses, \(R\)) are derived from the properties in D.

Parenthetically, note that accounting for such behaviors as adding arithmetic series in terms of rules is not the same as introducing mediating responses and response-produced stimuli. In the latter case, the basic idea is to provide a detailed account of the interrelationships involved in terms of (possibly complex) networks of associations. Rules treat such relationships at a more molar level. That is, rules by their very nature act on classes of effective stimuli and not on particular stimuli.

The basic question, of course, is which of these two alternatives better captures the essential characteristics of behavior on structured tasks. The first observation cited above, taken together with the relatively large amount of available data (e.g., Scandura, 1969a), indicates the behavioral reality of rules. Scandura found repeatedly that performance on any one instance of most structured tasks is directly related to performance on any other instance of the respective tasks. Behavior strongly tends to be either uniformly good or bad. (There is more that can be said on this point, but going into this here would detract from the main point.) Accordingly, it would seem that when an investigator is interested in working with structured tasks, the rule would seem to provide the more natural conceptual basis. Mediation accounts of such behavior tend to be ad hoc as well as complex and cumbersome. (In working with nonsense materials, on the other hand, where it is unclear as to what, if any, relationships exist among the instances, some resort to associations and their related theory may be more fruitful.)

This inadequacy of mediation accounts becomes one of principle unless one takes a more general view of stimulus and response than has generally been the case. In particular, no mediation theorist to the author's knowledge has explicitly considered as stimuli what amount, in a related context, to S-R pairs (i.e., associations). (Note: Any given
entity may serve as either a stimulus or a response. What the entity is called in any particular situation depends solely on the role it is playing--Hocutt, 1967.) To see this, it is sufficient to consider the associative connections involved in generating sums and differences in arithmetic, together with those connections which relate addition and subtraction. In this case, we would have as a minimum such connections as

\[
4 + 5 \rightarrow 9 \\
\downarrow \\
9 - 5 \rightarrow 4
\]

where the vertical arrow acts neither on the stimuli, 4 + 5 and 9 - 5, nor on the responses, 9 and 4, but rather on the associations themselves.

As a second and somewhat more subtle example, consider the task of adding "4" and "3" in column addition. If embedded in a problem like

\[
\begin{array}{c}
41 \\
+ 32
\end{array}
\]

the tens digit in the sum is "7." However, if the problem involves carrying, like

\[
\begin{array}{c}
47 \\
+ 35
\end{array}
\]

then the tens digit in the sum is "8." In effect, the response given to the complex "4, 3" depends on the context, in particular on the previous response. (In the first problem, the units digits "1" and "2" sum to "3" which does not involve carrying, whereas, in the second problem, the sum "12" of "7" and "5" does.) This implies that the effective stimulus in column addition includes not just the digits in a particular column but the previous response as well, specifically "carry" or "no carry." In effect, the stimulus in this case is a pair consisting of either "carry" or "no carry" paired with the tens digits "4" and "3." Thus, "carry, 4, 3" elicits the response "3," whereas "no carry, 4, 3" elicits "7." To see how these S-R pairs may be viewed as associations on associations, we need only observe that mediation theorists have no difficulty in talking about stimulus properties of responses (or, equivalently, in saying that the source of a given stimulus is the previous response). Hence, in this case, the stimulus properties of the response "carry," for example, may be thought of as eliciting the compound entity "4" and "3" as the response; it is the association "carry" → "4, 3," then, that serves as the stimulus (in the second problem) for the response "8."

As unfamiliar as this view may seem, this is precisely the sort of assumption that Suppes (1969a) had to make in proving that given any finite connected automaton (which for present purposes amounts essentially to a rule), there is a stimulus-response model that asymptotically becomes isomorphic to it. In order to account for rule-governed behavior, then, mediation theorists of necessity will have to generalize what to date has been the traditional view. ... An important generalization of this idea ... is the view that "associations on associations" are nothing more than a special case of "rules on rules."

\[^{4}\text{The main points are summarized in the next section.}\]
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 is to prove in detail that stimulus-response theory, or at least a mathematically 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 show that this linguistic claim is false. The specific constructions given below 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 works, but in general it does not, for reasons that become clear below. 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 (learning) that goes beyond the classical cases and yet represents perhaps the simplest nontrivial automaton. By nontrivial 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 stimuli 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.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LW</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RB</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RW</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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 nontrivial two-state automata.) 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.²

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 x 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 x S, but R x P(S), where 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 x S rather than S as the basic set of stimuli to which the axioms of automata theory 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, σ_j,n) rather than

\[ T = \{(r_i, σ_j)\}, \]

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 an axiom 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 x 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 ²Ed. Note: Due to space limitations, the technical details of Suppes' paper have been deleted.
Suppes

approach. Each response \( r \) produces a stimulus \( \sigma_r \), or more generally, a set of stimuli. Assuming again, for formal simplicity just one stimulus element \( \sigma_r \), 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 \( \sigma_r \) reduces directly and significantly the memory load on the animal. The importance of \( S_R \) for models of children’s language learning is less clear.

**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.

Given this theorem 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.

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. 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.
Suppes (1969a) has suggested that just as higher mathematics may be deduced from set theory, so may automata theory be deduced from stimulus-response theory. In other words, he seeks to destroy criticisms of stimulus-response theory based on the need for an internal state to mediate between stimulus and response by showing that finite memory can be conditioned by a suitable schedule of reinforcement. The present paper shows, first, that this result is simpler than may appear from Suppes' presentation, and second, that it is not true that this theorem vitiates the criticisms of stimulus-response theory.

1. THE PROOF

By a stimulus-response table is meant simply a table

<table>
<thead>
<tr>
<th>S_0</th>
<th>r_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>r_1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S_n</td>
<td>r_n</td>
</tr>
</tbody>
</table>

listing for each of \( n \) distinct stimuli \( s_j \), \( 1 \leq j \leq n \), a response \( r_j \) with which it is to be paired. The fundamental axiom of stimulus-response theory is that for each stimulus-response table there is a stimulus-response model that asymptotically becomes isomorphic to it in that there is a reinforcement schedule which will ensure with repeated trials, that for any \( j \), \( 1 \leq j \leq n \), the probability that the organism will respond to stimulus \( s_j \) with response \( r_j \) will tend to 1 as the number of trials tends to \( \infty \).

Consider a set \( R_1, \ldots, R_k \) of responses with the special property that part of the stimulus received by the organism at Trial \( t + 1 \) is an indicator
of the response at Trial \( t \)---for instance, the experimenter might arrange that at Trial \( t + 1 \) a light will shine above the lever pressed by a rat at Trial \( t \), or the organism may simply observe the position of one of its limbs. Thus imagine that the full stimulus \( (s) \) at a trial consists of a pair \( (\sigma_j, R_L) \), where \( \sigma_j \) is an external stimulus applied by the experimenter, while \( R_L \) is a stimulus which indicates that the organism's response was \( R_L \) at the previous trial.

In this case one might expect, by the fundamental axiom, to find a stimulus-response model that asymptotically becomes isomorphic to a stimulus-response table of the form

\[
\begin{array}{c|c}
(\sigma_1, R_L) & R_{11} \\
(\sigma_2, R_L) & R_{12} \\
\vdots & \vdots \\
(\sigma_k, R_L) & R_{1k} \\
(\sigma_1, R_L) & R_{21} \\
\vdots & \vdots \\
(\sigma_m, R_L) & R_{mk} \\
\end{array}
\]

in which the response of the organism at Trial \( t + 1 \) is determined not only by the stimulus externally applied at Trial \( t + 1 \), but also by the response of the organism at Trial \( t \). However, this will only be so (and it is the justification of the following statement that provides the only excuse for the detailed mathematics dispensed with in this presentation) if it can be ensured that every stimulus \( (\sigma_j, R_L) \) can be presented to the organism over and over again. Since the experimenter can only control \( R_L \) indirectly, this amounts to demanding that, at criterion, the organism may be induced to make any desired response if only the experimenter will provide the appropriate sequence of externally applied stimuli. If this is the case, it is said that the stimulus-response table is connected.

In this way, memory is introduced into the stimulus-response scheme, for a response now is determined in part by the previous response, which in turn depends partly on still earlier stimuli and responses. Thus following Suppes (1969a) one makes the illuminating discovery that stimulus-response theory does not preclude memory.


\(^2\)Excerpted in this volume, pp. 11-17.
In fact, finite-memory systems are well studied by some mathematicians and computer scientists under the label "finite automata." A finite automaton is specified by giving three finite sets—a set X of inputs, a set Y of outputs, and a set Q of internal states—and two functions—one which expresses the next state in terms of the present state and present input to the machine, and one which tells how the present output of the machine is determined by the present state. (An introduction to automata theory in the context of brain models may be found in Arbib, 1964, Ch. 1.) For simplicity—but note that despite a claim of Suppes to the contrary, it is with some loss of generality—the case where the present output equals the present state shall be studied only, so that only the set X of inputs or stimuli, the set Q of states or responses, and the function δ which determined the next state need be specified. Clearly, the specification of δ for a finite automaton (X, Q, δ) may be given by a table just like the second stimulus-response table above. Noting that an automaton yields a connected table just in case the automaton is connected in the sense that for each pair of states q and q' one may find an input sequence x₁ ... xₚ causing the machine to change from state q to state q', the following may be deduced.

1.1 Representation Theorem for Finite Automata

Let (X, Q, δ) be a connected finite automaton such that

(a) each element of X may be identified with a distinct externally applied stimulus from a set \{σ₁, ..., σₘ\};
(b) each element of Q may be identified with a distinct response from a set \{R₁, ..., Rₖ\};
(c) to each response Rᵢ there corresponds an indicator \(\hat{R}_i\) available to the organism at the next trial, and such that each pair \(σ_j, \hat{R}_i\) may serve as a distinct total stimulus.

Then there is a stimulus-response model that asymptotically becomes isomorphic to the finite automaton, in the sense that with increasing \(t\) the probability tends to 1 that if at Trial \(t\) the response corresponds to state \(q\), and at Trial \(t + 1\) the externally applied stimulus corresponds to input \(x\), then at Trial \(t + 1\) the response will correspond to the next state \(δ(q, x)\).

Thus the simple exposition of the result that Suppes (1969a) states as: "Given any connected finite automaton, there is a stimulus-response model that asymptotically becomes isomorphic to it" is completed.

This is an important result—but its importance can be dangerously overrated if Conditions a to c, which we have taken care to make explicit in the statement of the theorem, are neglected. The import of these conditions will be explored in the next section.

2. RESERVATIONS

Suppes notes that Miller, Galanter, and Pribram (1960) objected to the conditioned reflex as the building block for a scientific psychology, claiming that stimulus-response theories could not account for intention,
plan, and purpose. Instead, they formulated a theory of plans in terms of TOTE (test-operate-test-exit) units, a plan then being defined as a TOTE hierarchy. Since Chomsky (1963) has shown that a TOTE hierarchy may be viewed as a finite automaton, Suppes deduces that "Any TOTE-hierarchy in the sense of Miller and Chomsky is isomorphic to some stimulus-response model at asymptote." He would thus have us believe that intention, plan, and purpose are hence immediately subsumed into stimulus-response psychology. I claim that this conclusion is unwarranted.

2.1 The Argument by Time

The first reason for refusing to take seriously what shall be called the TOTE corollary is that it ignores the time required to approach asymptote for a stimulus-response scheme. It would not be unreasonable to explain a cognitive process in terms of a TOTE hierarchy which involves eight components with four states each, and this might have as many as $4^8 = 65,536$ total states. Assuming only 10 different externally applied stimuli, this yields a stimulus-response table with 655,360 entries--approximately 3 years' work even if a new entry is learned every minute of a 10-hour day. When one considers that such a TOTE hierarchy would be but one of thousands, one sees that no organism would live long enough to do all its learning this way. (Incidentally, I do not wish to imply that I find the TOTE hierarchy a perfect model for cognitive processing--only that I find it a better approximation than stimulus-response theory.)

The whole virtue of the hierarchical approach is that to master an eight-component hierarchy, the eight constituent TOTE units and a few rules about their interrelations need only be learned. One does not learn explicitly what to do in 655,360 different situations--instead one learns a few simple rules which cover all these contingencies.

It may well be that Suppes has demonstrated how learning starts, with his stimulus-response mechanism providing many of our simplest plans. But from then on, it seems, the ability to build on what has already been learned, treating complex cognitive structures as units at a higher level, is crucial to intelligent purposive behavior. Mental development is cumulative.

2.2 The Argument by Conditions a to c

There can be no quarrel with Condition a of the theorem. It is Conditions b and c that cause trouble. The above eight-component hierarchy would require the organism to discriminate 655,360 different total stimuli. In short, for many complex plans not only Condition b is doubted—that every distinct state of the automaton of a TOTE hierarchy finds distinct expression in a different response of the organism—but also Condition c must be rejected, holding that the input channels would be completely overloaded if they must carry not only representations of external stimuli but also complete information about the current state of execution of the TOTE hierarchy. Of course there are many TOTE units for which the relevant automata do satisfy the conditions, for example, if all relevant memory is represented in the immediate environment. It is important to know that given enough time, their behavior can be asymptotically attained by a stimulus-response model. In more
complex situations the author would hold that the state of execution is in fact encoded by the current output of all the neurons of the brain (setting aside questions of structural and biochemical changes)—but since only a minute fraction of these impinge on effectors or have efferent control of receptors, at any time, this is a far cry from accepting the general applicability of Conditions b and c. Again, I am prepared to admit that much of early learning in the infant does meet these conditions, but I cannot accept that they apply to most verbal behavior. (If one accepts the receptor surface of every neuron, no matter how central, as constituting part of the receptor surface of the whole organism, then we are all stimulus-response theorists, and the debate is meaningless.)

2.3 The Argument by Metaphor of Computers

The third argument simply notes that computers are programmed in a way which satisfies neither the conclusions nor the conditions of the representation theorem. To those, such as the present author, who feel that the computer metaphor has virtues far outweighing its dangers (if carefully handled—see Arbib, in press), this argument is most compelling. But to the staunch stimulus-response theorist, this argument is just begging the question and so it shall not be pursued here, save to note that the stimulus-response table for a digital computer, which is a sort of finite automaton, might well have $2^{100,000}$ entries—and that takes a lot of conditioning.

3. CONCLUSION

Suppes’ result seems most important in showing how many building blocks of behavior may be formed, but should not be construed as proof that intention, plan, and purpose may be subsumed into stimulus-response psychology. For these, some neural tissue analogue of a stored, internal, hierarchically organized program is required.
Arbib's (1969a)\(^2\) comments on my paper (Suppes, 1969a)\(^3\) on the stimulus-response theory of finite automata raise in a somewhat different form issues that continue to divide cognitive and stimulus-response psychologists. I am not persuaded by any of his arguments that a viable alternative to stimulus-response theory has yet been defined, although I am willing to admit that a fully adequate stimulus-response theory of complex learning and behavior has yet to be developed. The point of my paper was to show how the theory of finite automata and the theory of TOTE hierarchies in the sense of Miller and Chomsky could be subsumed in a rigorous way within an explicitly and precisely stated stimulus-response theory. Arbib's comments and arguments can be analyzed under five headings. The issues he raises are pertinent to making more definite the issues that divide stimulus-response and cognitive, especially linguistically oriented, theorists.

1. MATTERS OF PROOF

Arbib alleges that he has given a much simpler derivation of my main theorem: given any connected finite automaton, there is a stimulus-response model asymptotically isomorphic to it. He simplifies the proof by assuming that "for each stimulus-response table there is a stimulus-response model that asymptotically becomes isomorphic to it," but as Bertrand Russell said long ago, such procedures of postulation instead of proof have all the virtues of theft over honest toil. In other words, Arbib has not shown how the fundamental result follows from simple general assumptions about stimulus-response connections. To put the matter technically, he is faced with the problem of showing that his very powerful postulate, which corresponds to a theorem in my framework, is consistent with the simple assumptions about conditioning and sampling of stimuli that are an integral part of any standard stimulus-response theory.


2 Reprinted in this volume, pp. 19-23.

3 Excerpted in this volume, pp. 11-17.
Since it is not my main point in these remarks to deal with technical matters of proof, no more shall be said about these questions.

2. STIMULUS TRACES AND MEDIATING RESPONSES

Arbib seems to have a rather simple view of how an organism can remember its previous responses. He mentions the natural device used in experimentation with animals of having the previous response externalized by an appropriate stimulus. Exactly this approach has been used in the conditioning of pigeons to behave as simple automata, but it is not a general approach. Certainly it is not adequate to the problems of language learning to which my own research is primarily addressed.

Within classical stimulus-response theory that derives from Hull, for example, it is natural to talk about stimulus traces, and I mentioned this possibility. In terms of later stimulus-response theories, especially those associated with Osgood, Maltzman, Berlyne, the Kendlers, and others, it is natural to talk about mediating responses internal to the organism. Either of these concepts provides a natural framework to replace or to extend the very narrow externalization approach used by Arbib. Arbib objects to having some representation of the previous response in the organism for fear "that the input channels would be completely overloaded if they must carry not only representations of external stimuli but also complete information about the current state of execution of the TOTE hierarchy [p. 22]." This comment, like most of Arbib's other comments, however, is not backed up by a detailed analysis. It is not at all clear that this is a problem. Put in this general way, one may as well express concern over the internal processing also and simply say that it is found to be an overwhelming mystery of how the organism can work at all.

3. NUMBER OF STATES

I think the most serious and important issue raised by Arbib is the query of whether it is possible for an organism ever to have adequate time to learn the apparently large number of stimulus-response connections needed for it to become a suitably large automaton. This is a central issue that has been raised continually, particularly by psycholinguists critical of stimulus-response theories of language learning. The negative claims that Arbib makes concerning a TOTE hierarchy involving eight components with four states each are familiar, and his arguments also assume a familiar theological tone of saying it simply cannot be done.

I have come to make a distinction between negative dogma and negative proof. A wide range of psycholinguists and cognitive psychologists have asserted the negative dogma that stimulus-response ideas can never account for complex learning, because suitable conditioning connections could never be learned in sufficient time by organisms exhibiting complex behavior. This negative dogma seems to be an article of faith, and not an article of proof on the part of almost all who assert it, including Arbib. Certainly he does not give a proof that a device with 4^8 states is needed for any cognitive processing he cares to define. In view of his familiarity with the formal literature in the theory of automata on these problems, I suspect he will be somewhat more wary than many of the psycholinguists in venturing to give what would appear to be a detailed argument. For an example of a presumed negative proof as opposed to a negative dogma, but what is in fact a melange of confusions from end to end, see the proposed proof by Bever, Fodor, and Garrett (1968) that a formal
limitation of stimulus-response theory or associationism can be established.

It is generally recognized in mathematics and associated disciplines that negative arguments must be formulated in a more formal and explicit fashion than positive arguments. There is also a long history in mathematics and philosophy of establishing explicit and carefully defined systematic standards for evaluating the validity of a negative argument. Arbib does not work at all within that classical tradition. He makes a few casual statements about number of states, but offers no serious negative argument about the size of automata required for any cognitive processes understood well enough to be characterized in relatively exact terms. I will return to this point subsequently.

I think that the problem of the number of states required for various tasks is indeed a fundamental one, and I am not prepared to establish small upper bounds on the number of states needed for a very large number of tasks. On the other hand, I am skeptical of the large claims often made by psycholinguists or automata theorists. I would conjecture that the number of states required for any cognitive task, including language comprehension or production, will turn out to be a lot smaller than any of us originally thought. I readily admit that this is simply a conjecture, but I have enough confidence in it to insist that those who wish to claim that stimulus-response theory is inadequate because of the problem of the number of states that must be conditioned should offer negative proofs and not simply negative dogmas. The central focus of my own research efforts is to convert my own positive dogmas into positive proofs. (On the question of number of states, it is perhaps worth noting that a universal Turing machine can be constructed with only seven states.)

A simple example will illustrate an important point in talking about the number of states. If we flip a coin 100 times, the number of possible outcomes is $2^{100}$. Each of us can perform this experiment, but we could set the entire population of the world to flipping coins and not approach the number of possible outcomes. The number of actual states in a process, it is essential to note, is usually incredibly smaller, by orders and orders of magnitudes, than the number of possible states, and so it is with the relation between the number of actual and the number of possible conditioning connections.

4. REAL VERSUS METAPHORICAL LEARNING

Arbib argues in several places that learning simply could not take place according to stimulus-response conceptions because the number of states required is too large and the amount of time needed for conditioning is too restricted. I have already expressed my central argument against his claims about the number of states. I am willing to agree that if the number of states that must be conditioned individually is very large, then there will not be time for the task. What is interesting is the vagueness and weakness of the alternatives considered by Arbib. He expresses clear preference for TOTE hierarchies without considering these hierarchies as special cases of stimulus-response models, according to the first corollary of my main theorem. What he does not say or even sketch is how, within the framework of TOTE hierarchies, a theory can be given of the organism's acquiring cognitive skills. He does mention casually some references to neurons and biochemical changes in the brain, but no serious ideas are set forth in this physiological framework. The reader is left totally uninformed as to what serious alternative Arbib would propose,
with the single possible exception of his final remark about the computer as a metaphorical brain. Presumably, metaphorical brains engage in metaphorical learning. If this is what he is suggesting as a theory of learning, he cannot mean for his remarks in this direction to be taken seriously. An account of how organisms learn must be stated with sufficient definiteness and in sufficiently nonmetaphorical terms for it to be tested experimentally. The literature of artificial intelligence over the past decade shows clearly enough that computer theorists have as yet a poor idea of how to think about learning, or in particular, how to solve any special problems of learning in terms of computer hardware or software. Certainly Arbib himself makes no positive or definite suggestions of how to deal with any of the main difficulties facing theories of machine or human learning.

Indeed, it is surprising that in view of his opinions about the adequacy of psychological theories of learning he does not have more to say about the virtues of computer theories of learning. The reasons, however, are really clear enough. It does not take a very serious or systematic study of the literature of artificial intelligence to reach the conclusion that, although psychological theories of learning are certainly far from being sufficiently developed, so are theories of how computers should be programmed to learn.

5. PROBLEM OF HIERARCHIES

I have stated already that the issue about the number of states raised by Arbib is indeed a serious one. The other serious issue raised by him is the problem of hierarchies. It seems to be a matter of belief on the part of almost all cognitively oriented psychologists that conditioning theories must treat each simple conditioning connection as separate and equal. The concept of a hierarchy, it seems to be suggested, is contrary to the spirit of stimulus-response ideas. That a clear theoretical counterexample exists is shown at once by the corollary to a theorem, namely, the theorem that every TOTE hierarchy in the sense of Miller and Chomsky is isomorphic to some stimulus-response model at asymptote. The abstraction of this result may be unsatisfactory to some readers, but its meaning is clear. Any intuitive hierarchy that may be represented formally by a TOTE hierarchy, a finite oriented graph, a finite tree, or a finite automaton may be represented just as well by a stimulus-response model; in fact, better, for the stimulus-response model also includes an account of how the hierarchy is learned if the stimulus-response connections are not already coded in the genes.

Arbib concludes by stating that my results should not be taken to show that the concepts of intention, plan, and purpose can be subsumed under the central concepts of stimulus-response psychology. He does not challenge the formal result about TOTE hierarchies just discussed. He does challenge the stimulus-response account of how conditioning connections are built up in actual fact.

My own conclusion is this: What Arbib has given us is negative dogma not negative proof, and he has not stated a viable alternative to stimulus-response theory. The future task for stimulus-response theorists is evident. We must show in detail, for complex substantive examples, just how learning can take place according to stimulus-response ideas.
My comments have received a "reply" from Dr. Suppes (Suppes, 1969b). Rather than start an infinite chain of counterreplies, I invite the reader to take the note itself as reply, while drawing the following to his attention: One gains more insight by viewing Suppes' theorem as a simple corollary of the statement "for each stimulus-response table there is a stimulus-response model that asymptotically becomes isomorphic to it" than by giving a complex proof directly from axioms from which the above statement may be derived as a theorem. Suppes' theorem has only been proved subject to the conditions which have been made explicit in this paper, so that Suppes' use of his theorem to back up statements about stimulus-response theory is not always valid—in particular, we must disallow Suppes' TOTE corollary for all but those simple TOTE hierarchies which satisfy Conditions b and c. As for the rest, it is certainly not my belief that memory must be fed back through the input—I just explain the mechanism which Suppes and I used to deduce his theorem. As was stated in my paper, I do not even regard the TOTE hierarchy as an adequate model for learning. But the reader must not expect a theory of learning to be presented in a short theoretical note—I refer him instead to the author's forthcoming book The Metaphorical Brain (Arbib, 1972) and warn him that even that formulation will be provisional. In conclusion, I might say that I am charmed by the willingness of stimulus-response theorists—in their drive to avoid positing directly some m internal states to a finite automaton with p inputs—to accept $p^2(m^2+1)p$ conditioning states in the Markov chain they use in explaining learning.

---

In concluding Suppes' (1969b) reactions to Arbib's (1969a) reaction... to Suppes' (1969a) reaction to cognitive theorists, Suppes states, "The future task for stimulus-response (i.e., stimulus-sampling) theorists is...(to) show in detail, for complex substantive examples, just how learning can [emphasis added] take place according to stimulus-response ideas." My reaction to this comment is that it is precisely the wrong question to ask. It has much the same character as asking pre-Copernican astronomers to show in detail how planetary movements can be conceptualized with the earth (S-R associations) as the reference point. In my opinion, it is counterproductive in science to enter a new realm (e.g., complex human learning) with a strong commitment to principles involved in theorizing in what intuitively appears to be a qualitatively different area (e.g., simple conditioning).

Pre-existing foundations will prove useful only to the extent that (a) the basic concepts of the new theory can be reduced to those of the old and (b) the particular reduction in question is more "natural" than other possibilities which may exist. These points are well illustrated by extending Suppes' discussion of set theory. It is certainly true that mathematics can be reduced to set theory. It is also true, however, that this particular reduction has for some time been losing favor among research mathematicians. Many of them, perhaps a majority of the most productive, are now convinced that functions, or relations, between sets are far more important than the sets themselves. Granting this, Suppes' proof will prove useful only to the extent that: (a) automata provide a useful concept for theorizing about complex human behavior (i.e., the reduction has been proved; what is needed is a new theory) and (b) stimulus-sampling theory, or some other "conditioning" theory, provides the most natural reduction to S-R associations. In particular, the reduction from automata to S-R associations may follow more naturally from Arbib's axiom (which corresponds to Suppes' theorem) than from the assumptions made by Suppes regarding stimulus-sampling theory. (In fairness to Suppes, it should be emphasized that he mentions only the possibility of reducing complex forms of behavior to stimulus-sampling theory—cf. page 11 of this volume.)
Extending an implicit assumption by Suppes, I shall argue that automata (or some variant thereof) may provide an adequate basis for a theory of complex learning. Indeed, I shall suggest the form such a theory might take (for details see Volume I and Chapter 7). In the process, I shall also point out that it is more natural to reduce a theory of this type to S-R associations in a way which is more consistent with Arbib's axiom than with those assumed by Suppes as a basis for stimulus-sampling theory.

Let us first consider the form a theory might take if automata are taken as the basic building blocks. How might such automata be learned? If conditioning is the basic process, then clearly automata generally would be acquired gradually over long periods of time (and I might add be subject to the arguments proposed by Arbib). For present purposes, it is taken as axiomatic that the subject is not a blank slate, but comes into any learning situation with certain competencies (automata). These competencies form the basis for all subsequent learning. In effect, we assume that new knowledge is acquired by the interaction of existing automata in consort with a given environmental situation. A result of such interaction is new automata.

Suppose that a subject has already learned to act as automata A and B, where A might involve, "the ability to convert yards into feet," and B, "the ability to convert feet into inches." One question we might ask is: how does the subject learn to act like automaton C, "the ability to convert yards into inches"?

In the present view, automata A and B are somehow combined to form automaton C. This might be accomplished, for example, by a simple higher order rule (automaton) which says effectively to note the inputs and outputs of the automata (e.g., A and B) and to order them (the automata) so that the output of the first becomes the input of the second. Application of a higher order rule of this sort would result in an automaton C, which can be represented:

\[ \text{Automaton } C = \text{Automaton } A \rightarrow \text{Automaton } B \]

(The type of combination shown is well known to automata theorists as a series connection.)

Of course, automaton C could also be learned through a conditioning process. To see this, notice that according to Suppes' result, each of the automata can be represented as a stimulus-response table of the sort displayed by Arbib (page 20, this volume). Assuming automaton C is represented in a natural way, the characterizing table of C will obviously

---

3 In automata theory, there has been a general tendency to ignore the distinction which may be made between programs (algorithms, rules, effective procedures) and machines which use them (Scott, 1967). In traditional formulations of automata theory the two are almost totally confounded.

I feel that there are very good reasons in talking about complex behavior for making a sharp distinction between what a subject knows and what behaviors he is actually capable of as a consequence of built-in limitations of the human machine (cf. Scandura, 1971a).

Although I would have preferred to talk about rules and the functions computed by them, I have bowed to convention in this paper and have framed my arguments in terms of automata.
bear a very close relationship to the tables used to represent automata A and B. In particular, table C will include tables A and B, together with whatever stimulus-response pairs are needed to effect the transition from the outputs of automaton A to the inputs of automaton B. Modifying and extending Arbib's notation slightly, this may be represented:

![Table 1](image)

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (\sigma_1, \hat{r}_1) )</td>
<td>( R_{11} )</td>
</tr>
<tr>
<td>( (\sigma_1, \hat{r}_2) )</td>
<td>( R_{12} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( (\sigma_1, \hat{r}_k) )</td>
<td>( R_{1k} )</td>
</tr>
<tr>
<td>( (\sigma_2, \hat{r}_1) )</td>
<td>( R_{21} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( (\sigma_n, \hat{r}_k) )</td>
<td>( R_{mk} )</td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>( (s_1, \hat{r}_1) )</td>
<td>( r_{11} )</td>
</tr>
<tr>
<td>( (s_1, \hat{r}_2) )</td>
<td>( r_{12} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( (s_1, \hat{r}_e) )</td>
<td>( r_{1e} )</td>
</tr>
<tr>
<td>( (s_2, \hat{r}_1) )</td>
<td>( r_{21} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( (s_n, \hat{r}_e) )</td>
<td>( r_{ne} )</td>
</tr>
</tbody>
</table>
Notice that the transitional stimulus, $t_i$, is paired with each terminal state of $A$, and that each such pair is associated with a response, $r_{t_j}$, corresponding to the initial state, $r_{t_1}$, of $B$.

At first glance, the S-R approach might appear to have certain advantages in that it pinpoints precisely which S-R instances need to be learned. However, this apparent rigor and precision is actually misleading. To see why, consider a second learning situation of the same type in which the subject enters the situation with the capability of behaving like automata $A'$ and $B'$, where $A'$ involves converting degrees of latitude into minutes, and $B'$ converting minutes of latitude into seconds. The desired automaton $C'$, then, would correspondingly involve converting degrees of latitude into seconds.

Again, we can think of going from automata $A'$ and $B'$ to automaton $C'$, either within the S-R conditioning framework or by combining automata directly. In the former case, one constructs an S-R table for automaton $C'$ by introducing new S-R pairs to insure the transition from the table of automaton $A'$ to the table of automaton $B'$. In the latter case, automata $A'$ and $B'$ may be combined directly to form $C'$ in precisely the same way that $A$ and $B$ are combined to form $C$. The S-R view, however, tends to camouflage the (apparently) close relationship between the two learning tasks. In particular, the S-R instances needed to effect the transition from $A$ to $B$ and those needed to effect the transition from $A'$ to $B'$ have nothing directly to do with one another. It is not surprising, then, that S-R theorists seem (to my knowledge) never to have seriously considered higher order relationships of this sort. The similarity between the two learning situations only becomes clear when one looks at the automata as wholes. In this case, it is easy to see that the same higher order rule used to combine $A$ and $B$ to form $C$ would also serve to combine $A'$ and $B'$ to form $C'$. Furthermore, this same higher order rule (or automaton) would apply to any other pair of automata which can be combined in the same way.

It is possible, of course, that the apparent relationship might turn out not to be reflected in actual behavior. If this were the case, there would be no special reason to prefer the more cognitive interpretation. Most of the earlier problem solving research (e.g., Bartlett, 1932a; Duncker, 1945; Wertheimer, 1945), unfortunately, does not seem to deal very directly with this issue. The only really relevant experiment I know of is one of my own (Scandura, 1967b—cf. p. 52, Chapter 2) and, even here, the experiment was designed with something else in mind (cf. Scandura, 1971a).4

In this study, all of the subjects were taught to use a higher order rule (i.e., use of parentheses). Half were also taught to perform like one or more (lower order) automata (e.g., to determine the greatest integer in a number, denoted $[x] \rightarrow n$). The subjects who learned to perform like (lower order) automata were able with 80% certainty to combine these automata to form new automata (e.g., $[((x) + [y])] \rightarrow m$). The other subjects were uniformly unable to do so.

4 More recent and directly related experiments have been reported in Chapter 7 and Volume I.
For the sake of argument, let us further assume that new learning results from the action of higher order automata whose inputs (e.g., \((A, B)\) and \((A', B')\)) and outputs (e.g., \(C\) and \(C'\)) are themselves automata. The basic idea suggested by the above discussion is that any given (previously learned) higher order automaton acts on classes of \((n\text{-}uples)\) lower order automata (e.g., \((A, B)\) and \((A', B')\)) to produce new automata (e.g., respectively, \(C\) and \(C'\)). The specific nature of the psychological mechanism needed to effect the proper combination of such automata (to produce new learning) is not essential here. In view of the preliminary evidence cited above, it is sufficient to simply assume that one does exist.

Although this idea seems clearly to be more compatible with a conception of learning based on prior learning, I do not intend to imply that S-R conditioning theorists could not possibly handle the problem within an S-R framework. Indeed, they could condition individuals to become higher order automata by the same argument that Suppes (1969a) presented in his original paper.

On the other hand, if we take the position that knowledge begets new knowledge, then we might ask what are the simplest possible units of knowledge? This would certainly constitute one natural reduction. Noting the relationship Arbib has proposed between automata and S-R association tables (corresponding to Suppes' theorem), we see that the simplest automaton consists of a single association in which the stimulus itself is a pair. (Equivalently, by arguments presented by Suppes and myself—see pages 16 and 34 in this chapter, the stimulus may itself be an association.) In effect, the simplest unit in a theory of the form proposed would appear to be an association on an association. Further, in the theory there is no question how the S-R pairs get there; it is not a matter of conditioning. Even the newborn come wired-in for certain kinds of behavior, and I would propose that it is precisely these predispositions which form the basis for all future learning.

According to this view of learning, then, talking about how one might condition an individual to become an arbitrary automaton is a purely mathematical exercise and has no relevance to the phenomenon involved. That is, of course, unless the subject in question has never learned previously anything relevant to the automaton he is to become. But, then, I think this sort of learning really exists only in the mind of the S-R conditioning theoretician.

What Suppes has shown is that, ignoring time limitations and practical considerations involved in the process of conditioning, an organism may be conditioned to become an arbitrary finite connected automaton. This is worth knowing. But, to go on and imply that the major task of S-R theorists is to show for complex examples how learning can take place by S-R principles is unwarranted. A more profitable goal may be to search for new and better ways of conceptualizing complex learning.

---

^ See p. 11, this volume.
NEW DIRECTIONS FOR RESEARCH ON RULE LEARNING
Much of the research reported below was formulated and completed prior to the development of the SFL in its present form. Indeed, in the beginning, the research was based largely on a vague feeling that a new approach to research on meaningful learning was needed (Scandura, 1964a, b; Scandura, 1966a, b, c; Scandura and Behr, 1966). Both the language and the experiments evolved simultaneously, with the language helping to give impetus to the experiments, and the experimental results pointing to theoretical inadequacies in need of clarification.

The next article in this chapter includes some comments by Wittrock concerning the approach described below. This is followed by an excerpt from one of the editor's subsequent papers which effectively constitutes a reply.

1. PILOT RESEARCH ON RESPONSE CONSISTENCY

During the summer of 1962, Greeno and Scandura (1966) found, in a verbal concept learning situation, that essentially S either gives the correct response the first time he sees a transfer stimulus or the transfer item is learned at the same rate as its control. (This first experiment was reported in Chapter 1 and is repeated here for the convenience of the reader.)

If transfer obtains on the first trial (if at all), then responses to additional transfer items, at least under certain conditions, should be contingent on the response given to the first transfer stimulus. In effect, a first transfer stimulus could serve as a test to determine what had been learned during the original learning, thereby making it possible to predict what response S would give to a second transfer stimulus. To test this assumption, a total of about fifteen (highly educated) Ss overlearned the list shown in Figure 1.

---


2The learning (list 1) and transfer and control (list 2) stimuli were Underwood and Richardson (1956) nouns, the responses were nonsense syllables, and the lists were learned by a self-paced anticipation method. The transfer stimuli belonged to the same concept categories as did the learning stimuli.
Prior to learning the list, both the Ss and the experimenter agreed on the relevant dimensions and values—size (large-small), color (black-white), and shape (circle-triangle). The Ss were told to learn the pairs as efficiently as they could since this might make it possible for them to respond appropriately to the transfer stimuli. After learning, the Test One stimuli were presented and the Ss were instructed to respond on the basis of what they had just learned. Positive reinforcement was given no matter what the response. Then, the Test Two stimuli were presented in the same manner.

The results were clear-cut. All but three of these Ss gave the responses 'black' and 'large' respectively to the two Test One Stimuli (see Figure 1) and, also, responded with 'white' and 'small' to the Test Two Stimuli. On what basis could this happen? It was surely not a simple case of stimulus generalization; the responses did not depend solely on common stimulus properties. The first Test One Stimulus, for example, is as much like the fourth learning stimulus as the first (see Figure 1).

Perhaps the simplest interpretation of the obtained results, is that most of the Ss discovered the two underlying principles during list one learning and later applied them to the test stimuli. These principles might be stated, 'If (the stimulus is a) triangle, then (the response is the name of the) color' and 'If circle, then size.' The former principle may be characterized by letting I involve the property of being a triangle, D = the set of color properties, R = the set of color names, and O = the transform which maps color properties onto their names. The denotation of this principle would consist of the set of S-R pairs, \(\{(S_i, R_i) | i \in I\}\) where \(S_i\) is a colored triangle and \(R_i\) is the name of the corresponding color.

The results obtained in this miniature pilot experiment (which I have repeated a number of times) provide support for the contention that principle learning is an all-or-none affair (Scandura, 1965; Greeno and Scandura, 1966). The Ss either learned the principles or they didn't; in almost every case, Test Two performance was compatible with that obtained on Test One. This is not to say, of course, that all of the Ss learned the two principles indicated above. Apparently, some of the Ss
Empirical Research

ignored the similarities between the original pairs and learned them in rote fashion—i.e., as four distinct principles, each involving one stimulus (actually one equivalence class of stimuli) and the corresponding word response. Under such circumstances, random test performance would be anticipated.

The results of another pilot study, conducted at the University of Michigan during the summer of 1963 and reported by Scandura (1966e), were also revealing. In this case, high dominance nouns taken from the Underwood and Richardson (1956) list were used. Each of these nouns elicited a single adjective associate with a frequency greater than 50%. Eleven college Ss overlearned a list of eight such nouns associated with a total of four adjective categories; two nouns were associated with each adjective category. Both stimuli in a given category were assigned a common response. The four Test One and four Test Two stimuli were also high dominance nouns associated with the same four adjective categories. The task was put in the context of a game in which an explorer is lost and must determine which direction (the four responses) he should go given various hints (the transfer nouns). S had the option of responding to the transfer nouns with, 'I don't know,' when none of the four learned responses seemed appropriate. Without this control, appropriate responding to a transfer noun would have occurred by chance in about one out of four cases. Again positive reinforcement was given for all choices.

The results with these concept materials were equally revealing. In those cases where transfer potential was indicated, the responses to the second set of stimuli conformed to prediction in 47 of 52 cases. Furthermore, when asked, all but two of these Ss correctly identified the common adjective as the basis for their test responses.

The predictive value of the methodology described is dependent on knowing when S will continue to employ the same rule or responding set. Consistency of response to test stimuli may be influenced by feedback as well as by instruction variables operating between the first and second test responses. The effects of positive, negative, and neutral reinforcement of the first test response may be crucial. Telling S how he should respond or indicating that the 'rules have changed,' by hint or choice of test stimuli, may also affect response consistency. On the other hand, suggesting that the first response is appropriate apparently encourages use of the same responding set on the second test. As indicated above, S was told in our pilot experiments that he was correct regardless of how he responded to the test stimuli. He also was encouraged to respond on the basis of his prior learning. In effect, the experimental situation was designed to both control and capitalize on Einstellung for assessment and predictive purposes. Under these conditions, response consistency was near perfect.

Nonetheless, these results cannot be interpreted as unambiguously as in the first pilot study. With verbal materials it is almost impossible to identify all of the important stimulus dimensions. Nonetheless, in concept learning experiments, using the Underwood and Richardson (1956) materials, it does appear that the dominant adjective associate plays the predominant role.

With actual subject matters, assessment sometimes presents additional problems. In the first place, it is not always easy to specify uniquely the basis for an overt response. There is usually more than one path to the goal. Consider, for example, a situation in which S is asked to compute 35.449 + 35.551 as rapidly as possible. S can laboriously multiply
2. **PRINCIPLE LEARNING**

The question of relationships between S-R pairs seems so basic, and so obvious, that one wonders why it has not been studied extensively. Because it provides a simple context in which to contrast mediation and set-function formulations, this problem is discussed in some detail.6

Consider a paired associate (PA) context in which the relationships between four pairs are varied while the other factors are held constant. In Figure 2, such a manipulation is accomplished by selecting the two principles indicated by 'If black, then shape' and 'If white, then size.'

35 times 449 and 35 times 551 and then add the products or he can recognize this is an instance where the distributive rule would allow him to compute 35 \( (449 + 551) = 35 \times 1000 = 35,000 \). Clearly, it is not the sum alone which determines what is learned (i.e., the 'way' in which the problem is solved), but the time it takes to respond. If the correct answer is given in a short time, the distributive rule or some equivalent was probably used. Giving the answer in a relatively long time would likely indicate use of the usual computational algorithm. In both cases, the physical response alone would not be adequate to specify the rule. If S gives an incorrect answer or if the problem is so easy that there would be little time differential no matter how S does the computations, further complications would be introduced.

In short, the careful selection of test stimuli and responses is essential in order to assess knowledge. Ideally, these elements should be chosen so as to eliminate all *modi operandi* but the one in question. Although probably not attainable, this ideal can be approached in many cases.

Another problem involved in work with actual subject matters is that of complexity. More than one principle may, and usually does, enter into a single test response. To determine the knowledge underlying the response, it is often necessary to assess each principle individually, as in diagnostic work with school children.

In many test situations, there are few available responses from which to choose (as in True-False and Multiple Choice tests). Under these conditions, there are additional problems of assessment since there is a high probability of giving any particular response (by guessing) irrespective of what is learned. A similar problem obtains in assessing concept learning. There are at least three ways of minimizing this problem: (1) present more than one test stimulus, (2) include appropriate controls for comparison (e.g., Greeno and Scandura, 1966), and (3) provide an alternative to guessing as was done in the pilot study described above.

The assessment methodology employed in this research may be used in conjunction with two types of variable: (1) those which affect the probability of rule learning and (2) those which affect response consistency. Giving directions and presenting cues, hints, or other attention-getting devices provide examples of the former type of variable. The consistency with which S responds according to a learned rule may be influenced by feedback, as well as instruction variables operating between the first and second test responses.

When this research was conducted, I, in the four-tuple characterization of a principle, was viewed as a set of stimulus properties which determine when a particular rule \( (D, O, R) \) is to be applied. It now seems more reasonable to view the process by which rules are selected for use as higher order rules defined on the entire contextual situation, including the desired goal as well as the stimulus context.
Fig. 2. Sample paired-associate lists, together with S-R mediation and SFL representations of these lists. In the experimental list the pairs are interrelated; in the control list, they are not. Two principles are involved in the experimental list: (1) If black, then shape, (2) If white, then size.

Broken lines indicate associations to be learned; solid lines indicate previous associations. The symbols, rs, refer to both the mediating response and the response produced stimulus. In the S-R mediation representation of the experimental lists, rs, corresponds to 'black', rs1 to 'shape', rs2 to 'white', rs3 to 'size', rs4 to 'triangle', rs5 to 'circle', and rs6 to 'large', and rs7 to 'small'. In the control list representation, rs1 corresponds to 'circle', rs2 to 'small', rs3 to 'large', and rs4 to 'triangle'.

In the SFL representation both the principles and denotations are characterized. 0, in each case, is the mapping between D and R.
In the experimental list, two pairs correspond to each of the two principles involved. The stimulus properties (sizes, colors, and shapes) and the (internal) responses (names of shapes and sizes) in the control list are identical with those in the experimental list. Thus, any differences in the learnability of these lists would be hard to attribute to anything but the presence of relationships between pairs in the experimental list.

Assuming S and E agree on the relevant stimulus dimensions, S's task in learning the experimental list can be viewed as that of discovering the principle identifying (I) and response determining (D) attributes since O is essentially an identity map (i.e., naming) between D and R. On the other hand, S could learn the experimental list without noting any relationship between the pairs. Only one real alternative is available in learning the control list; the list was constructed so that no principle exists which involves more than one pair. To the extent that relationships between pairs are noted, the experimental list should be easier to learn.

The S-R mediation description of the list contingencies in Figure 2 leaves much to be desired. The representation of principle learning is relatively complex and would have been even more so had we not let 'rs' represent both mediating responses and their assumed stimulus properties. No single chain, for example, can adequately represent rule or principle learning in which more than one pair is involved. The one-to-one pairing between the $S_i$ and $R_i$ ($i = 1, \ldots, 4$) does not follow from a simple analysis of the S-R links in the longer three-stage chain. This chain does not make clear, for example, why $R_1$ is the response to $S_1$ rather than $R_2$. The more direct two-link chains involving the $rs_i (i = 1, \ldots, 4)$ serve this purpose. In effect, corresponding stimuli and responses are connected by two chains; those which do not correspond are connected by only one.

In view of this complexity, perhaps the most crucial limitation of S-R formulations may prove to be their inability to lead to practically important questions concerning meaningful learning. The S-R representations that would seem to be called for bear more than a passing resemblance to Copernican epicycles and related attempts to salvage concentric theory.

2.1. Learning principles in paired-associate lists.

Paired-associate learning (PA) has been studied as a function of many variables (e.g., meaningfulness, association value, pronounce-ability) but little attention has been given to relationships among different S-R pairs. The purpose of this exploratory effort was to determine relationships between (1) the number of S-R pairs related by a common principle, (2) learning rate, and (3) transfer (Scandura, 1967c).

The materials to be learned consisted of 12 pair lists. Each stimulus had a property relating to shape, border, shading, outline, and color. Four colors and eight values of each of the other four attributes were

---

7 It may appear that an appropriate control list could be constructed by pairing the experimental stimuli and responses in random fashion. Alas, this turns out not to be a critical control. The responses in the experimental list would all be names of properties of the experimental stimuli but this would not necessarily be the case in the control list. In effect, any differences between the groups could be attributed to pre-learned associations between the respective stimuli and responses in the experimental list rather than to relationships between the pairs.
used. The responses were descriptive labels attached to the non-color stimulus properties (e.g., circle). Of the 12 pairs in each list, six were instances of one principle (P6), three were instances of another (P3), two were instances of a third (P2), and one was an instance of a fourth (P1). The principles were constructed so that the same principle applied to all stimuli having a particular color. The response determining cue was either a shape, a border, a shading, or an outline. The four colors and the determining attribute dimensions (e.g., shape) were randomly paired to form four principles (e.g., If black, then shape) which appeared equally often under each condition. The PA list was learned by the anticipation method to a criterion of three consecutive errorless trials.

Prior to learning the original list, each of the 20 college Ss was pretrained so that he was familiar with the stimulus dimensions and could name each stimulus property. These responses were typed on a card and were always available to S. In addition, S was told that a pattern was involved which might facilitate his learning and guide his responses to the transfer stimuli.

The dependent variables were the average number of errors per instance (i.e., an S-R pair associated with a principle) for each S (on each of the four principles) and the number of appropriate responses to the transfer stimuli.

Except for a very small reversal between treatment P3 and P2, learning rate (i.e., the average number of errors per instance) decreased with the number of instances per principle: 5.0, 3.4, 3.5, and 2.7, respectively (F = 8.76, df = 3/76, p < .001). The difference between P1 and P2 was significant (F = 11.50, df = 1/76, p < .01), but none of the other adjacent means differed significantly. Under the experimental conditions, the rate of learning an S-R pair increased with the addition of a second S-R instance but increasing the number of instances still further apparently had little effect.

The number of appropriate responses to the transfer stimuli was also affected by the number of instances per principle. There were 27, 8, 15, and 9 appropriate responses (as indicated by the experimental principles) given to the P6, P3, P2, and P1 transfer stimuli, respectively. Although the trend was not entirely regular, a sign test indicated that the degree of principle learning was higher in treatment P6 than in the average of treatments P3, P2, and P1 (z = 2.6, p < .005). It might be argued that the difference in the number of appropriate responses was due to there being more responses per category in treatment

---

8 It might be argued that the difference in the number of appropriate responses was due to there being more responses per category in treatment...
Another analysis demonstrated that P6 transfer was related to learning rate. Of those 9 Ss who responded appropriately to both P6 transfer stimuli, 7 had below median (2.61) error scores, indicating more rapid learning; of those 11 Ss who responded appropriately to at most one test stimulus, 8 had above median error scores, indicating slower learning. An exact probability test (Finney, 1948) on the resulting 2 x 2 contingency table indicated a significant relationship between P6 transfer and learning rate (p < .035). The small number of Ss who gave two appropriate (transfer) responses with respect to the other principles precluded the possibility of obtaining significant relationships. Only 3, 5, and 2 Ss gave both desired responses to the P3, P2, and P1 test stimuli, respectively.

The list learning and transfer results were not entirely consistent. The inclusion of more than two instances did not affect learning rate, but it may have affected transfer. These results could reflect real differences or be simply artifacts of the situation. In either case, the overall pattern of results was sufficiently clear to make any interpretation in terms of stimulus or response generalization extremely difficult, if not impossible. Some resort to S-R generalization (Hull, 1943; Berlyne, 1965) or rule learning would appear necessary. For reason primarily of parsimony, a rule interpretation would seem preferable.

3. RULE GENERALITY AND CONSISTENCY IN RULE LEARNING

In many instructional situations, the question often arises as to how general the presentation of material ought to be. Some proponents emphasize that the more general a rule the more useful it will be: others, that the more specific the rule, the better the learning. There is a real need to better understand the psychological principles involved, but previous studies dealing with rule (or principle) learning (e.g., Craig, 1956; Gagné and Brown, 1961; Haselrud and Meyers, 1958; Kersh, 1958, 1962; Kittle, 1957; Scandura, 1964a, 1966a; Wittrock, 1963) have dealt only indirectly with this question.

Assuming that the answers hinge, at least in part, on learnability as well as general utility, and armed with the denotative characterization of a rule as a function, we (Scandura, Woodward and Lee, 1967) set out to explore this question. In particular, we were concerned with the effects of rule generality on learnability and transfer. We also explored the response consistency hypothesis with more complex materials. Two experiments were conducted, the independent variable in both cases being the scope (i.e., generality) of a rule statement. Scope was defined in terms of the corresponding denotation, one statement being more general than another if the denotation of the former included the latter.10

**P6.** When in doubt, the Ss may have tended to give a response from the most frequently experienced category. A comparison however, of the average number of P6 responses given to the P3, P2, and P1 transfer stimuli (16) was not significantly higher than the ten P3, P2, and P1 responses given to the P6 stimuli (p > .10).

9Although a legitimate distinction may be made between rules and principles, the distinction is fine and was not recognized until after the study was completed. The terms have been used synonymously throughout the paper except in the concluding section where the distinction is outlined.

10Notice that defining a rule as a function makes it possible to
Our original hypotheses were that: (1) the scope of a rule would be fully reflected in performance, there would be no success on extra-scope problems and little difference in performance on within-scope problems, (2) the learnability of a statement, as determined by within-scope performance, would vary inversely with scope, and (3) the rule taught would be used on all problems under conditions of nonreinforcement.

3.1 Experiment one

In the first experiment, each group of 17 college Ss (majors in elementary education) was presented with one of three ordered rules dealing with a variant of the number game called NIM. In the game, two players alternately select numbers from a specified set of consecutive integers, beginning with one, and keep a running sum. The winner is the one who picks the last number \((j)\) in a series with a predetermined sum. If, for example, this sum is 31 and the set consists of the integers 1-6, the players alternately select numbers from 1-6 until the cumulative sum is either 31 or above (in which case no one wins). There are rules which allow the player who goes first to always win.

Any such game can be characterized by an ordered pair of integers. The application of each winning rule was illustrated with a common \((6, 31)\) game. The least general rule \((S)\), adequate for winning only \((6, 31)\) games, was stated, '... make 3 your first selection. Then ... make selections so that the sums corresponding to your selections differ by 7.' Rule SG was adequate for solving \((6, j)\) games \(j = 1, 2, \ldots, n\) and was stated, 'the first selection is determined by dividing the desired sum by 7 and making the remainder your first selection... Then ... make selections so that the sums corresponding to your selections differ by 7.' The most general rule \((G)\) was adequate for solving \((i, j)\) games \(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\) and was stated, 'the first selection is determined by dividing the desired sum by one more than the largest integer in the set from which the selections must come and making the remainder your first selection... Then ... make selections so that the sums corresponding to your selections differ by one more than the largest integer in the set.'

All Ss, including two control groups, were tested on three problems. The first was within the scope of each rule, the second within the scope of all but rule \(S\), and the third only within the scope of rule \(G\).

The results were straightforward. Of the 13 Ss in group \(S\) who solved problem one, none solved problem two, and one solved problem three. The corresponding numbers for groups SG and G were, respectively, 5, 4, 0 and 5, 5, 4. Within the scope of each rule there were only chance differences in performance on the problems. On the other hand, only one S solved an extra-scope problem.

Consider a variety of other relationships between different rules. In particular, two functions (i.e., sets) may be disjoint (i.e., have no instances in common), overlap, or be identical in addition to being ordered (i.e., one being more general than the other). Such relationships might provide a basis for formalizing questions concerning the effects of prior learning on later learning.

11Exact probability tests on 2 \(\times\) 2 contingency tables were used to test the various hypotheses.
The relative interpretability of the three rule statements was determined by comparing group performance on problem one which was within the scope of each. Rule S proved to be easier to learn, under the self-paced conditions, than were rules SG and G (p < .005 in both cases). There was, however, no discernable difference in the interpretability of rules SG and G.

The third facet of this research was concerned with the consistency with which presented rules are applied. We wanted to determine whether the S and SG Ss would use the rule taught even when it was inappropriate (on the second and third problems). To make this possible, no information was given as to when the rules were and were not appropriate.

Of the 17 S Ss, 13, 9, and 8 used the rule taught on problems one, two, and three, respectively. The corresponding numbers in groups SG and G were 7, 7, and 5 and 6, 6, and 6. Although there was a slight tendency to not use the rules taught on problems two and/or three, where they were inappropriate, there were no significant differences in frequency of use.

These results certainly provided strong support for our original hypotheses: (1) performance on within-scope problems did not differ appreciably, even though the common illustration was more similar to problem one than the others, and successful problem solving was limited almost exclusively to within-scope problems, (2) rule S proved easier to interpret than rules SG and G, and (3) the rules taught tended to be used consistently on all problems whether they were appropriate or not.

About the only major unanticipated result in experiment one was that rule G proved as easy to interpret as rule SG. In view of the rather low proportion of successes in these groups, we were originally tempted to attribute the lack of such an effect to scale insensitivity near its lower extreme.

3.2 Experiment two.

To determine the generality of these findings, a second experiment, dealing with arithmetic series, was conducted simultaneously with junior high school Ss. In this experiment, both scope (S, SG, G%) and example (present, absent) were varied independently. Since most of the Ss were already familiar with arithmetic operations introduced and, to some extent, with number series generally (i.e., as in adding lists), it was

12 The first mentioned result has particular relevance for the psychologist since it crystallizes the fact that no generalization gradient is to be expected when the stimulus values are discrete rather than based on a continuous physical dimension. If there were such a gradient, performance on the first test problem which was more similar to the example, should have been superior to that on the other problems. Even S-R associationists are generally agreed that the lack of such an effect provides indirect support for a rule interpretation. To the extent that the variables involved in meaningful learning are discrete, a rule interpretation may prove more useful.

Furthermore, when the underlying stimulus dimension(s) are continuous, S-R theorists will need to consider the possibility that generalization gradients are simply artifacts of averaging individual differences in perceptual discrimination (and, hence, what rule is learned) over continuous dimensions (e.g., Lykken, Rose, Luther and Maley, 1966).
felt that examples might provide a basis for generalization, via discovery, to extra-scope problems. Another difference between this experiment and the first was that rule S, \(50 \times 50 = 2500\), was effectively an answer given treatment and applied to only one series. This series was used both as the common example and as problem one. In experiment one, rule S applied to a number of different game sequences.

Although the pattern of results shown in Table I paralleled those of experiment one in most respects, there were several important differences. First, the presence of the example (problem one) along with rule S resulted in significantly better performance on problem two than when rule S was shown alone, the only case in either experiment where nonnegligible success was noted on an extra-scope problem. This effect may have been due to the form of the combining operation, '50 x 50,' in the rule S statement. '50 x 50' is clearly an instance of the more general SG combining rule, 'n x n = n^2.' Presumably, the statement of rule S, together with the common illustrative series, \(1 + 3 + 5 + \ldots + 97 + 99\), provided the successful S Ss with enough cues to generalize. In particular, they may have discovered that this series had 50 terms. Hindsight suggests that this difficulty might have been overcome by simply stating the sum, 2500, of the illustrative series rather than '50 x 50.'

Second, Table 1 indicates that only three of the 19 G-with-example Ss solved problem three whereas 18 solved problem one and 14 solved problem two. The decrement between problems two and three was significant (\(p < .003\)). The reason for this difference was not immediately apparent especially since 15 of these Ss applied rule G to the third problem. A more intensive post hoc analysis of the situation, however, suggested that the result may have been due to a difference in ease of determining \(N\), the number of terms, for use in the G combining rule, \([(A + L)/2]^N\). \(N\) could be determined from problem series one and two by taking the average of the first and last terms. A careful examination of the test papers suggested that this led to the incorrect value (25, rather than 24) for \(N\) in the third series, \(2 + 4 + 6 + \ldots + 46 + 48\). In short, the difficulty was not in the rule but in finding the correct value of \(N\). Such difficulties may be circumvented in future experimentation by controlling for such unwanted differences.\(^{13}\)

13It may be desirable to think of properties such as \(N\), as being derived from lower order (i.e., more easily discernible) stimulus

---

<table>
<thead>
<tr>
<th></th>
<th>Rule</th>
<th>Rule and Example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N)</td>
<td>one</td>
</tr>
<tr>
<td>Group S</td>
<td>20</td>
<td>8(8)</td>
</tr>
<tr>
<td>Group SG</td>
<td>20</td>
<td>5(5)</td>
</tr>
<tr>
<td>Group G</td>
<td>19</td>
<td>3(5)</td>
</tr>
</tbody>
</table>
Third, although the results of experiment two were in the hypothesized direction, only the overall effect of scope on interpretability was significant. This led us to wonder whether interpretability of the rule statements depended solely on generality. Could the rule statements also have differed as to the difficulty of interpreting the actual terms or symbols used? After consideration of this possibility, interpretability was rejected as an important factor in experiment two since a recheck convinced us that we had succeeded reasonably well in stating each rule as clearly as possible. Perhaps a more likely interpretation is that the Ss' familiarity with arithmetic interacted with the materials used so as to reduce the effects of statement generality.

Fourth, only one of the Ss who was shown the rule, 50 x 50, applied it to problems two and three. This result can probably be attributed to an interfering effect due to prior familiarity with addition problems. The Ss may simply have mistrusted rule S. How could a rule like 50 x 50, having only one answer, be the sum of all three problem series? Most junior high school Ss would find it unreasonable that the series 1 + 3 + ... + 97 (problem one) and 1 + 3 + ... + 79 (problem two) have the same sum (50 x 50). Some such reluctance may also have obtained on problem one with group S-without-example. Nonetheless, we were surprised that only 8 of those 20 Ss, not presented with the illustrative series, gave the correct sum (2500 or 50 x 50) for problem one.

3.3 Implications and theoretical comment.

The results of these experiments demonstrate, in a rather conclusive fashion, the behavioral relevance of rule generality. For the most part, successful performance was noted only on tasks within the scope of verbally stated rules. When rules are presented in an expository fashion, it is normally too much to expect generalization to problems to which the rule does not immediately apply (however, see Scandura and Durnin, 1968).

Of perhaps even greater practical significance were the lack (there was one exception) of performance differences on within-scope problems and the consistency results. The former result demonstrates that (almost) any stimulus within the scope of a rule is equally as difficult to respond to correctly as any other. Furthermore, coupled with the consistency data cited above, the obtained consistency results suggest that only one (new) test stimulus is needed to determine whether, in fact, a given rule has been learned. No more information is gained by using additional test instances. These results could have far-reaching implications for the development of highly efficient measuring instruments.

In addition, the pronounced tendency of the Ss to attack all of the test problems in the same way, irrespective of whether the procedure used was appropriate, suggests that the ability (i.e., knowing how) to solve problems and knowing when to make use of this ability to solve problems are quite distinct. Testing for the latter ability necessarily must involve the presentation of extra-scope problems.

Perhaps even more important than the results of these exploratory experiments were the post hoc analyses they made both necessary and possible. In particular, the results of these experiments indicated that the roles properties. Thus, the rule, \((A + L)/2\), worked for problems one and two whereas \(L/2\) was required for problem three.
played by various aspects of a principle statement need to be more clearly specified than the form, 'If I', then R' indicates. It does not detail all that appears relevant. For one thing, it was not possible in the rule generality study to distinguish between the roles played by A, L, and N (where A, L, and N have a particular meaning) and the algebraic expression \[(X + Y)/2\]Z (where the variables have general relevance). The former variables relate to properties (D) of the series stimuli, while the latter is a ternary operation (O) by which another such property (e.g., sums) may be derived. I', of course, although it played no role in the rule generality study, is also critical. It tells when, in fact, a rule can and cannot be applied. Thus, the rule, N^2, is appropriate whenever an arithmetic series consists of consecutive odd integers beginning with 1 while \[(A + L)/2\]N works whenever there is a common difference between adjacent terms.

These observations suggest that a principle statement might be represented more appropriately by the form, 'If I', then O'(D') = » R', where I' refers to the set of stimulus properties which indicate when the rule denoted O'(D') should be applied, D' refers to the set of those properties which determine the responses, and O', to the operation from which the responses, denoted by R', may be derived from the properties referred to by D'. This representation led naturally to the four-tuple characterization of a principle (I, D, O, R) and to the previously mentioned distinction between a rule and a principle.

Although the actual symbols used in a statement may be an important factor as has been suggested above and as will be demonstrated in the next section, the hypothesis advanced in the rule generality study to the effect that scope and learnability are inversely related finds a formal rationale in the nature of the characterizing elements. Making operational use, for example, of the arithmetic series property (i.e., dimension), 'the difference between adjacent terms is some common value,' necessarily presumes that 'the difference between adjacent terms is two,' '... three,' 'etc.,' can all be correctly interpreted. The converse does not necessarily follow. A similar relationship exists with respect to the rules, 50 x 50 and N x N. To correctly apply the latter, more general, rule to any particular series requires the ability to determine any value of the dimension N, including 50. Being able to apply 50 x 50, however, does not.

It would appear that the more general the rule the more is expected of the learner. Whether such differences will be reflected in behavior, however, may depend not only on rule generality but the population involved, particularly on whether the Ss have the necessary requisite abilities.

In effect, differences in generality appear, on analysis, to be equivalent to differences in abstraction level. Thus, the number two is more abstract than the property two oranges because the former applies to a collection of sets only one (subcollection) of which has the latter property. For the same reason, the property represented by the place holder X is more abstract than the number two since it refers to a still higher order collection. Unfortunately, we have not yet conducted a study designed to provide definitive information on these points. For the present, this analysis remains hypothetical.
4. INTERPRETABILITY AND SYMBOLISM

To help clarify the role symbolism plays in mathematics learning, I recently completed a study (Scandura, 1967b) with the help of John Davis in which we varied both the symbols actually used to construct statements of a principle and the ability of an S to interpret these symbols. It seems almost axiomatic that the ability to interpret a statement of principle depends critically on the ability to interpret the symbols of which it is composed, be they mathematical symbols or elements of the native language (e.g., English). Nonetheless, in mathematics learning the use of mathematical symbolism is frequently, if not always, preferred to ordinary English. The reason why, however, is never made explicit.

The purpose of this study was to determine whether: (1) principles are more easily memorized when stated symbolically or when stated verbally and (2) the ability to correctly use constituent symbols and the required (grammatical) combining rules is a necessary and/or sufficient condition for applying a learned (i.e., memorized) principle statement.

4.1 Method

Four artificial principles, each unfamiliar to the 24 Ss (college majors in elementary education), were selected for study. Each principle was based on one of the following notions: greatest integer, sigma notation for sequential addition, vector, and partial derivative. Two statements of each principle were prepared; one was composed of unfamiliar mathematical symbolism and the other of carefully, yet succinctly, worded English. For example, the greatest integer rule was stated (in English).

(1) Take the greatest integer in X.
(2) Take the greatest integer in Y.
(3) Divide the result of step one by the result of step two.
(4) Take the greatest integer in the quotient obtained in step three.

The symbolic form of this rule was \([([X] + [Y])/2]).

Tasks were designed to train the Ss to interpret the constituent symbols. For example, one set of tasks involved the greatest integer function (i.e., \([x, \lfloor x \rfloor]\) all real \(x\)). In addition, all of the Ss were required to demonstrate proficiency in the use of parentheses as a means of signifying the order in which binary operations are to be taken. These conventional rules of grammar were involved in all four principles. Of course, neutral materials were used to teach and assess proficiency in the use of parentheses. In no case did the pretraining or assessment include either a complete rule or one of its instances.

A 2 x 2 factorial design, with repeated measures, was used. Each S effectively served as his own control. One factor was the form in which a given principle was stated, symbolic or English. The other factor involved the presence or absence of training on the constituent symbols. Of course, the principles were counterbalanced over treatments so that each was used equally often under each of the four treatments. All other unwanted factors were randomized, including presentation order.

Separate measures of learning rate and interpretability were obtained. Learning rate was determined by presenting each principle statement for a fixed period of time for study and testing to see if the Ss could completely reproduce them in written form. All four principle statements were shown once before the next go-through (trial) began. Easy of
learning was determined by the number of trials it took to learn each principle to a criterion of two perfect reproductions in a row.

Interpretability was measured immediately after all of the statements had been well-learned. To demonstrate his 'understanding' of the statements, S. was required to apply each of the corresponding (underlying) principles to two stimulus instances. For example, one of the problems used to determine whether S. could apply the integer rule was stated simply, 'If \( x = 8.64 \) and \( y = 3.24 \) then ...'. All four principles were tested once before the second set of test problems was given.

### 4.2 Results

The results demonstrated quite clearly that: (1) symbolic rules are learned more rapidly, whether the constituent symbols are familiar or not (\( p < .01 \))—there were only 2 exceptions (out of 24) to this generalization; (2) rules, stated in symbolic form, are applied successfully if and only if the Ss have been taught how to apply the constituent symbols (and the necessary grammatical rules)—there were only 4 exceptions to the sufficiency part of this generalization and none as regards necessity; (3) rules, stated in the native English language, are applied equally well whether or not training in the use of the corresponding mathematical symbols is given; and (4) English statements, once learned, are applied equally as well (in this study, somewhat better) as symbolic statements in which use of the constituent symbols has previously been mastered.

These results are not entirely surprising but they do, nonetheless, make explicit at least one aspect of the role symbolism plays in mathematics learning. The use of symbolism makes mathematics learning more efficient when the constituent symbols and grammatical combining rules have previously been mastered. Symbols, of course, also serve the practical function of requiring less space in printing. Nonetheless, these results suggest that under certain conditions it may be well to remember that ordinary English can be used to teach mathematical ideas.\(^{14}\)

### 5. EXTRA-SCOPE TRANSFER IN LEARNING MATHEMATICAL RULES\(^{15}\)

The main purpose of this research was to help identify the underlying causes of generalization from one instance of a rule or strategy to another. Our more immediate aim was to provide evidence relevant to the interpretation offered in the Rule Generality Study to explain the one case of extra-scope transfer obtained (i.e., from the rule statement

\(^{14}\)These results provide a rational basis for making one type of branching decision that, while intuitively obvious, needs to be made explicit in computer-assisted instruction. Given the objective of learning a particular principle and an expository mode of instruction, one might proceed as follows: (1) test to see if S. can make use of the constituent symbols; (2) if so, present the principle in the more efficient symbolic form; (3) if not, present the principle in English.

Although learning feedback has long been recognized as an important factor in promoting efficient learning, it has been unclear as what sort of feedback to measure. The present results suggest that specific sorts of feedback are needed in order to make specific kinds of decisions.

\(^{15}\)This section is based on a paper presented by John Durnin and the author at the American Educational Research Association Convention in Chicago on February 10, 1968.
'50 x 50' to the statement 'n x n'). In essence, it was hypothesized that if a given rule or strategy is a restriction of a more widely applicable rule then a statement of this restricted rule may very well provide a basis for generalization to extra-scope items, which are instances of the more general rule or strategy. A restricted rule statement may be viewed as one obtained by replacing the variables entering into the statement of a more general rule with the specific values of a particular instance. For example, the restricted rule statement, '50 x 50', can be obtained from the statement, 'n x n', by replacing 'n' with the specific value, '50.'

A secondary purpose of the study was to obtain further information on the consistency hypothesis. That is, we wanted to find out whether a correct response to one instance of a generalized rule would imply success with other instances of the more general rule. Since the scope of a rule was directly related to the number of variables entering into the rule (i.e., the number of dimensions which were allowed to vary), it was hypothesized that if transfer to one instance indicates that a particular rule (e.g., '50 x 50') has been generalized along one or more dimensions (e.g., to 'n x n'), then transfer should be expected to additional instances which differ from the original only along the same dimension(s).

5.1 Method

The task used was again the number game 'NIM.' Restricted statements of the three game winning rules, S, SG, and G, used in Experiment One of the Rule Generality Study, were constructed. These restrictions were applicable only to (6, 31) games. Statement S' was essentially identical to the rule S statement. Statement SG' was a restriction of SG in the sense that the rule was restricted to one value (i.e., 31) of the 'desired sum (m)' dimension. It was stated, 'The appropriate first selection is determined by dividing 31 by 7. The remainder 3 should be your first selection.'

Statement G' was a restriction of statement G in that the rule was restricted to one value (i.e., 31 and 6, respectively) of the 'desired sum (m)' and 'size of selection set (n)' dimensions. It was stated, 'The appropriate first selection is determined by adding one to six (1 + 6), and dividing 31 by this result. The remainder 3 of this division is the selection which should be made first ... It is important to notice that 7 = 6 + 1.'

Three of the four treatment booklets included one of the restricted statements (S', SG', and G') together with a common (6, 31) game which was provided for practice. The fourth booklet served as a control. The first two test problems, 1A and 1B, were (6, 31) games. Test problems 2A and 2B were (6, m) games, which differed along the 'desired sum' dimension, with m = 25 and m = 29, respectively. Test problems 3A and 3B were (n, m) games, which differed along both the 'desired sum' and 'size of the selection set' dimensions, with n = 5, m = 26 and n = 7, m = 33, respectively.

The Ss were 88 West Philadelphia High School students enrolled in an academic mathematics program. They were randomly assigned to three experimental groups (S', SG', G') and a control (C) so that each group
Empirical Research

included 22 Ss.

Each S completed one of the four treatment booklets and the test in that order. The experiment was self-paced and with only a few exceptions the Ss completed the experiment well within the time limit of 40 minutes.

The criterion measure was use of the appropriate pattern (AP). S was given credit for using the AP if he won the game and employed an appropriate game winning strategy. All of the tests conducted were applied to 2 x 2 contingency tables.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Problem 1A</th>
<th>Problem 1B</th>
<th>Problem 2A</th>
<th>Problem 2B</th>
<th>Problem 3A</th>
<th>Problem 3B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Ss</td>
<td>(6, 31)</td>
<td>(6, 31)</td>
<td>(6, 25)</td>
<td>(6, 29)</td>
<td>(5, 26)</td>
<td>(7, 35)</td>
</tr>
<tr>
<td>Group C</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group S'</td>
<td>22</td>
<td>16</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group SG'</td>
<td>22</td>
<td>19</td>
<td>20</td>
<td>7</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Group G'</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2 Results and Discussion

Table 2 shows that restricted rule statements may provide an adequate basis for generalization. The three experimental groups performed at essentially the same level on the (6, 31) games (Problems 1A and 1B), but there were 12 Ss in groups SG' and G', as compared to none in group S', who were successful on problems 2A and 2B. This difference was significant at the .01 level.

A cursory review of the literature suggests that the transfer observed in a number of other studies may also have involved generalizing a restricted rule statement. Maier (1945), for example, found that providing S with a problem solving procedure, as it applied to one problem (i.e., with a restricted statement), improved the level of performance on a second problem (which was presumably within the scope of a more general rule). Some such generalization mechanism may also be involved in what some investigators have called 'remote transfer.' Thus, in a recent study, Wittrock's (1967) non-replacement strategy group was presented with a restriction of a general strategy which was applicable to his remote transfer items. Apparently, what these Ss actually learned (i.e., discovered) was the more general strategy.

The performance of the G' Ss, however, suggests that transfer can not necessarily be expected to all problems within the scope of the rule from which the restricted statement is derived. Of the five Ss in group G' who were successful on problems 2A and 2B, none was successful on problem 3A and only two, on problem 3B.
These differences between problems 2A and 2B and problems 3A and 3B suggest that the level of performance on transfer problems may depend on the particular dimension(s) involved. Problems 2A and 2B required that the G* statement be generalized only along the 'desired sum' dimension whereas problems 3A and 3B required generalization along the 'size of selection set' dimension as well. Apparently, the G' Ss were more capable of making the former generalization than the latter.

To test the consistency hypothesis, those Ss who used the AP on problems 1A, 2A, and 3A and those who did not (non AP users) were compared as to AP use on problems 1B, 2B, and 3B, respectively. There were significantly more AP users on problem 1A who were AP users on problem 1B than was the case for non AP users on problem 1A (p < .001). The same relationship held for problems 2A and 2B (p < .001) and problems 3A and 3B (p < .001), respectively. There were only four cases out of a total of 131 in which a non AP user (in groups S', SC', and G') on an 'A' problem became an AP user on the corresponding 'B' problem. There was only one case (out of 67) where an AP user on an 'A' problem was not an AP user on the corresponding 'B' problem.

These results suggest that if transfer obtains on one new problem, which differs (from the training problem) along one or more dimensions, then transfer may be expected to other problems which differ along these same dimensions.

Of course, the ease with which a correspondence can be determined between statement cues and the determining properties of an illustrative problem undoubtedly depends heavily on individual differences as well as on the nature of the cue. A major task of future research will be to determine what the important individual differences are.

6. ATTRIBUTE AND OPERATION CUEING IN THE DISCOVERY OF MATHEMATICAL RULES

So far we have limited our illustration and discussion to reception learning; that is, learning which takes place via the interpretation of symbolic statements. Much, however, is learned by discovery. The typical instructional procedure involves presenting, one at a time, either stimuli or stimulus-response pairs corresponding to the to-be-discovered rule. To determine whether learning has taken place, the learner is usually asked to give the appropriate response to new stimulus instances. The former type of situation, in which S is tested repeatedly, is exemplified by:

The sum of the first 2 odd integers, 1 + 3 = ?
The sum of the first 3 odd integers, 1 + 3 + 5 = ?
...

The learner would be expected to discover that the correct sum may be obtained by simply squaring the number of terms in the (odd integer) series.

16Other variants of reception learning are possible, such as learning from diagrams or pictures (icons), as in geometry, and from concrete objects, as with Dienes' (1967) multiple embodiments.
Ostensibly to speed the discovery process, the teacher or auto-instructor might cue the critical aspects of the stimuli. Thus, italicizing the number of odd integers (as above) or printing them in red would presumably attract attention by making the cues more salient. Of course, it would be equally possible to identify the appropriate operation (i.e., squaring) or, to introduce various combinations of both determining (D) and operation (O) cues.

Although a good deal of verbal and non-verbal cueing goes on during the contemporary discovery lesson in mathematics, there have been very few attempts to uncover the basic mechanisms involved. Whereas, recently, there have been a few such studies concerned with concept learning (Haygood and Bourne, 1965; Wittrock and Keislar, 1965) and, earlier, with problem solving (Maier, 1930), no consideration has been given to rules. This is indeed, unfortunate since rules seem to underlie so much of mathematics learning.

With this in mind, the members of my research seminar on mathematics learning conducted a pilot study to determine the effects of verbal attribute and operation cueing on the rate of discovering mathematical rules (Scandura, 1966d). In particular, the study extended that of Haygood and Bourne (1965) in two ways: (1) rules were used instead of concepts and (2) the operations involved were arithmetic, rather than logical. The study was designed simply to determine whether identifying the determining attributes (i.e., the class D of stimulus attributes which determine the responses) or the appropriate combining operation (i.e., O) does, in fact, increase the rate at which arithmetical rules are discovered.

6.1 Method

To minimize the effects of individual differences, artificial materials were again used. The stimuli were four-tuples of numbers (e.g., (4, 8, 9, 3)) and the responses were simply new integers that could be derived uniquely from exactly three of the four original integers by some combination of two (of the four) elementary arithmetic operations. Three rules were used. The determining characteristics and operations, respectively, were (1) $A_1, A_3, A_4$; (2) $A_1, A_2, A_4$; (3) $A_2, A_3, A_4$; and (1) $X + Y - Z$, (2) $X \cdot Y + Z$, (3) $X \cdot Y + Z$ where the subscripts, $i = 1, 2, 3, 4$, in $A_i$ refer to position in the four-tuple and $X$, $Y$, and $Z$ to place holders.

A 3 x 3 design, with repeated measures on the second factor, was used. Factor one involved the type of cue given (none, determining attribute (D), operation (O)). The second factor was a composite of the rule in question and the order of presentation but gave some indication of

---

17 Mike Bundrick, John Davis, Rosalie Jensen, Bob King, Frank Pavlick, and Larry Smith.

18 Although the order in which stimuli are introduced is also used extensively to promote discovery, this factor was not considered in this preliminary effort.

19 Since the study was designed as much as a learning experience (for my graduate students) as one of advancing knowledge, no attempt was made to remove this confounding, by counterbalancing principles over
the effects of one discovery on the next. The 36 elementary education majors were randomly assigned to one of the three cue groups so that a given S was exposed to only one type of cue. Each S completed three discovery episodes.

The Ss were told that their job was to write that number which they thought corresponded to the four-tuple shown. They were also instructed, 'There is a procedure by which you can always determine the corresponding number when I show you the set.' Then, the control group was told, 'To help you discover this procedure as rapidly as possible, you should try to determine the three specific positions in the four-tuple and a rule which combines the numbers in these positions to yield the corresponding number.' The attribute group was told, '... determine a rule which combines the numbers in the (proper positions inserted) to yield the corresponding number.' The operation group was told, '... determine the three specific positions in the four-tuple from which the numbers X, Y, and Z are always taken where (proper rule inserted) yields the correct number.'

After S responded to a given four-tuple, either by writing a number or by indicating he 'didn't know,' the card on which the four-tuple appeared was turned over exposing the same four-tuple together with the correct number response. The Ss were given approximately 12 seconds to compare the four-tuple with its solution before the next four-tuple was shown. Ten such four-tuples comprised one problem and S's score was the total number of correct responses made.

The probability of giving a correct response, by chance, without discovering the corresponding rule was relatively small. This was evidenced by the fact that once a correct response was given, S almost invariably gave the correct response thereafter.

6.2 Results

The results are summarized in Table 3.

Both attribute and operation cueing induced significantly (p < .01) earlier discovery of all three rules. But, whereas a significant improvement, presumably due to practice, was noted between problems two and three (p < .01), there was essentially no difference in performance on the first two problems. This is surprising since so-called 'warm up' effects typically have their greatest effect in the beginning.

More intensive comparisons, however, indicated that problem one practice significantly (p < .01) improved operation group performance on problem two but actually hindered (significantly, p < .01) attribute group performance—classic cases of positive and negative transfer. At least two possible interpretations may be given for the latter finding. First, having discovered that some combination of addition and subtraction presentation order. To have gotten too deeply involved in the details of design so early in their graduate research training could well have been self-defeating. I wanted them to think of research, first and foremost, as a conceptual experience rather than as a composite of technologies.
Table 3

Mean Number of Correct Responses

<table>
<thead>
<tr>
<th>Problem</th>
<th>One</th>
<th>Two</th>
<th>Three</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>6.6</td>
<td>4.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Operation</td>
<td>3.5</td>
<td>7.2</td>
<td>8.7</td>
</tr>
<tr>
<td>Control</td>
<td>1.1</td>
<td>0.6</td>
<td>4.0</td>
</tr>
</tbody>
</table>

(A₁ + A₃ - A₄) worked on problem one, many of the attribute-cue Ss may have spent too much time trying various combinations of addition and subtraction on problem two. Second, the rule, A₁ · A₂ + A₄, needed to solve problem two, since it involved multiplication and division as opposed to addition and subtraction, may have been intrinsically harder than that needed on problem one. Of course, both interpretations may have some degree of truth. Having been given the appropriate rules, in both cases, the operation-cue Ss were not subject to such effects. Furthermore, whatever response set developed on the basis of discovering that positions 1, 3, and 4 were relevant on problem one was more than compensated for by the practice afforded.

Clearly, research aimed specifically at such questions is needed to provide definitive information. It is impossible to say, at this time, that we fully understand exactly what is involved or, even more important, what the boundary conditions for these findings are.

Since specifying boundary conditions has all too frequently been passed over or, at most, been paid ambiguous lip service in educational research, I should like to emphasize one point about these results. It is doubtful that attribute cueing will ever be shown to be unconditionally better than operation cueing or vice versa. What future research may be expected to do, however, is to specify the conditions under which each will be superior.

7. DISCOVERY LEARNING

One of the fundamental assumptions underlying several of the new mathematics programs is that discovery methods of teaching and learning increase the student's ability to learn new mathematics. Indeed, this assumption has guided the development of many new curricula in all of the subject matter fields. Attempts to demonstrate advantages or disadvantages of self-discovery, however, have either failed, been open to criticism on scientific grounds, or are seemingly inconsistent even when apparently well-controlled.

Research on discovery learning has been confounded by differences in terminology, the frequent use of multiple dependent measures, and vagueness as to what is being taught and discovered (Scandura, 1964b). While the difficulties due to the use of inconsistent terminology can often
be minimized by a careful reading of research reports, the use of multiple dependent measures often makes it impossible to unambiguously interpret experimental results. Several investigators, for example, have found that groups which are given an expository statement of a rule perform better on transfer tests than groups which are required to discover this rule for themselves from instances of the rule. The obtained differences in transfer ability, however, may well have been because the discovery groups simply did not discover the rule.

Gagne and Brown (1961) overcame the dependent measure problem by equating original learning and investigating only transfer differences on new problems. On the basis of an analysis of the learning programs used in the Gagne and Brown study, Eldredge (1965) hypothesized that the obtained results could have been due to a number of flaws in the programs used. Eldredge proposed that exposition and discovery situations may be better characterized as differences in order of presentation. Exposition may be defined as giving rules and then examples of these rules, whereas discovery may be defined as giving the examples and then the rules. Contrary to his hypothesis, however, his discovery group did evidence more transfer than his exposition group. Unfortunately, there were a number of difficulties with the study that make the results difficult to interpret.

The Set-Function Language was used as an aid in removing these difficulties (Scandura, 1966a; Roughead and Scandura, 1968). The resulting analysis of what is involved in discovering rules indicates that discovery learners learn 'something' by which they can derive solutions to an entire class of problems. Roughead and I called this 'something' a derivation rule. Thus, discovery learners who actually succeed in making a discovery, should be expected to perform better than expository learners on tasks which are within the scope of such a derivation rule. If the new problems presented have solutions beyond the scope of a discovered derivation rule, however, there would be no reason to expect discovery Ss to have any special advantage.

This study was concerned with two basic questions. First, can 'what is learned' by discovery be identified and if so, can that knowledge be taught by exposition with equivalent results? According to the SPL, all behavior is controlled by rules so that there might well be some identifiable rule which is equivalent to 'what is learned' by discovery. Specifically, we hypothesized that 'what is learned' by guided discovery in the Gagne and Brown study could be identified and, hence, could be presented by exposition. The second question was, how is 'what is learned' by discovery dependent on what the learner already knows and/or the nature of the discovery treatment itself? More particularly, we hypothesized that the discovery of a derivation rule can actually be hindered by having too much prior information.

Assuming transfer depends only on whether or not the derivation rule is learned, sequence of presentation should have no effect on transfer so long as the subject is forced to learn the underlying derivation rule. That is, presenting the derivation rule by exposition or by guided discovery either before or after presenting the desired responses should have no effect on performance on transfer tasks. On the other hand, if a discovery program simply provides an opportunity to discover (with hints as to the solution) but does not guide the learner through the derivation procedure, sequence of presentation might well have a large effect on transfer. Assuming the learner is capable and motivated, he
may well succeed in determining the appropriate responses and, in the process, discover a derivation rule. It is not likely, however, that a person would learn such a derivation rule if he already knew the correct responses.

We made three hypotheses: (1) what is learned by guided discovery can be presented by exposition with equivalent results; (2) presentation order is not critical when learners are effectively 'forced' to learn derivation rules, either by exposition or by guided discovery; and (3) presentation order is critical when the discovery guidance provided is specific to the respective responses sought, rather than relevant to a general strategy or derivation rule.

### 7.1 Method

The task was essentially identical to that used by Gagne and Brown and Eldredge and involved finding formulas for summing the terms in number sequences. That is, the stimuli were number series, like \(1 + 3 + 5 + 7\), and the responses were formulas in \(n\), the number of terms, for summing such series. For example, the appropriate formula for summing

\[1 + 3 + 5 + 7 + \ldots + 2n - 1\]

is \(n^2\).

Using the SFL as a guide (i.e., by identifying, in turn, \(D\), \(0\), and \(R\)) we were able to identify that derivation rule taught in the guided discovery program used by Gagné and Brown. On the basis of this knowledge, four programs were constructed: (1) the formula-given program simply stated the correct summing formula for each problem series confronted in the learning program; (2) the guided discovery program remained essentially as it was in the earlier studies; (3) the expository program consisted of a precise expository description of that derivation rule which was presumably equivalent to that learned by guided discovery—it consisted of a general procedure by which the desired formulas could be derived; and (4) in the opportunity-to-discover program, the problem sequences were presented along with encouragement and hints as to what the desired formulas were. These hints involved such statements as 'the formula has a "2" in it.' The same number sequences were used in each of these four programs.

Seven treatments were constructed by combining these four basic programs. After going through a common introductory program, one group of subjects simply went through the formula program. The other six groups received the formula program together with one of the other three programs. Two of these six groups received the guided discovery program together with the formula program; two additional groups received the expository and formula programs; and the final two groups received the opportunity-to-discover and formula programs. One group, in each of the resulting three pairs, received the programs in one order; the other group received them in the reverse order. Only the order of presentation was varied. After finishing their respective programs all of the subjects were tested on new series to see how well they could determine the appropriate summing formulas.

### 7.2 Results and Discussion

The results were rather clear cut. Essentially, the group given the
formula program only and the group given the formula program followed by the opportunity to discover program performed at one level. The other five groups performed at a common and significantly higher level. Two points need to be emphasized.

First, 'what is learned' during guided discovery learning can, at least sometimes, be taught by exposition—with equivalent results. Of course, there are undoubtedly a large number of situations where, because of the complexity of the situation, 'what is learned' during discovery cannot be clearly identified. It is still an open question, for example, whether still higher order derivation rules, which have a more general effect on the ability to learn, may be learned by discovery. If we believe that the answer to this question is in the affirmative, then there is no real alternative to learning by discovery unless or until we can identify just what is involved. Nonetheless, intuition-based claims that learning by self-discovery produces superior ability to solve new problems, as opposed to learning by exposition, has not withstood experimental test. The value of some forms of discovery to transfer ability does not appear to exceed the value of some forms of exposition. Apparently, the discovery myth has come into being not so much because teaching by exposition is a poor technique as such, but because what has typically been taught by exposition leaves much to be desired. As we identify just what it is that is learned by discovery in more and more situations, we shall be in an increasingly better position to impart that same knowledge by exposition.

The second point to be emphasized concerns the sequence effect. While the group that was given an opportunity to discover and then the formula program performed as well on the transfer problems as those given the derivation rule in a more direct fashion, the group given these programs in the reverse order (i.e., the formula-opportunity group) did no better than those Ss given the formula program alone. In effect, if a person already knows the desired responses, then he is likely not to discover a more general derivation rule.

An extrapolation of this result suggests that if S knows a specific rule, then he may not learn one which is more general even if he has all of the prerequisites and is given the opportunity to do so. The reverse order of presentation may enhance discovery without making it more difficult to learn more specific rules at a later time. In effect, prior knowledge may actually interfere in a very substantial way with later opportunities to discover. In spite of this fact there may be some advantages inherent in learning more specific rules. Although the available data are not entirely clear on this point, it is quite possible that specific rules may make it possible to determine responses more quickly than rules which are more general.

This sequencing result may have important practical and theoretical implications. The practical implications will be attested to by any junior high school mathematics teacher who has attempted to teach the 'meaning' underlying the various computational algorithms after the children have already learned to compute. The children, in effect, must say to themselves something like, 'I already know how to get the answer. Why should I care why the procedure works?' Similarly, drilling students in their multiplication facts before they know what it means to multiply, may interfere with their later learning what multiplication is. Let me make this point clear, because it is an important one. It is not that meaning should be taught first simply out of some sort of dislike for rote
learning—for certain purposes rote learning may be quite adequate and the most efficient procedure to follow. The point is that learning such things as how to multiply without knowing what multiplication means, may actually make it more difficult to learn the underlying meaning later on. The theoretical implications are even more interesting for the researcher and, in fact, may be crucial to any theory based on the rule construct. More will be said on this topic in the final article in this series on theoretical direction.
COMMENTS ON SCANDURA'S APPROACH TO RULE-GOVERNED BEHAVIOR

MERLIN C. WITTROCK

. . . Scandura's attempt to begin to build a useful model of mathematics learning is exactly the kind of intellectual endeavor that should be encouraged in educational research.... Nonetheless, I feel that at this point in its development, Scandura's model should be related to different sets of data to refine it and sharpen it to the point that he can use it to make predictions about the effects of learning different kinds of ... rules.

Let's discuss some data while we ask of Scandura's model the following question: What new interpretations, predictions, or questions does the model bring to the data? I would like to present an experiment on problem solving.

The experiment (Wittrock, 1967) was concerned with verbal mediation in problem solving with young children. The children were taught strategies for solving problems involving the choice of one of four concepts: color, size, shape, or number. The task was one of correctly matching pictures on the basis of one of the four concepts. The pictures were presented on a screen, with one picture centered towards the top and the other two pictures below it, one to the right and the other to the left.

The children were randomly assigned to three groups. One group was taught a non-replacement strategy, a second group, a replacement strategy, and the third group acted as a control. The children in the non-replacement group were taught to hang four concept cards, which corresponded to the four concepts previously mentioned, on four hooks in front of them. Since the lower left picture would always match the top picture in two ways and the lower right picture would always match the top picture in the other two possible ways, the child was taught to hang the cards to represent the ways the right and left bottom pictures matched the top picture. The child then chose the bottom picture he thought would be the correct match. After being told the correct response, the child then took from the hooks and turned face down the two cards which represented the incorrect bases for matching. He used the two remaining cards on the next set of pictures. Again he made a selection, found the correct match and turned face down the card which did not represent the correct basis for matching. The one remaining card, which should now be

---

1Printed with permission of the author.
the correct choice, was used to match the pictures in the remaining sets. In brief, each child in this group was taught a most efficient way to eliminate incorrect bases for matching cards from his population of four possible basis. He learned to discard the incorrect bases by turning face down and discarding the cards representing these incorrect bases. It was predicted that this group would perform the best of all the groups in the experiment.

In the replacement group the procedure was the same as in the non-replacement group; as were the cards and problems. However, the children in this group were not taught to remove the cards from the hooks and discard them when they were finished with them. Each child kept the cards in front of him at all times.

Of course this does not necessarily mean that the child learned a replacement strategy. This procedure was just one way to operationally define such a strategy. The control group was told the four possible bases for matching the pictures, but they were not given on cards.

After training each of the groups was given four tests. The first test, a test of learning, presented the same concepts and instances used during experimental training. The second test measured transfer to new instances of the concepts color and shape. Each child was allowed to use his four concept cards on these two tests.

The third test was identical to the second except that the children were not permitted to use their concept cards. The fourth test measured remote transfer. The concept and instances of this test had never previously been covered in the experiment and no cards were permitted on this test.

<table>
<thead>
<tr>
<th>Treatment Groups</th>
<th>Learning 40 items (with cards)</th>
<th>Transfer to new instances 40 items (with cards)</th>
<th>Transfer to new instances 40 items (without cards)</th>
<th>Transfer to new concepts 26 items</th>
</tr>
</thead>
<tbody>
<tr>
<td>School I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(20) Nonreplace-</td>
<td>29.4</td>
<td>29.6</td>
<td>30.1</td>
<td>17.8</td>
</tr>
<tr>
<td>ment group</td>
<td>4.4</td>
<td>5.2</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>(20) Replacement</td>
<td>25.1</td>
<td>24.9</td>
<td>26.4</td>
<td>16.9</td>
</tr>
<tr>
<td>group</td>
<td>6.3</td>
<td>7.1</td>
<td>7.6</td>
<td>5.5</td>
</tr>
<tr>
<td>(19) Control group</td>
<td>22.4</td>
<td>20.6</td>
<td>20.9</td>
<td>14.2</td>
</tr>
<tr>
<td></td>
<td>5.2</td>
<td>4.7</td>
<td>6.5</td>
<td>3.6</td>
</tr>
<tr>
<td>School II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(20) Nonreplace-</td>
<td>25.7</td>
<td>25.0</td>
<td>27.7</td>
<td>16.1</td>
</tr>
<tr>
<td>ment group</td>
<td>8.5</td>
<td>8.5</td>
<td>6.5</td>
<td>5.0</td>
</tr>
<tr>
<td>(20) Replacement</td>
<td>22.9</td>
<td>24.0</td>
<td>23.3</td>
<td>14.4</td>
</tr>
<tr>
<td>group</td>
<td>6.5</td>
<td>6.1</td>
<td>6.2</td>
<td>5.2</td>
</tr>
<tr>
<td>(20) Control group</td>
<td>21.7</td>
<td>20.9</td>
<td>20.1</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>7.2</td>
<td>8.0</td>
<td>6.4</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Table 1
Means and Standard Deviation of the Test Scores
The results were almost exactly as had been predicted from a mediational theory of learning. For the learning test the nonreplacement group and replacement group each had significantly higher means than the control group ($p < .05$). On the transfer-to-new-instances test with cards the mean of the nonreplacement group was statistically greater than the means of each of the other two groups. The most interesting finding, though, occurs in the test of transfer-to-new instances without cards. The mean for the nonreplacement group was significantly greater than the mean of the replacement group ($p < .05$), and the nonreplacement group mean was significantly greater than the control group mean ($p < .05$). The children in the experimental groups learned to internalize the strategies well enough to use them without the cards. On the last test, transfer-to-new concepts without cards, the mean of the nonreplacement group was greater than either the replacement group or control group ($p < .05$).

The children in the experimental groups learned to use the strategies. They transferred the strategies to new problems, and the transfer was best for the nonreplacement strategy. In brief, the results were in close agreement with the predictions, especially in the transfer-to-new-instances test without cards.

The important question is, how does Scandura's conception of a rule, as he articulated it, relate to these data? Specifically, is his conception of rules useful for predicting or explaining results from experiments on children's problem solving? What new ways to look upon these data does his conceptualization give us?
Perhaps I should emphasize first that the "model" referred to by Wittrock is not a theory in the sense that few psychological assumptions are made. Rather, it is a language in which to describe complex human behavior. The only psychological assumptions are that: (1) all behavior is rule-governed, and (2) it is possible to identify some rule or combination of rules which can account for any conceivable kind of behavior in which a subject might engage. At the time Wittrock made his comments, the "model" said nothing about how rules are learned. It only provided a way of accounting for behavior of which a subject is capable at any given point in his learning.

Nonetheless, it is possible to interpret Wittrock's (1967) experiment in terms very similar to those I used earlier in explaining Haygood and Bourne's (1965) results. In the Wittrock study, the effective response involved picking the appropriate pictures from the displays presented. As before, the effective stimuli may be determined by asking what properties of the total situation are required to generate these responses. In this case, in order to produce a correct response, the subject must not only see the pictures involved, but must also know the concepts from which he is making a selection.

While it is possible to construct a single rule for solving such a task, it is instructive to consider this as two separate but related tasks. The first task involves identifying a basis for responding (i.e., for determining an appropriate concept). The second task involves using the concept for picking the appropriate picture. More specifically, in the first task, the effective responses are concepts. The effective stimuli are the four concept labels together with previous matching situations and the correct responses to them. An effective procedure would go something as indicated by Wittrock—that is: Eliminate two labels on the basis of the first answer, then, on the basis of the second answer, eliminate one of the two remaining labels.

1 This comment refers to my discussion of rules in Chapters 1 and 2. The extension of this "model" to a theory is described in Chapter 7 of this Volume and Volume I.

2 See Scandura (1972).
The second task involves responding to a matching problem (that is, a picture display) on the basis of the given concept. In this case, the effective responses involve picking an appropriate picture and the effective stimulus consists simply of the display. A procedure (rule) for solving such tasks might be: Encode the top picture, check the left one; if the concepts match, pick the left picture; otherwise pick the other picture.

What Wittrock taught the subjects (i.e., in the strategy groups), of course, was a composite of the rules for solving these two tasks. On this basis, it is not surprising that the non-replacement strategy group did the best since they were taught the strategy directly. The replacement strategy group had to devise an effective procedure for eliminating concepts on their own. The control group, of course, had even less to go on.

I would only add to this that it is not at all surprising that what Wittrock called remote transfer was evidenced. Indeed, the remote transfer tasks could be solved by direct application of the rules he taught. The fact that many of the subjects succeeded on them should have been expected (even, or particularly, in the absence of mediation theory).

In view of more recent work, I would also point out that there is an important difference between teaching a single rule for solving problems, as I have done in analyzing the Haygood and Bourne (1965) study (Scandura, 1972) and using two (or more) distinct rules as I have above. In fact, the learning mechanisms described later on in Chapter 7 spell out mechanisms by which such (distinct) rules may be combined to form new (single) rules which are adequate for solving such problems.
3

GRAPH THEORETIC MODELS
A psychologist who wishes to consider the nature of structured learning faces a number of difficulties. One is a serious lack of systematic empirical information about how structured knowledge is acquired and used by individuals. Most of the studies of learning that have been carried out in controlled laboratory situations have involved materials that have as little structure as the experimenter could manage in constructing a task for his subjects. But the lack of empirical information is clearly not the only impediment to psychological understanding of structural learning. It may not even be the major difficulty. For it is not clear that large amounts of empirical information about structural learning could be assimilated to the cognitive structures that psychologists have developed to deal with experimental data. In order that findings should lead to understanding, we need concepts for representing and interpreting the findings. The problem that I am concerned with in this chapter is the problem of representing cognitive structures. This involves what we may call conceptual research—the development of concepts that may be useful as a framework for psychological theory about cognitive structures.

Much of this chapter will be taken up with a relatively formal presentation of one system of concepts, consisting of an extension of mathematical graph theory. I think that these concepts provide a representation that avoids some of the difficulties that our ordinary psychological concepts leave us. On the other hand, there are some rather important deficiencies in the framework that I will present and I will point these out at the end of the chapter.

The purpose of presenting the main concepts of this chapter is not to develop a definitive framework for interpreting data about structural learning. It is, rather, to demonstrate that concepts relevant to the complex features of structural learning can be dealt with rigorously. The first step beyond associationism can be taken in a number of directions. This chapter presents one. And perhaps any such first step can only be expected to serve as a transition to other developments.

To introduce this discussion, consider an example of students' understanding of the relationship among weight, volume, and density of solid objects. Consider the following problem: "A block of metal weighs 40 gm, and its volume is 16 cubic cm. What is the density of the metal?" Such a problem is well within the competence of a student who understands
the concepts of weight, volume, and density.

To begin, I want to apply what I think is about as rich a theory of this problem as I can manage using concepts that are now part of common psychological lore. First, most students who can solve this problem probably know the formula, "Density = weight over volume." Many psychologists would describe this as a serial chain of responses. It would be reasonable to say that this chained response is connected to a stimulus that is part of the stated problem—probably the part stating that density is the variable to be calculated.

Once the student remembers the correct formula, he has to carry out some additional operations to solve the problem. First, he has to substitute the correct values for "weight" and for "volume" in the formula, and he has to carry out division. A stimulus-response analysis of arithmetic calculation is possible, though it is somewhat cumbersome and a more molar analysis in terms of operations or rules probably is preferable.

However, this chapter will not deal with the theory of calculations. Rather, we want to focus on the process by which the student comes to decide what calculations to perform. According to the stimulus-response theory, this process is simply memory of the correct formula. This is not to deny that rote memorization is a good description of some of the learning that goes on in schools and other places about problems like this. But it does not represent the kind of learning that most school programs strive for—and often achieve. For example, most teachers would expect students who could solve the first problem also to solve a problem like the following: "A block of wood with density 0.8 gm per cubic cm has a volume of 15 cubic cm. How much does the block weigh?" In the first problem, density was calculated from volume and weight. In this case, weight is calculated from density and volume. Of course, a stimulus-response analysis is possible again, where the student remembers the response, "Weight = density times volume." and then carries out the appropriate substitutions and multiplication. But this kind of analysis seems not to provide the right kind of insight into what our students learn to understand about density, weight, and volume.

The thing that seems to be missing from the stimulus-response analysis may be stated in a number of ways. One way is to notice that the two responses, "Density = weight over volume" and "Weight = density times volume" are related by simple algebraic transformation. In a sense, a student who understands either of these formulas probably knows them both, and this close connection should be represented in a theory about his knowledge.

Another difficulty with the stimulus-response theory stated here is a lack of attention to the relationship among the response components. Considering the utterance "Weight = density times volume" as a serial response tells us that a student can solve a particular kind of problem, but the theory doesn't say anything special about this kind of response. The sequence of elements might just as well be "m, n, o, p, q, r," or "The Times They Are A-Changin,' or any other well learned sequence. The thing that the theory should point out is that the formula represents a meaningful relationship among the concepts of weight, density, and volume.
Of course, the idea of a serial response is not the only concept in ordinary psychology that we can use to describe a student's problem-solving competence. There are possible analyses that describe associative relationships among components, but I feel that these miss the important features more seriously than the treatment given here using a serial response. And there are discussions that emphasize the relational character of the student's understanding, using concepts such as "reciprocal operations" or "structural understanding." Concepts such as these are useful in pointing out that some relational understanding is involved in students' competence; however, the concepts are vague, and I think that we need something that ties the nature of the relationships down more specifically in order to make real headway in understanding the cognitive processes that are involved.

The theory that I will present can be viewed as a generalization of psychological association theory. Psychologists are accustomed to dealing with associations between pairs of cognitive elements. Consider the case of paired associates. Let the integers 1, 2, and 3 stand for a short list of stimuli, and let the letters a, b, and c stand for their respective responses. If a subject learned this list, we could represent his knowledge by the graph shown in Figure 1. The arrows labelled $u_1$, $u_2$, and $u_3$ represent the three associations. The graph represents the fact that a trained subject could move from any of the three stimuli to its correct response.

![Figure 1](image)

Now, however, consider a junior high school student who has learned to solve the simple physical problem of calculating distance when he is given a rate of speed and a time spent travelling. Suppose the problem gives the speed as 30 mph and the time as 2 hr. Then the student obtains the answer as

\[ 30 \times 2 = 60 \text{ mi} \]

However, if another problem gives the speed as 50 mph and the time as 1 hr, the student obtains the answer as

\[ 50 \times 1 = 50 \text{ mi} \]

Ordinary association theory seems to miss two important aspects of
the student's knowledge. First, it seems patently incorrect to suppose that a student has specific associations of the form

\[
\begin{align*}
2 \text{ hr at 30 mph} & \rightarrow 60 \text{ mi} \\
1 \text{ hr at 50 mph} & \rightarrow 50 \text{ mi}
\end{align*}
\]

and all the other pairs for which the student could show successful performance. What the student knows is that distance is obtained by multiplying the speed by the time, and he knows how to multiply numbers. It seems more appropriate to describe his knowledge as a general operation of the form

\[d = s \times t,\]

in which many different pairs of numbers can be substituted for \(s\) and \(t\) and the student can calculate all the answers. Scandura (1968) has pointed out that knowing a rule of this kind permits a subject to select an appropriate member of an infinite range of responses for each member of an infinite domain of stimuli.

The second difficulty in applying association theory comes from the intuition that the elements of this problem are \(d, s,\) and \(t\). But if we view the situation in this way, there are no connections from one element to another. What the student can do is to move from the pair of points \(\{s, t\}\) to the point \(d\). Apparently we need a way of describing knowledge in terms of connections from sets to points. An appropriate graph is shown in Figure 2. The line labelled \(u\) represents the student's knowledge of the correct operation used to calculate the answer, in this case, multiplication.

\[\text{Figure 2}\]

From the point of view of mathematics, the present theory can be viewed either as a generalization or a special case of ordinary graph theory. A graph is a set of points \(X\) and a mapping \(F: X \rightarrow X\). The present theory has connections from subsets of \(X\) to points in \(X\). Then ordinary graph theory can be obtained by restricting the present theory so that the only subsets considered are singletons. On the other hand, we could use the concepts of ordinary graph theory and let the set of vertices be the power set of \(X, \mathcal{P}(X)\), with a mapping \(F: \mathcal{P}(X) \rightarrow \mathcal{P}(X)\).
Then the present theory would be obtained by restricting the usual graph theory by permitting connections only from sets to singletons.

The material presented in this chapter will be referred to as a theory, but it is not an empirical theory because it is not testable. In my opinion, the psychological and mathematical concepts that are available do not seem to provide an appropriate framework for theoretical development regarding problem solving. Association theory is too austere, and the main alternative, Gestalt theory, is too vague. However, the present theory follows Gestalt theorists such as Duncker (1945) and Wertheimer (1959) in emphasizing relational properties in problem solving. It might be accurate to say that this paper attempts to use some of the insights of Gestalt theory, but with an increased degree of formal rigor.

It would be desirable to have a framework with concepts that referred to all the important aspects of problem solving, but still fit together in a coherent and elegant system. Of course, the present effort does not achieve that goal. We should expect future research to show that some of the concepts developed here are unimportant, and that other concepts need to be added to describe important aspects of problem solving, even in simple situations of the kind used as examples in this chapter.

Many of the concepts of the present theory are taken directly from Berge (1962), often with only minor rewording of Berge's definitions. However, in some cases the idea of connecting sets to points changes the nature of the concept needed. In several places the development here follows Berge's closely. However, I have not tried to point out the detailed relationships between my development and Berge's discussion. A reader unfamiliar with ordinary graph theory should be able to follow the present argument without undue difficulty.

1. GENERAL DEFINITIONS

1.1 A graph is a 4-tuple

\[ G = (X, A, R, U) \]

where \( X \) is a set whose power set is \( \mathcal{P}(X) \), \( A \subseteq \mathcal{P}(X) \), \( R \subseteq X \), and \( U \) is a set of pairs of the form \( u(A, x) \), \( A \in A \), \( x \in R \). The subsets comprising \( A \) are the initial vertices of \( G \), and the points comprising \( R \) are the terminal vertices of \( G \). The pairs comprising \( U \) are the arcs of \( G \). For an arc \( u(A, x) \), \( A \) is the initial vertex of \( u \) and \( x \) is the terminal vertex of \( u \).

1.2 The arcs of \( G \) specify a relationship \( F: A \rightarrow R \) where

\[ F(A) = \{ x : \exists u(A, x) \in U \}. \]

That is, \( F(A) \) is the set of points that can be reached in one step from \( A \).

1.3 An arc \( u(A, x) \) is incident out from \( A \) and incident into \( x \) if \( x \notin A \). Let \( B \) be any set in \( \mathcal{P}(X) \), not necessarily a member of \( A \). Then \( u(A, x) \) is incident into \( B \) if \( x \in B \) and \( A \not\subseteq B \) and \( u(A, x) \) is incident out of \( B \) if \( A \subseteq B \) and \( x \notin B \).

The set of arcs incident into \( B \) is denoted \( U_B^- \), and the set of arcs incident out of \( B \) is denoted \( U_B^+ \). We say that the arcs comprising \( U_B \)
converge on B, and the arcs comprising $U_B^+$ diverge from B.

1.4 A path $\mu$ is a sequence of arcs $(u_1, u_2, \ldots)$ such that the terminal vertex of each arc is a member of the initial vertex of the succeeding arc. The set of arcs used in $\mu$ is denoted $U_\mu^+$; the set of initial vertices of $\mu$ is $Q_\mu$ and the set of terminal vertices of $\mu$ is $R_\mu$. A path may be finite or infinite. A finite path may be denoted by its first initial vertex and its last terminal vertex:

$$\mu = (u_1, \ldots, u_k) = \mu[A_1, \ldots, x_k].$$

The length of a path is the number of arcs in the sequence. A path is simple if it does not use the same arc twice, and composite otherwise. A path is elementary if no terminal vertex is a member of an initial vertex of an arc used earlier in the path, that is, for every $u_j \in U_\mu$,

$$\bigcup_{i=1}^{j-1} A_i \neq x_j.$$

It may be noted that in an elementary path, all the terminal vertices must be distinct.

1.5 The set of paths incident into a set B is

$$M_B^- = \{u[A_1, x_k] : x_k \in B \text{ and } (A_1 - B) = \emptyset \text{ for all } A_1 \in Q_\mu\}$$

and the set of paths incident out of B is

$$M_A^+ = \{u[A_1, x_k] : A_1 \subseteq B, x_k \notin B, \text{ and } (A_1 - B) \neq \emptyset \text{ for } A_2, \ldots, A_k \in Q_\mu\}$$

The paths comprising $M_B^-$ converge on B, and the paths comprising $M_B^+$ diverge from B.

EXAMPLE: Consider the proof of a simple theorem in probability theory:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

One proof uses the axiom that for any disjoint sets, the probability of their union is the sum of their probabilities. Call this Axiom a, and then a proof proceeds as follows:

1. $P(A) = P(A \cap B) + P(A \cap \overline{B})$, by a.
2. $P(B) = P(A \cap B) + P(\overline{A} \cap B)$, by a.
3. $P(A) + P(B) = 2P(A \cap B) + P(A \cap \overline{B}) + P(\overline{A} \cap B)$, by 1, 2.
4. $P(A \cup B) = P(A \cap B) + P(A \cap \overline{B}) + P(\overline{A} \cap B)$, by a.
5. Theorem, by 3, 4.

A graph describing this reasoning process is shown in Figure 3. The point a represents the axiom, the numbered points correspond to the steps in the proof, and t stands for the theorem. $u_1, u_2,$ and $u_3$ represent the process of realizing that the sets $A, B,$ and $A \cup B$ can be
partitioned in the appropriate ways, \( u_4 \) is addition, and \( u_5 \) involves comparison of the right-hand sides of the expressions given as steps 3 and 4.

![Figure 3](image)

In Figure 3, \( u_5 \) is incident out of the set \( E \), and \( u_4 \) and \( u_3 \) are incident into \( E \). Then

\[
\begin{align*}
\mathcal{U}_E^- & = \{u_3, u_4\}, \\
\mathcal{U}_E^+ & = \{u_5\}.
\end{align*}
\]

The graph in Figure 3 is not a path. One path is \((u_1, u_4, u_5)\) with length 3 which is simple and elementary.

A set of paths incident into \( E \) is

\[
\mathcal{M}_E^- = \{(u_1, u_4), (u_2, u_4), (u_3)\}
\]

and the paths in \( \mathcal{M}_E^- \) converge on \( E \). The arc \( u_5 \) may be considered as a path incident out of \( E \).

\[
\mathcal{U}_E^+ = \{(u_5)\}.
\]

This example can also illustrate the use of the theory to describe alternative ways of solving a problem. For example, the axiom about the probability of disjoint unions often is stated in the form:

\[
\text{if } A \cap B = \emptyset, \quad P(A \cup B) = P(A) + (P(B).
\]

Then a new point \( a' \) can be used to represent the axiom, and one way to proceed after Point 3 of the earlier proof is

\[
\begin{align*}
4'. \quad P(A \cup B) & = P(A) + P(\overline{A} \cap B), \\
5'. \quad P(A \cup B) & = P(A \cap B) + P(A \cap \overline{B}) + P(\overline{A} \cap B), \text{ by 1, 4'}, \\
6'. \quad \text{Theorem, by 3, 5'}.
\end{align*}
\]
The graph for this proof is shown in Figure 4.

Figure 4

Still other variations are possible, of course. Referring to Figure 3 again, suppose that a student did not see how to combine the results given in Points 3 and 4 of the first proof. Then a teacher might have to explain by subtracting \( P(A \cap B) \) from both sides of the result in Point 3, giving an expression exactly like the right side of Point 4. This would introduce another step,

\[
5''. \quad P(A) + P(B) - P(A \cap B) = P(A \cap B) + (P(A \cap \overline{B}) + P(\overline{A} \cap B), \text{ by 3, 4},
\]

\[
6''. \quad \text{Theorem, by 4, 5''},
\]

and the graph is shown in Figure 5.

Figure 5
1.6 A circuit is a path \((u_1, ..., u_k)\) where \(x_k \in A_1\). A circuit is elementary if for each \(u_j \in U_u\), the path \((u_{j+1}, ..., u_k, u_1, ..., u_{j-1})\) is elementary.

A loop is a circuit of length 1.

EXAMPLE: The graph in Figure 6 is a circuit. Let \(p\) stand for the unit price of a commodity, and let \(n\) be the number of units sold. Then \(s\) can be the total sales. Let \(r\) be the percentage of total sales that constitute the profits of the company; then \(f\) can be the company's profits. Let \(t\) be the amount of profit per unit. Then the number of units sold can be calculated from \(f\) and \(t\). The operations are simply

\[
\begin{align*}
  u_1 : s &= p n, \\
  u_2 : f &= s r \\
  u_3 : n &= f / t
\end{align*}
\]

![Diagram](image)

Figure 6

1.7 A subgraph of \((X, A, R, U)\) is obtained by specifying \(A' \subseteq A\) and \(R' \subseteq R\) giving a graph

\[G' = (X, A', R', U')\]

where

\[U' = \{u(A, x) : u \in U, A \in A', x \in R'\}\] \((1)\)

A partial graph of \((X, A, R, U)\) is obtained by specifying a subset \(U' \subseteq U\), giving a graph
A partial subgraph of G is obtained by specifying $\mathcal{Q}' \subseteq \mathcal{Q}$; $R' \subseteq R$, and $U'' \subseteq U'$ as given in Equation 1.

EXAMPLE: Consider the graph shown in Figure 7. The intended interpretation of the points is as follows: $l$, $w$, and $h$ stand for the length, width, and height of a rectangular solid. $v$ and $m$ stand for the volume and mass of the solid. $d$ is the density. $a$ and $f$ stand for acceleration and force. The arc $u_1$ corresponds to the formula

$$v = l \cdot w \cdot h.$$ 

$u_2$ corresponds to the formula

$$d = \frac{m}{v}.$$ 

$u_3$ corresponds to the formula

$$f = m \cdot a.$$ 

$u_4$ corresponds to the formula

$$m = \frac{f}{a}.$$ 

The intended use of subgraphs is to describe systems where only some of the concepts or points of the graph are functional. Consider the following problem: "A rectangular solid is 6 cm long, 4 cm wide and 2 cm high. It weighs 40 gm. Find its density." The solution of this problem can be represented by a subgraph of Figure 7, since acceleration ($a$) and force ($f$) do not play a role in the problem. In addition, a psychological question is whether a student reading the problem would identify the weight given as the mass needed for calculating density.
Suppose that he did not. Then his knowledge about the situation might be described by a subgraph specified by
\[ U' = \{u_1\} \]
\[ R' = \{v, d\} \]
according to definition 1.7 we have
\[ U' = \{u_1\} \]
The main problem is that \( B \not \in A' \) so the set needed to get to \( d \) does not exist.

Now consider another student, given the same exercise. Suppose that this student understands that weight and mass are substitutable, but he fails to see that volume should be obtained as an intermediate calculation. We could describe his cognitive situation by obtaining a subgraph specifying
\[ R' = \{d\} \]
\[ A' = \{A, B\} \]
Then
\[ U' = \{u_2\} \]
The intended use of partial graphs is to describe systems where only some of the needed operations are available. Suppose, in addition to the information given earlier in the example, the student is told that when the solid is acted on by an unknown force, it accelerates at a rate of 24 cm/sec\(^2\), and is asked to calculate the force. Further, he is asked to check his answer by then calculating the mass using the force and acceleration. A student might have the arc \( u_3 \) available and apply the correct operation to calculate the force. However, it is possible (though hopefully, unlikely), that he learned that formula in such a way that he did not see how to work back to mass. Then \( u_4 \) would be unavailable. We could describe his knowledge of the situation using a partial graph specified by
\[ U' = \{u_1, u_2, u_3\} \]
Then there is no arc connecting \( D \) and \( m \) in the partial graph.

It should be noted at this point that relationships among the arcs of a graph certainly are important. For example, in Figure 7, arcs \( u_2 \) and \( u_4 \) both represent operations of arithmetic division, and arcs \( u_1 \) and \( u_3 \) both represent operations of multiplication. Relationships like these can lead to interesting hypotheses about the processes of solving and learning how to solve problems, and the relationships can be represented easily by considering subsets of the set of arcs \( U \). However, for formal purposes it seems best to recognize that the operations are different (for example, dividing mass by velocity is not exactly the same thing as dividing force by acceleration). Some informal discussion about relationships among arcs will be given in later sections, but
formal analysis of these relationships is beyond the scope of this article.

2. ANALYSIS OF PROBLEMS AND PROBLEM SOLVING

The concepts of a theory of graphs on sets can be used to analyze either specific problems or general relationships among concepts. For example, Figure 8 represents the kind of knowledge about distance, time, and speed that a junior high school student is expected to acquire. But a specific problem involving distance, rate, and time is represented by Figure 2. In this section some concepts involving specific problems will be presented.

2.1 A problem is a pair of sets \((V, W)\) where \(V\) is a set of values given in the problem and \(W\) is a set of unknowns to be obtained.

2.2 Let \(J = (V, W)\) be a problem, and let \(G = (X, \mathcal{Q}, R, U)\) be a graph with domain \(X \equiv V \cup W\). A graph of \(J\) in \(G\) is a partial graph or partial subgraph of \(G\), specified by an ordered set of arcs.

\[
U_J = \{u_1(A_1, x_1), \ldots, u_m(A_m, x_m)\}
\]

such that \(U_J \subseteq U\), \(W \subseteq \{x_1, \ldots, x_m\}\), and for each initial vertex \(A_1 \subseteq V \cup \{x_1, \ldots, x_{i-1}\}\). If a graph satisfying these requirements exists in \(G\) then we say that the problem is solvable in \(G\).

**EXAMPLE:** Let \(G\) be the graph shown in Figure 7. \(V = \{l, w, h, m\}\), \(W = \{d\}\), and \(J = (V, W)\). Then a graph of \(J\) in \(G\) is the partial graph specified by the arcs \(U_J = [u_1(A, v),\ldots, u_m(A, v)]\).
Both $u_1$ and $u_2$ are arcs of $G$. Furthermore, $A \subseteq V$, and $B = \{m, v\} \subseteq V \cup \{v\}$.

In applications, a person's knowledge about some concepts may be represented by a graph $G$. We may consider a problem $J$ presented to that person. The first question is whether the problem is solvable, given that person's knowledge. Other possible questions include the number of different solutions that this person might have for the problem, or the smallest number of arcs needed to obtain a graph of $J$ in $G$.

If a problem is solvable by a person, a second question is whether the person will find a method of solving the problem, and how he goes about searching for a method of solution. In the present framework this process can be viewed as trying to construct a graph of the problem using the components of $G$, where $G$ represents what the person knows. One way to view this is as a search among the components of $G$ for elements that will contribute to solution. In one kind of problem a simple search seems to provide a relatively efficient way to construct a graph, and it guarantees that a graph will be found if it exists.

2.3 Let $J = (V, W)$ be a problem, let $G = (X, \mathcal{A}, R, U)$ be a graph, and let $G_J = (X, A_J, R_J, U_J)$ be a graph of $J$ in $G$. For each $w_i \in W$, $G_J$ may join $J$ at $w_i$ or it may be separate from $J$ at $w_i$. $G_J$ joins $J$ at $w_i$ if $U_J$ includes an arc $u(A, w_i)$ such that $A \subseteq (V \cup W)$. Otherwise, $G_J$ is separate from $J$ at $w_i$. If $G_J$ joins $J$ at $w_i$ for all $w_i \in W$, we say that $G_J$ is completely joined to $J$.

If $G_J$ is completely joined to $J$, then $G_J$ can be found by the following algorithm. Let $U_G$ be the arcs of graph $G$. The symbols "+" and "-" indicate additions to and deletions from sets.

1. Set $V' = V$, $W' = R_G = W$, $A_J = U_J = \emptyset$
2. Does $W' = \emptyset$? If yes, exit. If no, go to 3.
3. Select an element $w_i$ from $W'$.
4. Search in $U_G$ for an arc $u_i(A_i, w_i)$ such that $A_i \subseteq V'$. If one is found, go to 5. If not, go to 2.
5. Set $W' = W' - w_i$, $V' = V' + w_i$, $U_J = U_J + u_i$, $A_J = A_J + A_i$.
6. Go to 2. When the program exits, $A_J$, $R_J$, and $U_J$ will contain the initial and terminal vertices and the arcs of $G_J$.

If the problem $J$ does not have a graph in $G$ that is completely joined to $J$, then the problem of finding a graph becomes more complicated. The algorithm given above takes each unknown and searches for a way to calculate it. If the solution of a problem is represented by a graph that is separated from the problem, then one or more quantities used to calculate an unknown is not mentioned in the problem. This means that the search described in the algorithm will not succeed in every case. One
possibility would be to carry out a more detailed search; for each arc 
\((A, w_i)\), if \(x \in A\) but \(x \notin (V \cup W)\), search for an arc \((B, x)\) such that 
\(B \subset (V \cup W)\). Such a procedure would guarantee solution for separated 
problems where each quantity needed to solve the problem can be obtained 
using quantities which are mentioned in the problem. (We might describe 
such problems as having a degree of separation equal to one.) But the 
search would take a long time. It seems likely that an individual would 
depart from a systematic search and use a procedure that would permit a 
solution in less time. Such a procedure might well be available, but 
the individual might have to give up the guarantee of finding a solution.

The algorithm and the accompanying discussion above deal with stra-
tegies that are anchored on the unknowns of a problem. A graph for a 
problem probably can be constructed using the information as a starting 
point. A simple search strategy anchored to the information probably 
would be quite unwieldy for all but the simplest problems. The reason 
is that each arc in the graph of a problem involves a single quantity to 
be calculated, but some combination of quantities must be used as infor-
mation. The unknowns can be used one at a time in searching for arcs, and 
for each unknown \(w_i\), the question is, "Can I find an arc incident into \(w_i\)?" 
However, the information has to be used in combinations, so sets \(A_i\) have 

to be considered, asking, "Can I find an arc incident out of \(A_i\)"? The 
question might have to be asked about each set of items of information, 
rather than just about each item. This means that the number of possi-
bilities can be extremely large, and for this reason it seems reasonable 
to expect search strategies to be anchored to the unknowns.

As an alternative to searching among the arcs of \(G\), a person may 
generate the graph using some schema. An example is dimensional analysis, 
commonly used for simple physical problems. For example, consider the 
following problem: "A man drove at a speed of 60 miles per hour for 
four hours. He used 15 gallons of gasoline. Calculate his mileage in 
miles per gallon." A graph of this problem is shown in Figure 9. Arc \(u_1\) 
corresponds to the formula \(d = st\), and \(u_2\) corresponds to \(m = d/g\).

![Figure 9](image)

The person's knowledge about these concepts probably is complicated 
enough to require a graph like Figure 10. It is conceivable that an 
individual would carry out a search of the arcs of this graph, but another 
possibility is to consider the problem in the form:

\[
s = \frac{60 \text{ mi}}{\text{hr}}, \quad t = 4 \text{ hr}, \quad \frac{60 \text{ mi}}{\text{hr}} \times 4 \text{ hr} = 240 \text{ mi} = \ell.
\]

\[
d = 240 \text{ mi}, \quad g = 15 \text{ gal}, \quad 240 \text{ mi} / 15 \text{ gal} = 16 \frac{\text{ mi}}{\text{ gal}} = m.
\]
In this method of solving the problem, the answer is obtained by considering combinations of the units of measurement. The graph of the problem seems to be constructed piecemeal, and the construction seems to move from the information to the unknowns.

3. DESCENDENCE RELATIONS

The first part of this section presents some concepts that could be used to describe a person's state of knowledge. The main ideas have to do with the extent to which a set of concepts is tightly organized in the person's understanding of them.

3.1 The relation $\leq$ will be associated with points on a graph as follows. For two points $x$ and $y$, $x \leq y$ if and only if $x = y$ or $x$ is a member of some $A \in \mathcal{A}$ and there is a path from $A$ to $y$. Read "$x \leq y$" as "$x$ directly precedes $y$," or "$y$ directly follows $x$.

If $x \leq y$ and $y \leq x$, then $x \equiv y$, which is read "$x$ is equivalent to $y$." If $x \leq y$ but not $y \leq x$, write $x < y$, which is read, "$x$ is a direct ancestor of $y$," or, "$y$ is a direct descendent of $x." Note that the relation $\leq$ is transitive.

3.2 For a set $B$, a vertex $z$ which directly follows every member of $B$ is called a direct majorant of $B$. If a direct majorant of $B$ is a
member of $B$, the point is called a direct maximum of $B$. If two points $y$ and $z$ are both direct maxima of $B$, we have $y \leq z$ and $z \leq y$; hence $y = z$.

A direct minorant of $B$ is defined as a point that directly precedes at least one point in $B$, and which is not the direct descendent of any point in $B$. That is, $y$ is a direct minorant of $B$ if there is a $z$ in $B$ such that $y \leq z$, and there is no $x$ in $B$ such that $x < y$. A direct minorant of $B$ which is an element of $B$ is called a direct minimum of $B$. Two direct minima of $B$ are not necessarily equivalent; they may not be connected by the relation $\leq$.

Note that if $u(A, x)$ is an arc of a graph then any point that is in $A$ but is not a terminal vertex is a direct minimum of the graph.

3.3 A graph (or path) is strongly connected if for every pair of points $x$ and $y$, either $x < y$ or $y < x$, or both. If every point precedes every other point, then for every pair $x \leq y$ and $y \leq x$. Such a graph is called an equivalence graph. Clearly, an equivalence graph is strongly connected. In general, a graph may be strongly connected but not be an equivalence graph.

Remark on Circuits: if a graph (or a subgraph) consists of a single circuit, they if $y$ is a terminal vertex and $x$ is any other point, $x \leq y$. Therefore, in a circuit, all terminal vertices are direct maxima and are equivalent. Any point that is not a terminal vertex directly precedes all the terminal vertices, but is not a direct descendent of any points. Therefore, in a circuit, all points that are not terminal vertices are direct minima.

If a graph (or subgraph) consists of a single circuit, then if the graph is strongly connected, it is an equivalence graph. To see this, let $x \in A$ be a point in a circuit that is not a terminal vertex. Since every initial vertex $A$ must contain a terminal vertex, there is a point $z \in A$ that is a terminal vertex, but $x \not< z$ and $z \not< x$. Therefore, if a circuit is strongly connected, then all of its points are terminal vertices. But in a circuit, every terminal vertex has to directly precede every other terminal vertex, and therefore, all the points are equivalent.

Remark on Equivalence Graphs: In any equivalence graph every point directly follows every other point. Therefore, there must be at least one arc $u(A, x)$ for each $x \in X$. In other words, $R = X$.

Consider an arbitrary point $y$ in an equivalence graph. Since $y$ directly precedes every other point in the graph, there is at least one arc $u(A, z)$ with $y \in A$. But for each such arc, $z$ directly precedes $y$. In other words, there is a path $u_1[A, ..., y]$ where $z \in B$. But then there is a path $u_2[A, ..., y]$ whose first arc is $u(A, z)$, and $u_2$ is a circuit. Furthermore, since every point in $X$ directly precedes every point in $X$, $u_2$ is a strongly connected circuit. It follows that if $G$ is an equivalence graph with domain $X$, then every point $x \in X$ is contained in a strongly connected circuit.

EXAMPLE: Figure 10 represents an equivalence graph. Note that every point is the terminal vertex of at least one arc. This particular graph contains two strongly connected circuits—one with the points $d$, $s$, and $t$, and the other with the points $d$, $s$, and $m$. 

88 Greeno
An equivalence graph represents what we might call "perfect knowledge" about some concepts. If an empirical scientist had a theory about some concepts and the theory was represented by a strongly connected graph, he would be able to specify the effect of changing any variable on the value of every other variable. This kind of knowledge exists for some, but not all, systems of variables. And even where it does exist, an individual's understanding may not be complete. Therefore, a representation is needed for connectedness in a weaker sense.

3.4 The relation \( P \) will be associated with a graph as follows: For two points \( x \) and \( y \), \( xPy \) if and only if \( x \leq y \), or \( x \) and \( y \) are both members of some \( A \in \mathcal{Q} \), or \( x \) directly precedes some point \( w \) and \( w \) and \( y \) are both members of some \( A \in \mathcal{A} \). Read \( xPy \) as "\( x \) precedes \( y \)" or "\( y \) follows \( x \)." If \( xPy \) and \( yPx \), write \( xQy \), which is read, "\( x \) and \( y \) have equal precedence." If \( xPy \) but not \( yPx \), write \( xTy \), which is read, "\( x \) is an ancestor of \( y \)" or "\( y \) is a descendendant of \( x \)." Clearly, if \( x < y \) then \( xPy \) and if \( x = y \), then \( xQy \), but not necessarily conversely. \( x < y \) implies \( xPy \) but \( x < y \) does not imply \( xTy \).

The relation \( P \) is transitive.

3.5 For a set \( B \), a majorant of \( B \) is a point that follows every point in \( B \). If a majorant of \( B \) is a member of \( B \), it is a maximum of \( B \). All maxima of \( B \) have equal precedence.

A minorant of \( B \) is a point that precedes every point in \( B \). If a minorant of \( B \) is a member of \( B \), it is a minimum of \( B \). All minima of \( B \) have equal precedence.

If a point \( x \) is a direct majorant (or a direct maximum) of a set \( B \), \( x \) is a majorant (or a maximum) of \( B \). However, many points that are direct minorants (or direct minima) of \( B \) are not minorants (or minima).

3.6 There is a link between two points \( x \) and \( y \) if \( xPy \) or \( yPx \) or both. A link is denoted by \([x, y]\). A chain is a sequence of links \([x_1, y_1], ..., [x_k, y_k]\) where in each link, \( y_i = x_{i+1} \).

3.7 Two points \( x \) and \( y \) are connected if there is a chain from \( x \) to \( y \). A graph (or path) is connected if all pairs of distinct points are connected. If a graph is strongly connected, it is connected, but not conversely.

**Example:** Using the relation \( P \) and the concept of a chain, we obtain a greatly weakened notion of connectedness. An example is in Figure 11. There are direct connections between only a few pairs of points. However, the graph is connected. For example, there is a chain between \( e \) and \( g \):

\[ 
\mu = ([e, b], [b, f], [c, g]).
\]
Comparing the notions of connectedness and strong connectedness, we might say that in a strongly connected graph, we can move from any point to any other point, where each step involves an operation. In a connected graph, we can move from any point to any other point, but some steps may involve simply realizing that two concepts "have something to do with each other." In fact, in order for two points not to be connected, they must be elements in completely distinct structures. This is emphasized in the following property.

Given a point $x$, denote by $Y_x$ the set of points that are connected to $x$, together with $x$ itself. A connected component is the subgraph specified by the set $Y_x$. The components constitute a partition of $X$, and a graph is connected if and only if it possesses only one component. (This property states Theorems 1 and 2 in Berge (1962), and his proofs can be applied here with a slight change in notation.)

**EXAMPLE:** The relation $\leq$ seems most likely to be applied in analyzing the order of operations in solving a problem. The relation $P$ seems to be potentially useful in analyzing the order in which information is used in solving a problem. Consider the graph in Figure 12. The problem represented is solved by a straightforward series of calculations, starting with

$$v = l w h,$$

and ending with

$$f = m a,$$
The problem leads to a single point, \( f \), which is the graph's single maximum. (In this case, \( f \) is also the direct maximum of the graph.) It also has a well-defined starting point, and the minima of the graph are the points in \( A \). By comparison, notice that the points \( l, w, h, d, \) and \( a \) are all direct minima. However, the relation \( \preceq \) does not apply to many pairs of points, even though there is a clear ordering in their use. For example, \( \preceq \) does not apply to the pair \( (h, d) \) nor to \( (w, a) \).

The lower panel of Figure 12 represents a problem that does not have a well-defined starting point. The definition of \( P \) seems to be appropriate in relation to problems like this, since \( P \) does not apply to pairs like \( (w, f) \), and therefore the graph does not have a minimum. The indifference of order between operations \( u_1 \) and \( u_2 \) is expressed by the fact \( v \notin m \) and \( m \notin v \). Again, the graph has a direct maximum, in this case \( d \).

### 4. SUMMARY AND DIFFICULTIES

The theory presented above permits connections from sets to points,
rather than merely from points to other points. A typical simple structure is in Figure 2, where u is an arc connecting the pair of points (s, t) to the point d. This structure could be used to represent a student's ability to work from values of time and speed, and calculate distance. The points of the graph represent concepts or variables, and the arc represents an arithmetic operation: multiplying speed by time.

A representation like Figure 2 begins to allow us to represent the relational structure of responses that are used in simple problem solving. Another move is needed to represent the complex reciprocal relationships that often occur. When a student knows how to get distance from speed and time, he probably knows how to get speed from distance and time, and how to get time from distance and speed.

One kind of representation for this is in Figure 8, where speed and time connect to distance, as in Figure 2, but also distance and time connect to speed, and distance and speed connect to time. This kind of structure is called a circuit; any two pieces of information can be used to obtain the third one.

It seems safe to say that most of the teaching that we try to do involves circuits. Certainly in the case of simple applied mathematics of the kind represented in these graphs, we intend that students should understand the structures in a symmetrical way. On the other hand, our teaching may not always be successful in this regard, and the graph-theoretic framework provides a convenient representation of the cases where reciprocity is not achieved. Suppose that we discovered a student who could calculate distance, given speed and time, and could calculate speed, given distance and time, but who could not calculate time, given distance and speed. This could be represented by a partial graph of the graph in Figure 8. The concept of a partial graph is well defined in the framework of graph theory, and in a psychological theory, one of its uses would be to represent structural knowledge that is faulty or incomplete.

Another use of partial graphs is in describing a method of solution for a particular problem. The graph in Figure 8 represents the knowledge that we might hope to communicate to a student about weight, volume, and density. Now suppose that we give the student a problem like "Find the distance traveled if a man drives 4 hr at 50 mph." The method of solving this problem is represented by the graph in Figure 2. Notice that Figure 2 shows a partial graph of Figure 8. If a student really has something corresponding to Figure 8 in his mind, then before he can solve a problem he has to extract the relevant partial graph from the graph that contains his knowledge about the relevant concepts.

The idea of extracting relevant concepts and operations from a complex graph provides one theoretical representation of a process that many writers have mentioned, but few have illuminated. The process is involved when a person who has been thinking about a problem suddenly realizes that he will be able to solve it, and often this occurs before the details of a solution are clear in his mind. In a complex problem, requiring many steps, the task of finding a combination of sets and arcs may take some time and effort. But it may be reasonable to represent this as a process of finding the components of a partial graph. The individual's feeling that he "sees how to solve the problem" would correspond to the achievement of completing a partial graph that connects the information.
The discussion of this chapter has dealt with the problem of describing a kind of structured knowledge that is only a little more complicated than simple associations. I have argued that the best job we can do with ordinary psychological concepts has fatal weaknesses in that the description fails to capture the relational nature of the knowledge and does not permit any natural way of capturing the reciprocal nature of this knowledge. And I have described a system of concepts that seems to be helpful in providing a description of at least the simple cases that I have used as examples. The main move needed to get to this theory is a concept of connections between sets and points, and this extended version of graph theory seems to be capable of describing interrelated structures. It also is suggestive of hypotheses about the way in which we and our students go about finding a method of solution to a problem.

The graph theoretic concepts used in this paper are only a short step beyond those of ordinary association theory in psychology. There are some features of problem solving and understanding that need a richer set of concepts than these. One aspect of the graph theoretic system seems particularly weak to me. This aspect is the absence of any direct way of specifying the nature of connections between particular sets and points. The system works mainly on the basis of connections that are present or absent. The kind of connection is not expressed in an intrinsic way. Of course, the arcs of the graph can be labelled, and one could construct a system of labelling and classifying arcs that would carry the needed information about the different kinds of connections in the graph. But this seems like a superficial way of handling the problem. It would be preferable to have information about the kinds of connections built into the structure of the graph itself.

A second difficulty with the theory of graphs on sets is that it is extremely cumbersome. A graph like Figure 3 is clumsy enough—but with as many as nine or ten concepts and all of their interconnections the diagram would become virtually impenetrable. It would be desirable to have a representation that communicated the information more simply.

A set of concepts that seems promising in relation to these problems is closely related to a theory that Lamb (1966) and Reich (1968) have developed in another context, called a relational network. This theory was developed to account for language comprehension and generation, using a system called stratificational grammar. Our adaptation of the ideas uses the basic structural concepts, though the objects in the system are quite different.

An example of the kind of structure in a relational network is shown in Figure 13. The elements s, t, and d are connected by a relation that is labelled by the symbol at the intersection. The structure in Figure 13 represents the same knowledge as does Figure 3, only much more simply. From any pair of points, the network permits a move to the third point. But the operation used depends on the direction that one needs to move through the intersection. Going through in the direction of the arrow points involves multiplication. Going through in either of the other directions involves division.
The idea that I am suggesting is that we may be able to give a reasonably economical representation of a person's structural knowledge using a relational network. Of course, there will be a very large number of concepts, and they will be connected by different kinds of relationships. The different relationships merely require different symbols at the intersections of the network.

I have not examined the idea involved in the relational network as thoroughly as I have the idea based on graph theory. However, I think that the network probably will represent the same kinds of processes as the theory of graphs on sets, and will do it more simply. It also seems to allow a representation of another aspect of problem solving that can only be represented rather awkwardly in the theory of graphs on sets. This involves solving a problem of a kind that a person has not solved before. The discovery of new relationships can occur by design in a successful classroom effort. Consider the following theoretical sketch of what may happen when a teacher uses the discovery method successfully. The phenomenon to be explained is the occurrence of a new relational structure. Keep in mind that the student already has a supply of basic operations, and has connections that involve these operations. I think it is possible that the way in which a student is brought to discover a new relationship involves similarities and analogies with structures that he already knows.

As an example, suppose that a student has learned about speed as a rate of distance travelled per unit of time. And he knows that in that case, the formula relating the variables is

\[ \text{speed} = \frac{\text{distance}}{\text{time}} \]

Now consider the process of getting a student to discover the relationship

\[ \text{density} = \frac{\text{weight}}{\text{volume}}. \]

This might be accomplished by explaining that density is a rate of weight per unit of volume. Then by generalizing from the structure that he already knows, the student could discover the new formula.
What I have tried to do in this chapter is to sketch two ways in which cognitive structures may be represented in psychological theory. It goes without saying that these informal speculations are a long way from a satisfactory theory. A great deal of refinement in relation to experimental research and classroom observation will be needed before either of the primitive frameworks mentioned here becomes a theory. But I believed that both frameworks—and especially the relational network—have sufficient intuitive appeal to be worthy of the effort needed to test their usefulness as conceptual bases for the psychological representation of structured knowledge.
SOME PRELIMINARY EXPERIMENTS
ON STRUCTURAL LEARNING
JAMES G. GREENO

One task of a theory of structural learning is to give a description of subprocesses involved in the learning of structured materials. The experiments that I will describe here were designed to provide some preliminary information relevant to this kind of analysis for two cases of structural learning. The first study dealt with junior high school students learning to solve story problems about motion and speed. The second experiment dealt with college undergraduates learning about elementary statistics.

These studies involve exploration of three possible factors in learning how to solve the kinds of problem that we used in the experiments. The factors were partly suggested by the conceptual frameworks described earlier, and partly by intuition. One factor is understanding of the abstract mathematical concepts involved in the problems. A second factor involves techniques of applying the mathematical ideas in specific problem contexts. And a third factor involves understanding of the physical concepts that the problems involve. A slightly more specific description of these three factors is given in Table 1.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Motion/Speed</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract Mathematical Concept</td>
<td>Reciprocity of Multiplication and Division</td>
<td>Sets and Functions</td>
</tr>
<tr>
<td>Problem-Solving Technique</td>
<td>Construction of Equal Ratios</td>
<td>Exponentiation and Combinatorial Notation</td>
</tr>
<tr>
<td>Physical Concept</td>
<td>Simultaneous Measurement of Distance and Time</td>
<td>Generation of Random Variables</td>
</tr>
</tbody>
</table>
In both of the experiments we used transfer designs to assess the effectiveness of various pretraining procedures in facilitating instruction in problem solving. In each case, a pretest was given to all students to see how easily they could solve problems of the type to be studied. Then the subjects were divided into groups with each group receiving a different experimental treatment. Next, all subjects were given the same instruction in the method of solving a kind of problem. Finally, subjects were tested to see how much they had learned during the common instruction.

A graph showing the experimental design is shown in Figure 1. The idea of the design is to permit inferences about the kinds of cognitive structures involved in learning how to solve the problems. For example, if a group receiving training in abstract mathematical concepts performed better on the post-test than the control group, we could infer that improved preparation in the abstract concepts facilitated structural learning in that situation; hence that acquiring the abstract cognitive structure was a significant component of the structural learning that went on.

![Figure 1](image)

1. EXPERIMENT I

This experiment was carried out using seventh-grade mathematics classes at University School in Bloomington, Indiana. I am grateful to the administrators and teachers of University School for their cooperation, and to David Lloyd and E. Glen Carr for their assistance in conducting the experiment.

Three different techniques were used to prepare students for instruction regarding word problems about time and motion. The techniques used were: (a) presentation of some relatively abstract materials dealing with the reciprocal nature of multiplication and division; (b) presentation of a problem-solving technique based on making two ratios equal to each other; and (c) demonstrations using model trains to provide physical illustrations of the relationships among time, speed, and distance. A fourth
group (d) received none of the treatments and provided a control.

1.1 Procedure

Three classes of average seventh-grade mathematics students were used. On the first day a brief pretest was given with seven problems. The problems were of varying difficulty, ranging from, "A man drove at a speed of 60 miles per hour for 4 hours. How far did he drive?" to, "The distance between Bloomington and Chicago is 240 miles, and there are two airline flights between the two cities. One flight is nonstop and takes one and one-half hours. The other flight stops for one-half hour in Terre Haute, but also takes one and one-half hours including the stop. How fast does each plane fly?"

On the second day, students in each class were divided into four experimental groups, matched insofar as possible on the basis of their scores on the pretest. These groups were taken to separate rooms and each group received one of the following training procedures:

(a) The abstract training was given using a concept called a "product set," which was defined as a set of three numbers, one of which was the product of the other two. The experimenter-instructor explained this definition and took students through several steps, including discrimination of sets with and without the defining properties of product sets, generating sets with the defining properties, and recognition that when one number is the product of two others, either of the other two numbers is the ratio of the product and the remaining number. The final test of understanding was to present several pairs of numbers and have each subject complete the product sets in all three possible ways—that is, from the numbers a and b, product sets can be made as follows: (a,b,ab), (a,b,a/b), (a,b,b/a).

(b) The group given training in a technique was instructed in how to set up and manipulate equations of the form $a/b = c/d$ in solving a variety of word problems involving rates. The instruction emphasized that each side of the equation equaled the same number. Students were to find what the numerical value of one of the ratios was and then try to find a value of an unknown that would make the ratio containing the unknown equal to the other ratio. The instruction identified the notion of "per" or "rate" with a ratio having one in the denominator. Problems and examples were worked out for the students, and they filled in worksheets including a total of 10 ratio equations and four word problems using the technique.

(c) The third group observed time and rate relationships in demonstrations using two 110 guage model train engines. The trains moved down parallel stretches of straight track marked off in one foot intervals (see Figure 2). A timer, visible to the students, started as one of the engines passed the starting point. The engine travelled at nearly constant speed until it passed a second photoelectric cell which stopped the timer and the engine or engines. The speeds of the engines were variable from .5 ft/sec to 3 ft/sec. A speed regulator with a dial calibrated from 1 to 3 ft/sec was available for the students to set the speed of one of the trains. The photocell that stopped the train(s) could be placed at any of the footmarkers along either track. The accuracy of the calibrations was adequate; if the train speed was set at 1 ft/sec
and the train travelled the full 12 ft, the time was within .5 sec of 12 sec. Thus, the speed and distance settings always produced a time that was accurate to the nearest second.

In an experimental session, the experimenter-instructor pointed out the various features of the apparatus and demonstrated the effects of varying speed and distance (less speed or more distance produced a greater time). Students' attention was called to the correspondence between the clock-time and the position of the train as it moved down the track. Tabulations of distance, speed, and time were kept on a blackboard. Students were asked to calculate the time that would be used for a given distance and speed, and to calculate the distance needed to use a given time with a given speed, and to calculate the speed after a trip with a given distance and a measured time. These calculations were all very simple, and nearly all the students were able to do them quite easily. Problems involving both trains moving simultaneously at different speeds were solved toward the end of the session.

The control group spent the hour that was used for the experimental treatments in a study hall doing their own studying or reading.

Each experimental procedure was used for about one-fourth of the students in each of the three classes. The experimenter who presented a given procedure was changed from one class to the next, so that each of the three main experimenters was exposed to different groups of students.

On the third day of the experiment, students returned to their classrooms and were given instruction in solving the time and motion problems that were in the pretest. One experimenter (J.G.) acted as instructor for all three classes. The instruction emphasized the formula: "distance = speed times time." An effort was made to relate this instruction to the content of all three of the experimental treatments. After about 25 minutes of instruction, the students took a final test containing
seven problems. These final problems were as similar to those in the pretest as possible in that the same operations were required for solution but different kinds of objects were used (for example, boats instead of airplanes) and, of course, different numerical values were used.

1.2 Results

Both the pretest and the final test were scored by giving one point for each correct answer, and one-half point for problems worked correctly but with arithmetic errors or with minor errors of labelling. We apparently constructed a set of problems of about the right difficulty. On the pretest, the scores ranged from zero to seven, with a mean of 3.6.

<table>
<thead>
<tr>
<th>Condition</th>
<th>All Problems</th>
<th>Hard Problems</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract Training</td>
<td>0.93</td>
<td>0.27</td>
<td>15</td>
</tr>
<tr>
<td>Problem-Solving Technique</td>
<td>0.96</td>
<td>0.15</td>
<td>13</td>
</tr>
<tr>
<td>Physical Demonstration</td>
<td>1.57</td>
<td>1.21</td>
<td>14</td>
</tr>
<tr>
<td>Control</td>
<td>1.00</td>
<td>0.57</td>
<td>14</td>
</tr>
</tbody>
</table>

The matching of students in the different groups was quite accurate; the mean scores in the four groups were 3.7, 3.1, 3.7, and 3.9. However, the differences were large enough to persuade us to use difference scores in evaluating the treatments. Table 2 shows the mean improvement in the number of problems solved in the four experimental groups. The data for all problems included the seven problems of the test. The data for hard problems included four of the seven problems that were more complicated. The differences between the group receiving the physical demonstration and the other groups probably were reliable. Nonindependent t tests were carried out, and the results were as follows: Comparing Demonstration vs. Control, t(26) = 2.12, p < .05; Demonstration vs. Abstract Training, t(27) = 3.18, p < .01; Demonstration vs. Technique, t(25) = 3.43, p < .01. No other comparison showed a significant difference between conditions.

2. EXPERIMENT II

This experiment was carried out using students in a section of Introductory Laboratory Psychology at Indiana University as subjects. I am grateful to the instructor of the section, Coleman Merryman, for his cooperation, and to E. Glenn Carr and David Lloyd who planned and carried out the experiment.

Three different techniques were used to prepare students for instruction in elementary statistics. The techniques used were: (a) presentation of some relatively abstract materials dealing with sets and sample spaces;
(b) training in the use of combinatorial notation and exponents; (c) physical demonstrations of frequency distributions using dice and cards. A fourth group (d) received none of the treatments and provided a control condition.

2.1 Procedure

Fifty-seven students in Introductory Laboratory Psychology served as subjects. A pretest was given to all of the students, including 15 questions involving general mathematical operations such as raising numbers to powers and manipulating fractions; 10 items concerning probability in situations like coin tossing and dice rolling; and five questions involving concepts of inclusion and intersection based on a Venn diagram. Scores on the pretest ranged from 12 to 28 with a mean of .20.

Subjects were divided into four groups, matched on scores obtained in the pretest. Experimental training procedures were given to all of the students in a group at the same time.

(a) Students receiving abstract training received a mimeographed handout containing definitions of "set," "subset," "sample space," "union," "intersection," and "Cartesian product." In addition, the number of elements in a set was discussed, and coin-tossing and dice-rolling examples were mentioned and used to illustrate the ideas. The experimenter went through the handout with the group of students, and each student worked a set of 10 exercises.

(b) In the group given technical training, exponents were introduced, including the facts \( X^0 = 1 \), \( X^m \cdot X^n = X^{m+n} \), and \( X^m/X^n = X^{m-n} \). Factorials were introduced, and combinatorial notation was defined. The students worked on two sets of exercises, one involving exponents and the other involving binomial coefficients and probabilities.

(c) The group receiving physical demonstrations observed the experimenter throw three dice for 108 trials, they guessed the suit of a cut card 36 times, and they observed poker chips drawn from a bag 108 times. In each case a record was kept of the results and simple statistics were computed.

(d) The control group was given free time during the session when the other groups received the experimental training.

Following the experimental sessions, the course instructor gave his standard lectures in introductory statistics. These included discussions of chance guessing in tests, simple outcomes of coin-tossing and dice-rolling, the sign test, the rank test, and simple descriptive statistics including concepts of mean, median, mode, range, skew, and cumulative frequency, and various ways of representing data. A 43 point test was given over these materials.

2.2 Results

The items in the final test were divided into six sets of eight items each and separate scores were obtained for each set. The first eight items were about chance guessing. A second set were about coins and
Some Preliminary Experiments on Structural Learning

A third set dealt with inference based on the sign and rank tests. A fourth set dealt with concepts of descriptive statistics. A fifth set involved the number of elements in various sample spaces. And the last set dealt with ways of representing data, including two questions specifically about the binomial distribution. The test was multiple choice, and the answers were scored as follows: the subject was given one point for the best answer, he was given one-half point if he chose an answer that was partially correct, and he was penalized one point if he chose an answer that was clearly false. There was a partially correct alternative for each of the 48 questions, and there was a clearly false alternative for 20 of the 48 questions.

Table 3
Mean Scores on Sections of Final Statistics Test

<table>
<thead>
<tr>
<th>Training Condition</th>
<th>Pretest</th>
<th>Guess</th>
<th>Coin-Dice</th>
<th>Inference</th>
<th>Descrip. Space</th>
<th>Repres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>20.7</td>
<td>3.6</td>
<td>4.3</td>
<td>2.3</td>
<td>5.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Technique</td>
<td>20.7</td>
<td>3.7</td>
<td>4.6</td>
<td>2.4</td>
<td>5.1</td>
<td>3.2</td>
</tr>
<tr>
<td>Physical</td>
<td>20.4</td>
<td>3.4</td>
<td>3.6</td>
<td>1.9</td>
<td>5.2</td>
<td>2.6</td>
</tr>
<tr>
<td>Control</td>
<td>20.5</td>
<td>2.6</td>
<td>3.5</td>
<td>2.3</td>
<td>5.3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

The mean scores for each group on each set of items are given in Table 3. There is no evidence that the experimental training procedures facilitated learning with regard to the questions on inference, concepts of descriptive statistics, or sample spaces. The technical training using combinatorial notation, factorials and exponents may have facilitated learning regarding the problems about chance guessing, coins and dice, and representing data. The group with technical training was significantly better than the control group on the questions about guessing (p < .05) and nearly so on the questions about coins and dice, and about representing data (p < .10). The group with abstract training seems to have done about as well as the group with technical training on the questions about guessing and about coins and dice, but this group's scores were not significantly higher than the control group's scores.

3. CONCLUSIONS

The main purpose of these studies was to investigate whether gross problem solving performance could be manipulated using variables that seem reasonable in relation to the conceptual framework previously presented. Despite the informal classroom conditions of the experiments and the lack of precision in the measures of experimental effects, the results do suggest that the concepts and distinctions discussed previously may provide useful insight into problem-solving processes.

In relation to the concepts introduced in this paper, the three kinds of experimental training used in the experiments might be described as follows. The abstract training might provide general structures which
could have facilitated learning of the specific structures needed to solve the problems used in the experiments. This apparently did not occur. Partly as a result of these data, we are inclined to think that general structures may contribute mainly to transfer between different but related kinds of problems.

Technical training of the kind used in these studies might be related to the formation of specific structures—the particular relational operations used in solving the problems presented in the experiments. This apparently was effective in the experiment dealing with statistics, though not in the experiment on time and motion problems. One reason for the difference might be that in the statistics experiment the technical training was quite specific to the kind of problem to be used in the experiment. In the time and motion experiment the technical training involved a rather general notion of rates, including the time-motion situation only incidentally. Another possibility is that the operators used in the statistics problems probably were unfamiliar for most of the subjects, while the seventh grade students all knew how to multiply and divide quite well before we worked with them.

One possible effect of the physical demonstrations is to provide facilitation of the search or construction of a graph of a problem. The demonstrations using electric trains for time and motion problems apparently facilitated learning, and two features of the demonstrations seem to support the idea that they aided the students in forming the interface between the verbal problem and their mathematical solutions. One feature was the use of elementary problems in conjunction with the demonstrations. In the statistics experiment, students did not solve problems as they observed the demonstrations, and this might have contributed to the lack of effect in that experiment. Another possibility is based on a theoretical speculation. It seems reasonable to suppose that the process of finding a method of solution to a problem depends on the extent to which the various concepts used in the problem fit together in an integrated cognitive structure. In the demonstrations of time and motion the subjects watched the simultaneous change of time and position, and related this to speed in an explicit way. I am inclined to think that the opportunity to integrate these variables in direct observation may have sensitized the subjects to relationships among the concepts of time, speed, and distance in a way that facilitated translation of the verbal problems into forms in which solutions could be found.

There is no question whether the data of these preliminary experiments justify any firm conclusions about the nature of problem solving or methods of training students to solve problems. They do not. However, the results are useful in permitting a test of the relevance of the graph-theoretical concepts presented earlier to observations that can be obtained in experiments. If we accept these results at face value, we might be led to the opinion that elementary statistical ideas are hard for underclassmen in college because they are unfamiliar with the relevant operations and therefore have difficulty in acquiring the appropriate specific structures needed to solve problems. And it seems possible that word problems about simple physical systems are hard for junior high school students because they have difficulty in finding or constructing the method of solution for specific problems. These conclusions may or may not hold up in future investigation, and if they do it will be important to test their generality and refine their concepts. But the fact that they can be stated and seem to communicate possibilities about cognitive structures seems to auger for the potential usefulness of the concepts presented earlier.
4

PIAGETIAN MODELS
INTELLECTUAL GROWTH AND
UNDERSTANDING MATHEMATICS
KENNETH LOVELL

In the view of the Bourbaki group of mathematicians, mathematics is the
study of structures or of systematic patterns of relationships. This
being the case, the topic discussed in this paper is, I suggest, funda-
mental to the mathematics educator in the school years K through Grade
12.

This paper is divided into three parts. In the first I wish to indicate
very briefly, some salient features of the conceptual framework inside
which I shall discuss the later sections. In the second part I review
some of the evidence obtained from research carried out into pupils'
understanding of mathematical ideas. Finally, I deal with some of the
implications for teaching in the classroom.

1. ASSUMPTIONS MADE IN THE PAPER

I turn then to part one. The Thirty-Second Yearbook of the N.C.T.M.
(Jones, 1970) deals with the history of Mathematics Education in the U.S.
and Canada. Scattered throughout the book are references to psychologi-
cal and educational theories and their impact on mathematics education.
The name of Piaget occurs on seven pages and some of his books are named,
but the volume does not spell out the precise ways in which his work is
of value to the mathematics educator. Personally, I have been greatly
influenced by his work, details of which have been published in a vast
array of books and papers over almost 50 years, although it would be
true to say that it is only in the last 10 years that there has developed
a widespread interest in his work in the United States. I believe that
his position regarding the acquisition of certain kinds of new knowledge
is of more value to the mathematics teacher than any other position at
the moment, although I equally affirm that his theory does not cover all
the facts and that one day it will be replaced by or subsumed under a
more all-embracing one. Some of the strengths and inadequacies of this
theory in relation to 1970 mathematics learning have been given by Lovell
(1966) and Beilin (1970). While it would be wrong for me to outline
Piaget's cognitive developmental system here, I must just make six points
which are the assumptions, so to speak, on which the remainder of my
paper rests.
108

Lovell

(1) Piaget has been concerned with the development of the general ways of knowing, or the intellect—whatever term one uses. It seems that such cannot be taken directly from the blackboard, textbook or film by mere perception, or acquired by drill; rather these general ways of knowing have to be actively constructed by the child through interaction with the environment. When forged, these are never forgotten in mental health. For example, the child never forgets that a subclass is subsumed within a class, that if A > B and B > C then A > C. Against this, Piaget distinguishes certain kinds of particular knowledge that derive primarily from the interaction between the individual and specific aspects of the environment; indeed, all kinds of teaching procedures may be employed. But the quality of the general ways of knowing or of the intellectual structures, determines the manner in which the particular knowledge is assimilated. Moreover, such knowledge may be forgotten at any time and have to be revised.

(2) At the core of the central mechanism of intelligence, there are in Piaget's view the basic operations of uniting, seriating, equalizing, putting into one-to-one correspondence, etc. These ultimately stem from actions in the first 21 months of life, internalized with the help of language but not deriving from the latter, thus yielding implicit mental actions which eventually emerge as reversible and integrated structures around 7 to 8 years of age. The distinction he makes between physical experience and logical-mathematical experience is also important. In the latter, knowledge comes not from the objects themselves but from the actions performed on objects, as when the pupil finds that five groups of three objects yield the same total as three groups of five objects. He has to keep a constant check on the coordination of his actions to avoid contradictions. In short, the pupil reflects on his activity in an auto-regulatory sense. It is this kind of experience which aids intellectual growth although it is not the only type of experience to do so as we shall note later. Moreover, intellectual structures are reorganized and linked with others even in the absence of environmental stimulants.

(3) For Piaget, thinking is not a representation nor a descriptive event. Rather it is an action which early on overtly, and later covertly, transforms one reality state into another thereby leading to knowledge of the state. In his view, to understand a state one must understand the transformations from which the state results. It is, of course, the implicit mental actions or the covert transformations which are important in mathematics education.

(4) The concept of stage is a key one for Piaget. It indicates successive developmental periods of intelligence such as sensorimotor, pre-operational, concrete and formal operational. Each stage is characterized by a relatively stable structure that incorporates earlier structures into a higher synthesis. Pinard and Laurendeau (1969) have made an explicit and detailed defense of the stage construct as used by Piaget. Experience leads me to suggest that we accept this concept of stages with the associated changes in levels of understanding in respect to new ideas to be assimilated. Of course, I well recognize transitional stages at the borderline between one broad stage and the next, and I well recognize the unsolved problem of horizontal differentials. I also testify that the elicitation of thought characteristic of the formal operational stage is affected by familiarity with content, and credibility.

(5) In the view of the Geneva school, language plays a role but not a central role in the growth of thought; although I would submit that we
are only at the beginning in understanding the relationship between language and thought. At the level of formal thought it is clear that propositional logic needs some inner verbal support, but the power of propositional logic is not fundamentally due to this support.

(6) When the mathematical idea to be learned depends on a level of logical thought beyond that which the child possesses, the idea is either partially learned, or learned with much difficulty, and his grip on the idea is tenuous (Beilin, 1970).

2. REVIEW OF RESEARCH

Having listed six major assumptions, I turn to the second part of my paper. I will attempt to review the evidence which was accumulated over the last 25 years or so in respect to pupils' intellectual growth and their understanding of mathematics. A radical selection of the literature must be made, but I will try to give the gist of what has been determined.

Piaget and Szeminska's (1941) classic book *The Child's Conception of Number* gave details of experiments and data obtained relating to cardinal and ordinal correspondence, to the ability of the child to differentiate and coordinate the ordinal and cardinal aspects of number, and to the beginnings of basic additive and multiplicative properties of number. In essence their view was that the concept of the natural numbers comes from the fusion of two logical entities, namely class and asymmetrical relation; with classes, relations, and numbers evolving around 7 to 8 years of age in a tightly locked mutually interdependent way. This book sparked off much research and some acute observation in the classroom. The upshot is that an understanding of classes, relations and numbers are elaborated at more or less the same time, but at the time of their emergence the child's level of performance on related tasks is uneven (Dodwell, 1962). The cardinal-value property of the elements of a set is present before the ordinal-value property, and it is certain that there is no sudden fusion of cardinal and ordinal properties, for numbers large and small. There is a gap in time between the child understanding the relationship \( N \) and \( N \div 1 \) for small numbers, say, 5 and 9, and for larger numbers, say, 82 and 83.

Three pieces of evidence may be adduced which are relevant to the views just expressed. Apostel, Mays, Morf and Piaget (1957) showed in a study involving the division of the elements of a set into subsets, that the number of elements was at first conserved only with \( n \) small, but by 8 years of age or soon after there was conservation regardless of the number of elements. Second, the recent study of American children by Almy (1970) showed that during the emergence of concrete operational thought the level of the child's performance is irregular. Of 914 second grade pupils of average age 7 years 4 months all of whom were tested individually on a number of tasks, 366 were clearly operational on three tasks involving conservation of number and quantity, 181 were operational on four tasks of seriation, and 253 on two tasks involving ordination. Third, all experience at Leeds suggests that in the case of able pupils, once the logical instrument is available in one type of task it soon becomes available in related tasks, whereas in the case of less able pupils, horizontal differentials persist much longer. Thus as logical thought emerges, the child's level of performance is uneven at first, the unevenness disappearing more quickly the abler the pupil.
Evidence on another important issue has been provided by Van Engen and Steffe (1966) at the University of Wisconsin. They found that of 100 first grade pupils almost all could correctly work using symbols, $2 + 3 = 4 + 5 =$. Yet only 54 correctly stated no preference for separate or combined piles of candies when 5 were used, and only 45 when 9 candies were employed. It would appear that many pupils in the first grade can memorize facts employing symbols, but are unable to abstract the concept of addition from physical situations. Likewise Steffe (1966) showed that only 60 out of 341 first grade pupils conserved numerousness over 4 items in each of three tasks when the numbers did not exceed 8, and that 128 pupils responded incorrectly to at least one item of the 4 in each of three tasks. But on a test of addition facts with totals not greater than 8, pupils of the second group had an average score of 76%, and pupils of the first group an average score of 91%. These data again warn us that addition facts can be learned without a firm abstraction of number.

Continuing with our study of number I now turn to two British investigators. Brown (1969) studied the growth of pupils' understanding of the properties of the set of natural numbers. Employing both paper and pencil, and individual tests using concrete materials, he concluded that children pass through a number of stages in grasping each property and that understanding is reached at the following ages in children of average ability using small numbers: closure at 7, identity at 7-8, commutativity at 8-9, associativity at 8-9 and distributivity at 10-11 years. But pupils' performance could be advanced or retarded by up to four years compared with the norm, and he found that the operational stage in respect to all the properties tested occurred at the earliest at 9 years in his sample.

Willington (1967) gave a number of individually administered tasks to pupils of average measured intelligence of elementary school age. One of the tasks involved the understanding of the equalization of differences leading to averages, and another the understanding of the combination of odd and even numbers. Once again stages in the growth of pupils' understanding were established. A full understanding of average was available to half the 9-year-olds and all the 10-year olds; but at 10 only a little over half of the pupils could generalize that the sum of any number of even numbers is itself even, whereas the sum of a number of odd numbers varies between odd and even.

In attempting to summarize the findings in respect to number, one must be very careful, since studies are carried out with limited samples in different countries. But it may be fairly said that there is limited operational use of number to the 8th birthday. From then onwards pupils acquire a greater grasp of larger numbers with the properties of the set of natural numbers being grasped by the majority of pupils at 10-11 years of age--some years after small numbers are first conserved.

The growth of spatial concepts (Piaget and Inhelder, 1948) and geometrical concepts (Piaget, Inhelder and Szeminska, 1948) is, like the concept of number, a particular application of intellectual growth. But we are now dealing with infralogical operations which arise around the same time as logico-arithmetic operations but which are distinguished from the latter in that they involve proximity and continuity. That such concepts depend upon the growth of logical thinking is obvious if we consider, say, measurement--a notion which has wide applicability. A child must grasp that the whole is comprised of a number of parts added together, and understand the principle of iteration, that is, the repeated appli-
cation of a unit to another entity, such as length of a line or area of a surface. Spatial and geometrical concepts then are not derived from an apprehension or 'reading' of the physical properties of objects, but are actions performed on objects in thought. Hence a conceptual space emerges from 7 years of age onwards in which the child understands spatial properties and relations, quite different from the perceptual space of the very young child who can recognize perceptual distinctions. It is true that when using concrete operational thought the image may still have a place in supporting spatial reasoning, but when formal operational thought is employed, spatial reasoning is a purely abstract affair.

In Piaget's view, there arise with the beginning of the onset of concrete operational thought, simultaneously and in parallel fashion so to speak, projective structures (rectilinearity, perspectives, coordination of viewpoints), and metrical structures (measurement, systems of reference, and measurement in two and three dimensions).

A number of studies have enabled us to check on some of the Geneva findings; for example, Lovell, Healey and Rowland (1962), Shantz and Smock (1966), Lovell (1970), and above all, the outstanding study by Laurendeau and Pinard (1970) at Montreal. By and large the stages in the development of the structures, proposed by Piaget, are found but there are differences. The age range for the elaboration of a particular structure is considerable even in children of comparable background and ability as judged by teachers or by test results. Again, the situation, the actual apparatus employed and the mode of response (for example, drawing versus selection of prepared drawings) affects the level of thought elicited and hence the level of behavior observed. The form of analysis may also influence the interpretation given to the experimental findings. Finally, we must note that a pupil's response can be irregular across related tasks for a considerable period, that is, horizontal differentials are much in evidence. No doubt differences in specific experiences of particular children, explain some of the irregularities and inconsistencies. But it is now clear that the tasks are subtle, that the relevant ideas have to be carefully devised and that analysis has to be thoughtfully considered. Elicitation of thought seems to be a more tricky problem than is the elicitation of thinking about logico-arithmetic concepts. Nevertheless, it is perfectly clear that it is the emergence and growth of logical structures which underpins pupils' ability to elaborate the spatial structures in question. The excellent study of Laurendeau and Pinard mentioned above involved a number of spatial tasks given to 50 children at each age from 2 to 12 years. From an examination of the relevant correlation coefficients, and even more from a careful examination of the scalogram analyses, the authors conclude that there is indeed a consistency or coherence in the way the particular spatial concepts in question develop. The various stages in the tasks are reached in a regular rather than in a chance order.

I would also like to mention in passing a study which Lunzer (1963) carried out both in Geneva and in Manchester (England) and which yielded similar results in both cities. In essence he found that it was not until the onset of formal operational thought at around 14 years of age, that the majority of those tested were able to dissociate, completely, area and perimeter of square/rectangle, and realize that under certain changes area is conserved and not perimeter, while under other changes the reverse is true. His work also suggests that it is not until formal operational thought is emerging that a geometrical situation can be handled by going outside the given limits of a figure. This is consonant
with the views of Piaget, Inhelder and Szeminska (1948) in respect of the construction of lines outside the figure when copying a triangle.

The concept of volume in its three aspects, internal volume, volume as occupied space, and complementary or displacement volume, is not well understood even at the end of elementary school. The evidence first established by Piaget, Inhelder and Szeminska (1948), has been broadly confirmed using different techniques by Elkind (1961a, 1961b), Lovell, and Ogilvie (1961) and Uzgiris (1964). It is true that individual pupil experience may well underline situational differences and account for the observed inconsistencies across different material. But, by and large, displacement volume is not well understood until the beginning of the emergence of formal operational thought. Pupils can, of course, through the iteration of a unit volume be led to calculate the internal volume, and volume as occupied space, in 5th grade, but volume as displacement is a more difficult idea. This may not be of great consequence for pure mathematics per se but it is of great importance to science teachers.

The only other infralogical concept emerging during the elementary school years that I can stop to consider is that of time. The emergence of temporal operations is fundamental for mathematics as it occurs in, say, the notion of speed or that of rate of change. Unfortunately, the child can use time words, and tell the time, long before he can handle temporal operations. Because of this he may appear to have elaborated temporal operations when he has not.

Research carried out by Piaget (1946), and by Lovell and Slater (1960), has indicated that it is not until between 3 and 9 years of age that pupils begin to carry out the following:

(i) Put events into a sequence according to their order of succession.

(ii) Mark off intervals of time between ordered points on a time scale, and place smaller ones within larger ones.

(iii) Choose some time interval as a unit and use it as a unit for measuring some other time intervals.

In (ii) and (iii) we are, of course, carrying out the operations of subdivision and displacement or iteration as in any other form of measurement. These three operations develop more or less at the same time, but the apparatus or situation used does influence the pupil’s ability to evoke temporal operations at first.

Once again we see that the understanding of time at this level depends upon the growth of the logical instrument. Even so, the pupil’s understanding of time at this age is confined to intuitional situations. It will be well into high school before 75 per-cent of pupils realize that time on the clock is a purely arbitrary convention and correctly answer, given good reasons, the question, “When we advance the clocks by one hour in springtime, do we grow one hour older?”

So much for the elementary school stage. Let us now consider some mathematical ideas elaborated during high school years. From around 12 years of age in the brightest pupils and from 14 to 15 years in ordinary pupils, we see the emergence of formal operational thought. The chief characteristic of such thinking is the ability to invert reality and possibility.
thereby leading to the ability to use a combinatorial system and hypothetico-deductive thought. It may also be characterized as second degree operations for now the pupil can structure relations between relations as in, say, metric proportion which involves the recognition of the equivalence of two ratios.

Piaget, Inhelder and Szeminska (1948) argued that it is easier to study the growth of the scheme of proportion in geometric than in nongeometric forms, for before the child can think about similar figures he can perceive whether the figures having different dimensions are similar. Our work at Leeds has indicated that pupils' responses in respect to the construction of a rectangle similar but larger than a model, can be placed more or less into the categories which Piaget suggested but the ages at which the stages are reached have been somewhat higher. However, Inhelder and Piaget (1958) warn us that the scheme of metrical proportion in non-geometrical form depends eventually on the emergence of the growth of formal thought and that it comes later than teachers would wish. The studies of Lovell (1961), Lunzer (1965), Lovell and Butterworth (1966) with British pupils, also Steffe and Farr (1963) and Gray (1970) with American pupils, have all confirmed that apart from very able 12-year-olds, it is from the beginning of junior high school onwards—the actual age depending on the ability of the pupil—that facility is acquired in handling metric proportion. Many pupils may not be able to do this until 14 to 15 years of age and some never. This is a matter of great consequence: it has repercussions in the teaching of physics and chemistry. This inability to handle metric proportions until these ages again clearly shows the dependence of the growth of mathematical understanding on the growth of the general ways of knowing.

Allied to proportion is quantitative probability. To tackle the latter the pupil has to be able to handle, in addition to metric proportion, the permutations and combinations in which a set of elements are grouped (Piaget & Inhelder, 1951). The recent study of Shepler (1969) involving a good teaching program showed that when probability questions could be answered using multiplicative classification almost 100% of sixth grade pupils, of mean measured IQ 117, obtained correct answers. But in questions involving estimated probability using large numbers, around 1/4 to 1/2 got the answers correct although some may have done so by a rote procedure since in each problem only two numbers were given and pupils would know that the probability could not exceed 1. Such questions cannot be solved by simple multiplicative classification, and require formal thought. Some bright sixth graders should have been approaching this stage, and the study does bring out what aspects of probability are assimilable by the majority of pupils in the upper classes of elementary school. I say this in spite of the recent study by the Roumanians, Fischbein, Pampu and Manzat (1970) reported in an American journal. They took three groups of able pupils aged 5 to 6, 9 to 10, and 12 ½ to 13 ½ years of age. Subjects were asked to choose out of two sets of marbles each of 2 colors, the set which they believed offered more chances of drawing a marble of a given color. A short period of instruction was given. In the pre-school pupils judgments based on simple binary relations were prevalent, while most of the sixth grade pupils based their decisions on relations between ratios. Instruction did not produce essential changes at these levels, but instruction did bring a shift toward the type of answer given by sixth graders in the case of the 9-10 year olds. Notice, however, that the choice had to be made in the case of smallish numbers and simple ratios; e.g., 2 white 4 black and 3 white 9 black; 12 white 4 black and 20 white 10 black; 3 white 4 black and 6 white 8 black. The instruction methods used
employed grouping the marbles. The pupil with flexible concrete oper- tional thought would be able to group the above groups as 1 white 2 black and 1 white 3 black; 3 white 1 black and 2 white 1 black; 3 white 4 black and 3 white 4 black. Even so, these able 9-10 year olds obtained only 75% of possible correct answers after instruction.

Work has also been carried out on the growth of pupils’ understanding of the mathematical function. Piaget et al (1968) carried out a study of functions which were linked with the scheme of proportionality, for only those functions in which laws of variation play a part were considered. A function was regarded as the relation between the magnitude of two quantities, the variation in one bringing about a variation in the other in the same proportion. The view taken of a function was, therefore, much narrower than the one currently held in mathematics. However, using ingenious experiments they have produced evidence which suggests that the pupil only slowly acquires the ability to understand a function in this limited sense. At first it is only putting into correspondence two values, e.g., the smaller the wheel the less distance travelled; or it may appear in the form of a causal dependency, e.g., the harder the surface the higher the ball bounces. But with the onset of formal operational thought the ratios between successive pairs of ordered values of a variable can be handled.

The current mathematical definition of a function is, of course, more general than that considered by Piaget et al (1968) as I have already indicated. In school mathematics, function is used in the sense of single valued function so that the function \( b = f(a) \), defined on \( A \) as domain and with a subset of \( B \) as range, gives a mapping of the set \( A \) into the set \( B \) such that for each \( a \in A \), there is a unique image \( f(a) \in B \). There are only two studies available as far as I am aware into the growth of the concept of a function. One was carried out by Thomas (1969) at Columbia University and one by Orton (1970) at Leeds, England. Some details of the latter's work will be published by the N.C.T.M. later this year. Both used a large number of tasks individually administered to pupils. Orton's work involved pupils of average and above average mathematics attainment of high school age. Part 1 tasks tested a wide range of situations, and presented relations in all of the major representations, by diagram, by graph, by ordered pairs, by table and by equation. The formation of the appropriate range for a given rule and domain, were also considered to be important tasks, in addition to the recognition of a function. Part 2 tasks, given only to older pupils, tested their ability to handle the composition of functions, use the f-notation, and tackle harder relations throughout.

In the Part 1 tasks the stages in the growth of a concept of a function corresponded closely to those found by Thomas. Responses to Part 2 tasks also yield four stages in the growth of understanding, but these could not be aligned to the American work since the latter did not explore the understanding of the composition of functions to the same extent. Suffice it is to say that these studies have given us a far better idea of the difficulties that pupils have in the growth of their understanding of the concept of function.

An interesting study in Britain by Reynolds (1967) has thrown some light on pupils' developing grasp of proof in mathematics. Such understanding will, of course, always be important regardless of the nature of the curriculum. He investigated among abler pupils of high school age the
understanding which they have in respect to assumptions, generalizations, and proof by converse, reductio ad absurdum, and deduction. His conclusions were that Piaget's formulations regarding stages of thinking accounts for a good deal of the variability in the nature of pupil responses, for replies indicative of concrete operational thought appeared regularly, while responses indicative of formal operational thought increased with age. But his work also showed that there were discrepancies between the replies obtained, and what might be expected from Piaget's views on the nature of formal thought. For example, Reynolds' work gives evidence that the degree of structure of a problem is important in this respect. In a well structured problem such as the Geneva school used, the assumptions, variables and universes of discourse are easily identified and the pupil has no need to introduce assumptions and hypotheses from outside. In the former instance solutions offered are closer to those expected from Piaget's cognitive developmental model.

Finally, in this section I would like to mention the concepts of point and limit, two notions of fundamental importance in mathematics. Piaget, Inhelder and Szeminska (1948) have indicated that children proceed through a number of stages in their growth of understanding of 'point', and that it is not until the onset of formal operational thought that point becomes thought of as homogeneous regardless of the original shape from which it was derived and, of course, without shape or surface area. Again Taback (1969) studied aspects of the concept of a limit among American children aged 3, 10, and 12 years using a number of individually administered tasks. He tells us that his subjects were drawn from independent schools, were very sophisticated in expressing themselves and came from homes in which education was respected and books were available. It would be reasonable to assume that such pupils with a chronological age of 12 years would have a mental age of around 14 years on the average. According to Taback only about one-third of the 12-year-olds could appreciate an infinite number of points within a neighborhood; that in respect of convergence, those questions which demanded a level of thought liberating the pupil from physical materials could be answered only by the 12-year-olds; and that taking the study as a whole, with few exceptions, only the 12-year-olds could conceptualize an infinite process.

This review has, I hope, given sufficient evidence that it is the development of the general ways of knowing which determines the manner in which taught material is understood.

Before I conclude this part of my paper I must say something of educational technology. I suggest we have been too much concerned with hardware and too little about teachware; i.e., too little about the preparation and presentation of material in such a way that the pupil will be motivated and helped to act on, transform and construct. In that branch of educational technology with which I have been most closely associated, computer based learning, this certainly has been the case although I believe that we are now seeing the light. Not that hardware is unimportant; indeed it is vital that the teacher be given the freedom to teach his subject matter in the way he considers best.

As I have suggested, it is the growth of the general ways of knowing that determines the manner in which new knowledge in mathematics is assimilated. I suggest that the purpose of computer based learning systems is for each individual, to provide opportunities for the evocation, organization and strengthening of the available strategies of thinking.
in mathematics, or other subject areas, thereby permitting an increase in level of attainment. In this I believe such systems may have great possibilities. At the same time, however, I believe we must have an open mind about the extent to which they will aid the growth of the general ways of knowing.

Again on the research side any models which are derived from hypothesizing or from pupil response, must for the foreseeable future, be looked upon from the point of view of enhancing adaptive material and from the point of view of controlling practice. Such models will cover very limited areas of work and should be looked upon merely as models for optimizing instruction within that area. My colleagues at Leeds, Pat Woods and J. R. Hartley (1971) are publishing this month details of such a model in respect to the addition of the natural numbers. Using criteria of probability of success and rate of working for each column of the task in vertical format, analysis of variance revealed main effects of digit size and number of rows. Following a formal development of the model, a least squares analysis derived a function which, for the experimental data, related these variables to the criteria. These are used by the computer to generate examples so that a pupil works at any specified level of success. But it must be looked upon only as a model for optimizing instruction over a very limited area, and in no sense be looked upon as a model reflecting intellectual development in the area.

A final point to note is that on such evidence as we have, teachers are slow to leave the computer system to do the things it is best suited to do and to turn themselves to more creative teaching.

3. IMPLICATIONS FOR TEACHING

I now turn to the third part of my paper. So far I have talked a great deal about research. What does it all suggest to the mathematics educator? Let me reply in this way. In 1961 an American named Mayer published a book with the title The Schools (1961). It was published here by Harper and in London by Bodley Head Press. While I cannot agree with all he wrote in the book there was one sentence which I found arresting. It was this: "What future teachers need, and cannot now find, is the course which attempts to explore the profound aspects of the deceptively simple material they are going to teach, which analyzes case by case the types of difficulty that children find in approaching such material, which suggests tools and techniques and methods of presentation that may help children overcome the difficulties."

How stands the position today--10 years later--in respect to mathematics teaching? I suggest that now we know--thanks to the Piaget-type research--much more about the profound aspects of the deceptively simple material in mathematics that children are called upon to learn. Again, if we take the trouble we can analyze in far greater detail the difficulties that children have in approaching such material. We also know that the development of the general ways of knowing will determine the manner in which the mathematical ideas are assimilated. Of course we have only just made a beginning in these matters, and far more knowledge is required. But I suggest that we can begin to write down the educational implications of all that I have said--at least the implications which relate to classroom organization and general teaching techniques--ever bearing in mind that understood knowledge results from action and the transforming of one reality state into another.
However, before I discuss such organization and techniques I would like to mention two other important findings of the last decade or so in respect to scholastic educability. First, there is now abundant evidence that the value judgments of parents, the function of language within the home and parental attitudes to education, affects scholastic educability. The evidence is so widespread that it does not need referencing. The mathematics teacher can certainly attempt to change parental attitudes but it's a hard job. Influencing what happens in the home, especially in pre-school years, is beyond the limits of the individual teacher and needs a national policy in order to attempt changes. Second, there is increasing evidence, e.g., Pidgeon (1970), that teacher expectation affects pupil performance. The strong suggestion for the mathematics teacher is that he should think well of his pupils and set standards of work which are high for them—that is, high for a particular individual or small group.

Realizing then that home attitudes and teacher expectation greatly affect pupils' desire to act on and transform reality thereby yielding new knowledge, we pass to consider classroom organization and general teaching approaches that appear to aid these transformations based on a Piagetian cognitive-developmental model. In respect to the elementary school we may suggest:

1. A move from a formal classroom atmosphere with much talk by the teacher directed to the whole class, to the position where the pupils work in small groups or individually, at tasks which have been provided.

2. The opportunity for pupils to act on physical materials, and to use games in the manner suggested by Dienes. It is the abstractions from actions performed on objects and not the objects themselves that aid forward knowledge of mathematical ideas. Not until flexible formal operational thought is available in mathematics can the latter be learned using words and symbols only, and intuitive data dispensed with.

3. In the Genevan view social intercourse using verbal language is an important influence in the development of concrete operational thought. Through exchanges, discussions, agreements, oppositions, both between children, and between adults and children, the child encounters viewpoints which must be reconciled with those of his own. There is now exchange and interpersonal as well as intrapersonal coordination. These cooperative aspects of exchange are important, for the pupil is forced to organize his thoughts into a coherent structure and also forced to elicit the strategies of thinking available to him. This argues a strong case for much teacher/child and child/child interaction in mathematics teaching. And since language helps the child organize his experience and carry his thinking, the case is made for dialogue and action to go alongside one another. Likewise, P. I. Galperin (1957), the well known Russian educator/psychologist, and a person very different from Piaget, argues for the use of numerical language and actions as opposed to language and things only.
4. Since mathematics is a structured and interlocked system of relations expressed in symbols and governed by firm rules, the initiative, and the direction of the work must be the teacher's responsibility. This was often overlooked in the progressive education movement. This does not mean that pupils should never have a choice of activities, and it does not imply that teachers should ignore naturally occurring but relevant situations. Indeed, it was 75 distinguished American mathematicians who reminded us in the American Mathematical Monthly (Mayer, 1962) that children wished to use mathematics as a tool with which to explore the world and not to play a game with arbitrary rules. In other words, our tasks should have the proper degree of structure and be seen by children to have relevance to real life.

5. Alongside the abstraction of the mathematical idea from the physical situation, there must be the introduction of the relevant symbolization and the working of examples, involving drill and practice and problems, on paper.

When we consider pupils over 12 years of age the position becomes more complex. With very backward children, the class organization and active approach used earlier have to be continued, although new mathematical ideas have to be introduced. The nature of the activities and materials will also change to become appropriate to the pupils' emotional and physical development. But the structures that these pupils will elaborate will be those derivable from interaction with first hand reality (Lovell, 1966). Alas, in the case of the weakest school educable pupils, most learning in mathematics will take place using algorithms which we must give them to cope with real life situations.

In the case of ordinary pupils we can slowly move them to new topics which depend upon the emergence of formal thought. But there rests upon us the absolute necessity of introducing these structures through concrete realizations. Let me give you an example of what I mean. Take an envelope. Let its center be 0 and suppose OK and OY are axes in the plane of the envelope, through O, and parallel to two adjacent sides of the envelope. Suppose OZ is the axis through 0, perpendicular to the plane of the envelope. The 12 or 13-year-old can build up a table showing the effects of rotating the envelope about the three axes—any one operation being followed by a second. He can understand the structure displayed by the table; hence he can understand the structure of a mathematical group in this one concrete realization. The same pupil can equally well handle the addition of the set of integers mod 4 when the operation is "add two numbers of the set," draw up an appropriate table and understand its structure, namely that of another group. But the recognition of the relationships between structures or the generalization of the structure requires flexible formal operational thought so that the group structure can now be conceived generally without a concrete realization (Dienes and Jeeves, 1970). Each successive concrete realization of an abstract structure is likely to increase the pupil's awareness of the structure and when formal thought is available increase the chances of the structure being generalized. Dienes (1963) also lays stress on this general point in another of his books: Experimental Study of Mathematics Learning.
4. STUDY OF MATHEMATICS LEARNING

I further suggest that in the case of ordinary pupils, our knowledge of their slow move to formal thought is such that small group and individual work is still necessary, permitting opportunities for individual or small group assignments, and dialogue between teacher/pupil and pupil/pupil.

The abler the pupil, the more quickly flexible formal operational thought will be in evidence, and the more quickly can he move away from intuitable data and consider third level abstractions devoid of concrete realizations. Not only are these structures more easily attainable in these circumstances but so are the relationships between structures and hence generalization. More class teaching with verbal and symbol exposition is now possible although the need for constant discussion between teacher and pupil and between pupil and pupil remains. And pupils still need the opportunity to formulate their own questions and discuss their own answers to them. However, if there are considerable differences in attainment even in an overall high attaining class, the small group approach remains necessary. As Gagne reminded us, problem solving demands masses of structurally organized knowledge. Such knowledge is not possessed equally when pupils differ in attainment, whatever their potential may be.

Having talked about the classroom and techniques in general terms I now wish to make four further points.

1. There is the question of discovery methods. I concur with Ausubel (1964) that both verbal learning and problem solving through active methods can be rote or meaningful. As he points out, active methods are not meaningful unless they rest on a base of understood concepts and that the operations involved are also meaningful. Obversely, if the pupil can relate new material, given by verbal exposition, in a substantive and nonarbitrary way to what has gone before, the learning will be meaningful for the child will have transformed one reality state into another.

2. The ways in which we now look at mathematical ideas demand a greater degree of verbal explication than was the case in the more traditional mathematics program. Teachers and pupils now require greater powers of verbal comprehension and explication in mathematics than formerly.

3. British teachers often find it hard to change from class teaching to small group work. Materials have to be prepared, groups organized, and language and action must proceed together. It is very hard work. No doubt U.S. teachers find the same difficulties, but they must be encouraged to make the move. Again, British primary school teachers are not equally good teachers across the whole range of subjects they have to tackle; they cannot handle the content and teaching methodology with equal facility across the board. They have indeed a difficult task as have U.S. elementary school teachers. I suggest, therefore, that while we encourage American elementary school teachers as much as possible in respect to mathematics teaching, and raise the overall standard so to speak, we must not expect that all will show equal competence in this area. Some will be relatively better
teachers of other subjects, and this must be accepted.

4. I would like to raise the question of whether we should wait until pupils are ready to assimilate ideas fully and formally so to speak, as mathematicians would have them assimilated ideally. In my view the answer is 'No.' Indeed, it is impossible to say with precision at present when pupils are ready. In the first place the emergence of new forms of thought are patchy and irregular at first. Some ideas of comparable structure and level of abstraction are understood better than others, and there are individual differences as well. Second, the more familiar pupils are with content, the more readily, within limits, can formal operational thought be elicited. Third, topics can be introduced in different ways and at different depths, so to speak, so that the teacher may well start with the assumption that pupils' understanding of the idea will be limited at that point in time. But the teacher can lay a framework, can make pupils feel more 'at home' with the idea, and when the teacher returns to the topic later and with a different treatment, the subject matter will be reorganized and seen in a different light, for at the later date the general ways of knowing will also have advanced. It is knowledge of the subject matter and of his pupils, which allows the teacher to distinguish the level of thinking of the child in relation to this particular topic, and he will not attempt to force an understanding not yet available to the pupil.

5. CONCLUSION

In this paper I have ranged far and wide. I hope I have conveyed a note of cautious optimism in respect of mathematics education. I do not believe that by some miracle we can accelerate the growth of pupils' thinking so that what was done in college can now be done in first grade. Knowledge of Piaget-type research does not make mathematical ideas per se any easier for children to learn. But I do believe that if we will but take the trouble, and accept certain limitations for some pupils, we can bring more understanding and greater enjoyment to pupils in respect to mathematics. I look to a slow and steady improvement over the years—not to a revolution. The Coleman report in this country—s ait to be the second largest social science research project ever mounted—clearly showed that the differences in educational achievement between groups is there at the beginning of school. These differences will not be removed overnight.

When the Chairman of the Convention Program Committee sent the details of this conference, his letter struck a somewhat pessimistic note. You will remember it ran: "These are troubled times for educational research and its practitioners. Federal support of and confidence in educational research is equivocal and diminishing. State and local educational problems are already of gigantic proportions and enlarge daily." In response to this I have tried to show that with respect to mathematics, research has given us much more knowledge of the profound aspects of the deceptively simple material pupils have to learn, more knowledge about the difficulties children have and the stages through which they pass in coming to grips with mathematical ideas, and indications concerning the form that classroom organization and teaching strategies should take. Moreover
still greater knowledge could help us in curriculum development itself. If we will but take the trouble to make serving teachers rethink their position, and reshape the education of our teachers-to-be, there are grounds for some mild optimism. Not that mathematical ideas themselves can be made easier, but we may be able to produce an atmosphere in which pupils are more likely to assimilate and enjoy the ideas in question.
In this paper we discuss some theoretical issues connected with the phenomenon of activity cycles frequently found in unstructured interviews with 3-4 year old children. Although the approach we take is Piagetian, we believe that the problems raised are important in any structural theory of knowing.

1. EXPERIMENTAL BACKGROUND

We first describe very briefly the necessary experimental background. The interviews in question are built around an apparatus, the balance beam (1). The balance used by us is 20 cm. high, 75 cm. long, has four evenly spaced hooks on each side and uses large washers as unit weights. Experimenter and subject both sit at the same side of the table facing the apparatus and discuss how it works, make predictions, manipulate it; the child is asked to make it balance, to make that side go down, to tell what happens when the beam is released, etc. Every effort is made to ensure a smooth and natural flow of interaction between experimenter and child. The experimenter avoids setting up problem situations; rather these should arise naturally during the interaction. The child is encouraged to take the initiative and express himself in any way he can; he may, within limits, manipulate the apparatus, play his own games, etc. The whole interview is videotaped, transcribed in minute detail, and then analyzed in various ways (cf. Witz, 1971b, and "Witz, 1971c).

2. CYCLES OF ACTIVITY

A common phenomenon under the free interview conditions just described is the occurrence of what one might call "cycles of activity." The general form of such cycles is shown in Figure 1: The cycle begins with an action, or a statement suggesting an action, A (e.g., the subject puts on a washer, or says, "Let's put one on here"); there is some

---

1 Based in part on a talk delivered at the Second Annual Interdisciplinary Meeting in Structural Learning, April 2-3, 1971, University of Pennsylvania, Philadelphia, Pa., see Witz (1971c). Research supported by a grant from the Basic Studies Branch, U.S. Office of Education. I would like to thank Jack Easley, co-principal investigator on this project, for his encouragement and advice.
expected (anticipated, intended, desired) effect, B (e.g., one side of the beam goes down); and the cycle terminates when the expected effect occurs. It is then immediately followed by another, similar cycle (with a similar action A, and a similar expected effect B). During the cycle the child may talk about why he is doing A, or what he expects to see happen or what he wants to accomplish (e.g., "It will crash"), and he may go through various auxiliary actions A' (e.g., remove the experimenter's hand when he is holding the beam, or say, "Take your hand off," etc.). Cycles usually occur in groups of two to six; a typical example is shown in Figure 2. First the child successively puts seven washers on the hook on the extreme left; while doing so, he talks about the fact that the beam will go down, and when the beam is released he watches intently. This marks the end of the first cycle, and he immediately initiates the second: While picking up some washers he says, "You can put more on," and proceeds to hang two washers on each of the three other hooks on the left, moving from left to right. The remaining four cycles are taken up with removing the washers from the hooks in the reverse order. In general, the important points are the following.

1. All cycles are initiated by the subject—they are not responses by the subject to instructions or questions of the experimenter.

2. Within a cycle, there must be definite evidence that the intended or expected event, B is in fact intended or expected. In other words, cycles must be distinguished from trial and error experimentation and play.

(1) and (2) together justify the view that cycles reveal existing cognitive structures. Further:

3. Any two consecutive cycles show both common features as well as variations in their A and B parts.

Both significant common features and variations will be called category constraints. To illustrate, in Figure 2 a common constraint across all six cycles is that the subject always hangs two or more washers on a
TIMMY 4/7
Complete Cycles,
First Group
Activity Structure I

3. ACTIVITY STRUCTURES

To account for a block of consecutive cycles like the one in Figure 2, we conceptualize, following Piaget, an organization of schemes which generates the cycles, and which is responsible for both the repetition as well as the variation in the cycles. Such a system of schemes will be called an activity structure. More precisely, we postulate a specific mode of organization of action schemes a and sensory experience schemes b into irreducible units a < b which will be called activity elements and are assumed to underlie the individual cycles; an activity structure is then a system of related activity elements which is assumed to underlie the whole group of cycles. (It is understood that the same activity structure may underlie different groups of cycles.) This general approach is strictly analogous to Piaget's procedure. Activity elements, or rather <, constitutes an irreducible mode of internal organization in the same way in which, for example, circular reaction is an irreducible mode of internal organization. We cannot say that in a cycle like Figure 1, the child does A "in order to achieve the goal" B, or that he does A because "he desires or wants to bring about" B. Within an a < b element, the a and b schemes are not really separable; there is thorough motor-integration (e.g., between releasing the beam and catching it when it comes down), and younger children are often unable to verbalize anything in the situation even if pressed; they simply act. There is little evidence of "reflection" or "thinking ahead to B," much less of recalling a "rule" or "general principle."

Further, in an activity element a < b, the a and b schemes do not necessarily correspond to specific actions (A) and events (B), but rather to broad ranges of actions and events. Different activity elements thus
may represent internal structure at what might be described from the outside as different "levels of generality." These levels are determined by the nature of the transition between the cycles. A direct transition from one cycle to another, which is subject to essentially the same constraints as the first, and in which no new aspects are brought in by the child, indicates that both cycles are underlain by the same activity element. But a direct transition from one cycle to a second, in which the child introduces new aspects or violates or varies some of the constraints from the first cycle, but in which at the same time several significant constraints from the first cycle are preserved, indicates the existence of a more general activity element $a_1 \rightarrow b_1$ which underlies at least the second and perhaps both cycles. Applying these criteria to the cycles in Figure 2, we obtain the structure in Figure 3. There are three levels of $a$-scheme organization:

Timmy 4/7

Activity Structure I

![Figure 3](image)

\[ a -- \text{putting substantial weight on one hook on one side}, \]
\[ a_1 -- \text{getting a lot of weight on the hooks of one side}, \]
\[ a_2 -- \text{taking the weights on one side off}, \]
\[ a_11 -- \text{operating on one side by putting on and taking off weights}, \]

and there are two identifiable levels of $b$-scheme organization:

\[ b -- \text{seeing the balance beam go down and come to rest}, \]
\[ b_1 -- \text{seeing the beam go down or return to normal}; \]

the partial order among the activity elements is justified by the indicated transitions.

Some indications of how these criteria are operationalized are given in Witz, 1971c
The procedure illustrated in this example -- of keeping the same mode of organization but acknowledging broader and broader coherent ranges of activity and expected events, corresponding to different levels of generality of internal organization--reflects of course our feeling that the same phenomenon is involved at various levels of generality and that it would be artificial to single out a single level (e.g., "unit movements" and single phrases). More decisively, we have concrete evidence that initially completely separate activity elements become subsumed under elements at higher levels of generality, and that thus two initially separate structures become combined into a larger structure, in precisely the way provided for by Figure 3. The subject in Figures 2 and 3 (Ti, age 4-7), for example, starts out with the activity structure shown in Figure 3 (structure I), and a second structure concerned with putting on washers in symmetrical configurations (structure II). Throughout the first half of the interview, these two structures are entirely separate. But after certain manipulations by the experimenter, Ti suddenly begins a series of three cycles, alternating from I to II and then back to I (see Witz, 1971c for more details). By our rules, there must be a new activity element, the top element in Figure 4, under which all elements of I and II are subsumed. The two original structures, I and II, therefore, have become incorporated into a single new, more comprehensive structure, which represents a new concept of what the balance really is and which marks a significant advance toward the adult concept. The essential point is that internal "level of generality" organization has been tied to dynamics in the child's own activity (transitions), not to a priori or objective relationships between classes of observable behaviors.

A structure of activity elements like Figure 3 or 4 can be formally modeled by a system of functions partially ordered in various ways (e.g.,
Scandura, 1968) or by partial feature specification, in feature matrices, or in other ways. But the essence of our data and of the terms of our analysis is that concepts are built up by dynamic means, and not by abstract operations on mappings or features. If we call the partial order between the activity elements of an activity structure "subsumption", then any explanation of subsumption by abstract operations on functions or features divorced from the real dynamics of the flow of activity commits one to a non-dynamic, non-active, and, worse, extensionalist view of concept development, when only an intrinsically dynamic or active view will do justice to the observations. But while the system of terms of our analysis is just large enough to spell out this point, it fails to provide an urgently needed intrinsic model for the activity elements themselves. Having abandoned simple dimensions, rules, mappings and features, we do not understand, for example, what it is that allows an activity element to crystallize at a certain level of generality, or why it is that two activity structures can be or are integrated and subsumed under a larger structure in some situations while two others cannot or are not. And there is the problem of openness in the expected effects: If the expected effect B fails to materialize, the child has no problem recognizing the substitute event B' as a substitute of B. And so on.

There is no doubt that activity structures represent a major way of knowing in small children. A basic problem for us is to explain the rise of a stable objective world in the child; the suggestion made in Figure 4, then, is that the rise of this objective world is accomplished by integration of motor activity into more and more encompassing structures.
SIMULATION MODELS
When the magician pulls the rabbit from the hat, the spectator can respond either with mystification or with curiosity. He can enjoy the surprise and the wonder of the unexplained (and perhaps inexplicable), or he can search for an explanation.

Suppose curiosity is his main response—that he adopts a scientist's attitude toward the mystery. What questions should a scientific theory of magic answer? First, it should predict the performance of a magician handling specified tasks—producing a rabbit from a hat, say. It should explain how the production takes place, what processes are used, and what mechanisms perform those processes. It should predict the incidental phenomena that accompany the magic—the magician's patter and his pretty assistant—and the relation of these to the mystification process. It should show how changes in the attendant conditions—both changes "inside" the members of the audience and changes in the feat of magic—alter the magician's behavior. It should explain how specific and general magician's skills are learned, and what the magician "has" when he has learned them.

1. THEORY OF PROBLEM SOLVING--1958

Now I have been quoting—with a few word substitutions—from a paper published in the Psychological Review in 1958 (Newell, Shaw, & Simon, 1958). In that paper, titled "Elements of a Theory of Human Problem Solving," our research group reported on the results of its first two years of activity in programming a digital computer to perform problem-solving tasks that are difficult for humans. Problem solving was regarded by many, at that time, as a mystical, almost magical, human activity—as though the preservation of human dignity depended on man's remaining inscrutable to himself, on the magic-making processes remaining unexplained.

In the course of writing the "Elements" paper, we searched the literature of problem solving for a statement of what it would mean to explain human problem solving, of how we would recognize an explanation if we found one.

Failing to discover a statement that satisfied us, we manufactured one of our own—essentially the paragraph I paraphrased earlier. Let me quote it again, with the proper words restored, so that it will refer to the magic of human thinking and problem solving, instead of stage magic.

What questions should a theory of problem solving answer? First, it should predict the performance of a problem solver handling specified tasks. It should explain how human problem solving takes place: what processes are used, and what mechanisms perform these processes. It should predict the incidental phenomena that accompany problem solving, and the relation of these to the problem-solving process.... It should show how changes in the attendant conditions—both changes "inside" the problem solver and changes in the task confronting him—alter problem-solving behavior. It should explain how specific and general problem-solving skills are learned, and what it is that the problem solver "has" when he has learned them [p.151].

1.1 A Strategy

This view of explanation places its central emphasis on process—on how particular human behaviors come about, on the mechanisms that enable them. We can sketch out the strategy of a research program for achieving such an explanation, a strategy that the actual events have been following pretty closely, at least through the first eight steps:

1. Discover and define a set of processes that would enable a system capable of storing and manipulating patterns to perform complex nonnumerical tasks, like those a human performs when he is thinking.

2. Construct an information-processing language, and a system for interpreting that language in terms of elementary operations, that will enable programs to be written in terms of the information processes that have been defined, and will permit those programs to be run on a computer.

3. Discover and define a program, written in the language of information processes, that is capable of solving some class of problems that humans find difficult. Use whatever evidence is available to incorporate in the program processes that resemble those used by humans. (Do not admit processes, like very rapid arithmetic, that humans are known to be incapable of.)

4. If the first three steps are successful, obtain data, as detailed as possible, on human behavior in solving the same problems as those tackled by the program. Search for the similarities and differences between the behavior of program and human subject. Modify the program to achieve a better approximation to the human behavior.

5. Investigate a continually broadening range of human problem-solving and thinking tasks, repeating the first four steps for each of them. Use the same set of elementary information processes in all of the simulation programs, and try to borrow from the subroutines and program organization of previous programs in designing each new one.
6. After human behavior in several tasks has been approximated to a reasonable degree, construct more general simulation programs that can attack a whole range of tasks—winnow out the "general intelligence" components of the performances, and use them to build this more general program.

7. Examine the components of the simulation programs for their relation to the more elementary human performances that are commonly studied in the psychological laboratory: rote learning, elementary concept attainment, immediate recall, and so on. Draw inferences from simulations to elementary performances, and vice versa, so as to use standard experimental data to test and improve the problem-solving theories.

8. Search for new tasks (e.g., perceptual and language tasks) that might provide additional arenas for testing the theories and drawing out their implications.

9. Begin to search for the neurophysiological counterparts of the elementary information processes that are postulated in the theories. Use neurophysiological evidence to improve the problem-solving theories, and inferences from the problem-solving theories as clues for the neurophysiological investigations.

10. Draw implications from the theories for the improvement of human-performance—for example, the improvement of learning and decision making. Develop and test programs of application.

11. Review progress to date, and lay out a strategy for the next period ahead.

Of course, life's programs are not as linear as this strategy, in the simplified form in which we have presented it. A good strategy would have to contain many checkpoints for evaluation of progress, many feedback loops, many branches, many iterations. Step 1 of the strategy, for example, was a major concern for our research group (and other investigators as well) in 1955-56, but new ideas, refinements, and improvements have continued to appear up to the present time. Step 7 represented a minor part of our activity as early as 1956, became much more important in 1955-61, and has remained active since.

Nor do strategies spring full-grown from the brow of Zeus. Fifteen years' hindsight makes it easy to write down the strategy in neat form. If anyone had attempted to describe it prospectively in 1955, his version would have been much cruder and probably would lack some of the last six steps.

1.2 The Logic Theorist

The "Elements" paper of 1953 reported a successful initial pass through the first three steps in the strategy. A set of basic information processes for manipulating nonnumerical symbols and symbol structures had been devised (Newell & Simon, 1956). A class of information-processing or list-processing languages had been designed and implemented, incorporating the basic information processes, permitting programs to be written in terms of them, and enabling these programs to be run on computers (Newell & Shaw, 1957). A program, The Logic Theorist (LT), had been written in
one of these languages, and had been shown, by running it on a computer, to be capable of solving problems that are difficult for humans (Newell, Shaw, & Simon, 1957).

LT was, first and foremost, a demonstration of sufficiency. The program's ability to discover proofs for theorems in logic showed that, with no more capabilities than it possessed--capabilities for reading, writing, storing, erasing, and comparing patterns--a system could perform tasks that, in humans, require thinking. To anyone with a taste for parsimony, it suggested (but, of course, did not prove) that only these capabilities, and no others, should be postulated to account for the magic of human thinking. Thus, the "Elements" paper proposed that "an explanation of an observed behavior of the organism is provided by a program of primitive information processes that generates this behavior [p. 151]," and exhibited LT as an example of such an explanation.

The sufficiency proof, the demonstration of problem-solving capability at the human level, is only a first step toward constructing an information-processing theory of human thinking. It only tells us that in certain stimulus situations the correct (that is to say, the human) gross behavior can be produced. But this kind of blind S-R relation between program and behavior does not explain the process that brings it about. We do not say that we understand the magic because we can predict that a rabbit will emerge from the hat when the magician reaches into it. We want to know how it was done--how the rabbit got there. Programs like LT are explanations of human problem-solving behavior only to the extent that the processes they use to discover solutions are the same as the human processes.

LT's claim to explain process as well as result rested on slender evidence, which was summed up in the "Elements" paper as follows:

First, ... (LT) is in fact capable of finding proofs for theorems--hence incorporates a system of processes that is sufficient for a problem-solving mechanism. Second, its ability to solve a particular problem depends on the sequence in which problems are presented to it in much the same way that a human subject's behavior depends on this sequence. Third, its behavior exhibits both preparatory and directional set. Fourth, it exhibits insight both in the sense of vicarious trial and error leading to "sudden" problem solution, and in the sense of employing heuristics to keep the total amount of trial and error within reasonable bounds. Fifth, it employs simple concepts to classify the expressions with which it deals. Sixth, its program exhibits a complex organized hierarchy of problems and subproblems [p. 162].

There were important differences between LT's processes and those used by human subjects to solve similar problems. Nevertheless, in one fundamental respect that has guided all the simulations that have followed LT, the program did indeed capture the central process in human problem solving: LT used heuristic methods to carry out highly selective searches, hence to cut down enormous problem spaces to sizes that a slow, serial processor could handle. Selectivity of search, not speed, was taken as the key organizing principle, and essentially no use was made of the computer's ultrarapid arithmetic capabilities in the simulation program. Heuristic methods that make this selectivity possible have turned out to be the central magic in all human problem solving that has been studied to date.
Thus, in the domain of symbolic logic in which LT worked, obtaining by brute force the proofs it discovered by selective search would have meant examining enormous numbers of possibilities—$10^{\text{hundreds or thousands}}$. LT typically searched trees of 50 or so branches in constructing the more difficult proofs that it found.

1.3 Mentalism and Magic

LT demonstrated that selective search employing heuristics permitted a slow serial information-processing system to solve problems that are difficult for humans. The demonstration defined the terms of the next stages of inquiry: to discover the heuristic processes actually used by humans to solve such problems, and to verify the discovery empirically.

We will not discuss here the methodological issues raised by the discovery and certification tasks, apart from one preliminary comment. An explanation of the processes involved in human thinking requires reference to things going on inside the head. American behaviorism has been properly skeptical of "mentalism"—of attempts to explain thinking by vague references to vague entities and processes hidden beyond reach of observation within the skull. Magic is explained only if the terms of explanation are less mysterious than the feats of magic themselves. It is no explanation of the rabbit's appearing from the hat to say that it is "materialized."

Information-processing explanations refer frequently to processes that go on inside the head—if you like—and to specific properties of human memory: its speed and capacity, its organization. These references are not intended to be in the least vague. What distinguishes the information-processing theories of thinking and problem solving described here from earlier discussion of mind is that terms like "memory" and "symbol structure" are now pinned down and defined in sufficient detail to embody their referents in precisely stated programs and data structures.

An internal representation, or "mental image," of a chess board, for example, is not a metaphorical picture of the external object, but a symbol structure with definite properties on which well-defined processes can operate to retrieve specified kinds of information (Baylor & Simon, 1966; Simon & Barenfeld, 1969).

The programmability of the theories is the guarantor of their operation-ality, an iron-clad insurance against admitting magical entities into the head. A computer program containing magical instructions does not run, but it is asserted of these information-processing theories of thinking that they can be programmed and will run. They may be empirically correct theories about the nature of human thought processes or empirically invalid theories; they are not magical theories.

Unfortunately, the guarantee provided by programmability creates a communication problem. Information-processing languages are a barrier to the communication of the theories as formidable as the barrier of mathematics in the physical sciences. The theories become fully accessible only to those who, by mastering the languages, climb over the barrier. Any attempt to communicate in natural language must perforce be inexact.

There is the further danger that, in talking about these theories in ordinary language, the listener may be seduced into attaching to terms their traditional meanings. If the theory speaks of "search," he may
Simon and Newell

posit a little homunculus inside the head to do the searching; if it speaks of "heuristics" or "rules of thumb," he may introduce the same homunculus to remember and apply them. Then, of course, he will be interpreting the theory magically, and will object that it is no theory.

The only solution to this problem is the hard solution. Psychology is now taking the road taken earlier by other sciences: it is introducing essential formalisms to describe and explain its phenomena. Natural language formulations of the phenomena of human thinking did not yield explanations of what was going on; formulations in information-processing languages appear to be yielding such explanations. And the pain and cost of acquiring the new tools must be far less than the pain and cost of trying to master difficult problems with inadequate tools.

Our account today will be framed in ordinary language. But we must warn you that it is a translation from information-processing languages which, like most translations, has probably lost a good deal of the subtlety of the original. In particular, we warn you against attaching magical meanings to terms that refer to entirely concrete and operational phenomena taking place in fully defined and operative information-processing systems. The account will also be Pittsburgh-centric. It will refer mainly to work of the Carnegie-RAND group, although information-processing psychology enlists an ever-growing band of research psychologists, many of whom are important contributors of evidence to the theory presented here.

2. THEORY OF PROBLEM SOLVING—1970

The dozen years since the publication of the "Elements" paper has seen a steady growth of activity in information-processing psychology—both in the area of problem solving and in such areas as learning, concept formation, and language behavior. Firm contact has been made with more traditional approaches, and information-processing psychology has joined (or been joined by) the mainstream of scientific inquiry in experimental psychology today. Instead of tracing history here, we should like to give a brief account of the product of the history, of the theory of human problem solving that has emerged from the research.

The theory makes reference to an information-processing system, the problem solver, confronted by a task. The task is defined objectively (or from the viewpoint of an experimenter, if you prefer) in terms of a task environment. It is defined by the problem solver, for purposes of attacking it, in terms of a problem space. The shape of the theory can be captured by four propositions (Newell & Simon, 1971, Ch. 14):

1. A few, and only a few, gross characteristics of the human information-processing system are invariant over task and problem solver.

2. These characteristics are sufficient to determine that a task environment is represented (in the information-processing system) as a problem space, and that problem solving

...
Human Problem Solving

takes place in a problem space.

3. The structure of the task environment determines the possible structures of the problem space.

4. The structure of the problem space determines the possible programs that can be used for problem solving.

These are the bones of the theory. In the next pages, we will undertake to clothe them in some flesh.

2.1 Characteristics of the Information-Processing System

When human beings are observed working on well-structured problems that are difficult but not unsolvable for them, their behaviors reveal certain broad characteristics of the underlying neurophysiological system that supports the problem-solving processes; but at the same time, the behaviors conceal almost all of the detail of that system.

The basic characteristics of the human information-processing system that shape its problem-solving efforts are easily stated: The system operates essentially serially, one-process-at-a-time, not in parallel fashion. Its elementary processes take tens or hundreds of milliseconds. The inputs and outputs of these processes are held in a small short-term memory with a capacity of only a few symbols. The system has access to an essentially infinite long-term memory, but the time required to store a symbol in that memory is of the order of seconds or tens of seconds.

These properties—serial processing, small short-term memory, infinite long-term memory with fast retrieval but slow storage—impose strong constraints on the ways in which the system can seek solutions to problems in larger problem spaces. A system not sharing these properties—a parallel system, say, or one capable of storing symbols in long-term memory in milliseconds instead of seconds—might seek problem solutions in quite different ways from the system we are considering.

The evidence that the human system has the properties we have listed comes partly from problem-solving behavior itself. No problem-solving behavior has been observed in the laboratory that seems interpretable in terms of simultaneous rapid search of disjoint parts of the solver's problem space. On the contrary, the solver always appears to search sequentially, adding small successive accretions to his store of information about the problem and its solution.³

³Claims that human distractability and perceptual capability imply extensive parallel processing have been refuted by describing or designing serial information-processing systems that are distractable and possess such perceptual capabilities. (We are not speaking of the initial "sensory" stages of visual or auditory encoding, which certainly involve parallel processing, but of the subsequent stages, usually called perceptual.) For further discussion of this issue, see Simon (1967) and Simon and Barenfeld (1969). Without elaborating here, we also assert that incremental growth of knowledge in the problem space is not incompatible with experiences of sudden "insight." For further discussion of this point, see Newell, Shaw, and Simon (1962) and Simon (1966).
Additional evidence for the basic properties of the system as well as data for estimating the system parameters come from simpler laboratory tasks. The evidence for the 5 or 10 seconds required to store a symbol in long-term memory comes mainly from rote memory experiments; for the seven-symbol capacity of short-term memory, from immediate recall experiments; for the 200 milliseconds needed to transfer symbols into and out of short-term memory, from experiments requiring searches down lists or simple arithmetic computations.

These things we do learn about the information-processing system that supports human thinking—but it is significant that we learn little more, that the system might be almost anything as long as it meets these few structural and parametrical specifications. The detail is elusive because the system is adaptive. For a system to be adaptive means that it is capable of grappling with whatever task environment confronts it. Hence, to the extent a system is adaptive, its behavior is determined by the demands of that task environment rather than by its own internal characteristics. Only when the environment stresses its capacities along some dimension—presses its performance to the limit—do we discover what those capabilities and limits are, and are we able to measure some of their parameters (Simon, 1969, Ch. 1 and 2).

2.2. Structure of Task Environments

If the study of human behavior in problem situations reveals only a little about the structure of the information-processing system, it reveals a great deal about the structure of task environments. Consider the crypt-arithmetic problem

\[
\begin{align*}
\text{DONALD} & \quad + \quad \text{GERALD} \\
\text{ROBERT} &
\end{align*}
\]

which has been studied on both shores of the Atlantic, in England by Bartlett (1958), and in the United States in our own laboratory (Newell, 1967; Newell & Simon, 1971, Part II). The problem is to substitute numbers for the letters in the three names in such a way as to produce a correct arithmetic sum. As the problem is usually posed, the hint is given that D = 5. If we look at the protocols of subjects who solve the problem, we find that they all substitute numbers for the letters, in approximately the same sequence. First, they set T = 0, then E = 9 and R = 7, then A = 4 and L = 3, then G = 1, then N = 6 and B = 3, and, finally, O = 2.

To explain this regularity in the sequence of assignments, we must look first at the structure of the task itself. A cryptarithmetic problem may be tackled by trying out various tentative assignments of numbers to letters, rejecting them and trying others if they lead to contradictions. In the DONALD + GERALD problem, hundreds of thousands of combinations would have to be tried to find a solution in this way. (There are \(9! = 362,880\) ways of assigning nine digits to nine letters.) A serial processor able to make and test five assignments per minute would require a

\[4\] Some of this evidence is reviewed in Newell and Simon (1971, Ch. 14).
Human Problem Solving

month to solve the problem; many humans do it in 10 minutes or less.

But the task structure admits a heuristic that involves processing first those columns that are most constrained. If two digits in a single column are already known, the third can be found by applying the ordinary rule of arithmetic. Hence, from D = 5, we obtain the right-most column: 
\[ 5 + 5 = T \text{, hence } T = 0, \text{ with a carry of } 1 \text{ to the next column.} \]

Each time a new assignment is made in this way, the information can be carried into other columns where the same letter appears, and then the most-constrained column of the remaining can be selected for processing. For the DONALD + GERALD problem (but not, of course, for all cryptarithmetic problems), it turns out that the correct assignments for T, E, R, A, L, and G can be found in this way without any trial-and-error search whatsoever, leaving only N, B, and 0 for the possible permutations of 6, 3, and 2.

Not only does this heuristic of processing the most-constrained columns first almost eliminate the need for search, but it also reduces the demands on the short-term memory of the problem solver. All the information he has acquired up to any given point can be represented on a single external display, simply by replacing each letter by the digit assigned to it as soon as the assignment is made. Since the assignments are definite, no provision need be made by an error-free processing system for correcting wrong assignments, nor for keeping track of assignments that were tried previously and failed. The human information-processing system is subject to error, however, hence requires back-up capabilities not predictable from the demands of the task environment.

Hence, from our knowledge of properties of this task environment, we can predict that an error-free serial information-processing system using the heuristic we have described could solve the DONALD + GERALD problem rather rapidly, and without using much short-term memory along the way. But if it solved the problem by this method, it would have to make the assignments in the particular order we have indicated.

The empirical fact that human solvers do make the assignments in roughly this same order provides us with one important piece of evidence (we can obtain many others by analyzing their thinking-aloud protocols and eye movements) that they are operating as serial systems with limited short-term memories. But the empirical data show that there are few task-independent invariants of the human processor beyond the basic structural features we have mentioned. Since the problem solver's behavior is adaptive, we learn from his protocol the shape of the task environment of DONALD + GERALD—the logical interdependencies that hold among the several parts of that problem. We also learn from the protocol the structure of the problem space that the subject uses to represent the task environment, and the program he uses to search the problem space. Though the problem space and program are not task-invariant, they constitute the adaptive interface between the invariant features of the processor and the shape of the environment, and can be understood by considering the functional requirements that such an interface must satisfy.

2.3 Problem Spaces

Subjects faced with problem-solving tasks represent the problem environment in internal memory as a space of possible situations to be searched in order to find that situation which corresponds to the solution. We must distinguish, therefore, between the task environment—the omniscient
observer's way of describing the actual problem "out there"—and the problem space—the way a particular subject represents the task in order to work on it.

Each node in a problem space may be thought of as a possible state of knowledge to which the problem solver may attain. A state of knowledge is simply what the problem solver knows about the problem at a particular moment of time—knows in the sense that the information is available to him and can be retrieved in a fraction of a second. After the first step of the DONALD + GERALD problem, for example, the subject knows not only that D = 5, but also that T = 0 and that the carry into the second column from the right is 1. The problem solver's search for a solution is an odyssey through the problem space, from one knowledge state to another, until his current knowledge state includes the problem solution—that is, until he knows the answer.

Problem spaces, even those associated with relatively "simple" task environments, are enormous. Since there are $9! = 362,880$ possible assignments of nine digits to nine letters, we may consider the basic DONALD + GERALD space to be $9!$ in size, which is also the size of the space of tic-tac-toe. The sizes of problem spaces for games like chess or checkers are measured by very large powers of ten—$10^{120}$, perhaps, in the case of chess. The spaces associated with the problem called "life" are, of course, immensely larger.

For a serial information-processing system, however, the exact size of a problem space is not important, provided the space is very large. A serial processor can visit only a modest number of knowledge states (approximately 10 per minute, the thinking-aloud data indicate) in its search for a problem solution. If the problem space has even a few thousand states, it might as well be infinite—only highly selective search will solve problems in it.

Many of you have tried to solve the Tower of Hanoi problem. (This is very different from the problem of Hanoi in your morning newspaper, but fortunately much less complex.) There are three spindles, on one of which is a pyramid of wooden discs. The discs are to be moved, one by one, from this spindle, and all placed, in the end, on one of the other spindles, with the constraint that a disc may never be placed on another that is smaller than it is. If there are four discs, the problem space comprised of possible arrangements of discs on spindles contains only $3^4 = 81$ nodes, yet the problem is nontrivial for human adults. The five-disc problem, though it admits only 243 arrangements, is very difficult for most people; and the problems with more than five discs almost unsolvable—until the right heuristic is discovered.

Problems like this one—where the basic problem space is not immense—tell us how little trial-and-error search the human problem solver is capable of, or is willing to endure. Problems with immense spaces inform us that the amount of search required to find solutions, making use of available structure, bears little or no relation to the size of the entire space. To a major extent, the power of heuristics resides in their capability for examining small, promising regions of the entire space and simply ignoring the rest. We need not be concerned with how large the haystack is, if we can identify a small part of it in which we are quite sure to find a needle.
Thus, to understand the behavior of a serial problem solver, we must turn to the structure of problem spaces and see just how information is imbedded in such spaces that can be extracted by heuristic processes and used to guide search to a problem solution.

2.4 Sources of Information in Problem Spaces

Problem spaces differ not only in size—a difference we have seen to be usually irrelevant to problem difficulty—but also in the kinds of structure they possess. Structure is simply the antithesis of randomness, providing redundancy that can be used to predict the properties of parts of the space not yet visited from the properties of those already searched. This predictability becomes the basis for searching selectively rather than randomly.

The security of combination safes rests on the proposition that there is no way, short of exhaustive search, to find any particular point in a fully random space. (Of course, skilled safecrackers know that complete randomness is not always achieved in the construction of real-world safes, but that is another matter.)

Nonrandomness is information, and information can be exploited to search a problem space in promising directions and to avoid the less promising. A little information goes a long way to keep within bounds the amount of search required, on average, to find solutions.

Hill climbing. The simplest example of information that can be used to solve problems without exhaustive search is the progress test—the test that shows that one is "getting warmer." In climbing a (not too precipitous) hill, a good heuristic rule is always to go upward. If a particular spot is higher, reaching it probably represents progress toward the top. The time it takes to reach the top will depend on the height of the hill and its steepness, but not on its circumference or area—not on the size of the total problem space.

Types of information. There is no great mystery in the nature of the information that is available in many typical problem spaces; and we now know pretty well how humans extract that information and use it to search selectively. For example, in the DONALD + GERALD problem, we saw how information was obtained by arithmetic and algebraic operations. Now, abstracting from particular examples, can we characterize the structure of problem spaces in more general terms?

Each knowledge state is a node in the problem space. Having reached a particular node, the problem solver can choose an operator from among a set of operators available to him, and can apply it to reach a new node. Alternatively, the problem solver can abandon the node he has just reached, select another node from among those previously visited, and proceed from that node. Thus, he must make two kinds of choices: choice of a node from which to proceed, and choice of an operator to apply at that node.

We can think of information as consisting of one or more evaluations (not necessarily numerical, of course) that can be assigned to a node or an operator. One kind of evaluation may rank nodes with respect to their promise as starting points for further search. Another kind of
evaluation may rank the operators at a particular node with respect to their promise as means for continuing from that node. The problem-solving studies have disclosed examples of both kinds of evaluations: for node and operator selection, respectively.

When we examine how evaluations are made—what information they draw on—we again discover several varieties. An evaluation may depend only on properties of a single node. Thus, in theorem-proving tasks, subjects frequently decline to proceed from their current node because "the expression is too complicated to work with." This is a judgment that the node is not a promising one. Similarly, we find frequent statements in the protocols to the effect that "it looks like Rule 7 would apply here."

In most problem spaces, the choice of an efficient next step cannot be made by absolute evaluation of the sorts just illustrated, but instead is a function of the problem that is being solved. In theorem proving, for example, what to do next depends on what theorem is to be proved. Hence, an important technique for extracting information to be used in evaluators (of either kind) is to compare the current node with characteristics of the desired state of affairs and to extract differences from the comparison. These differences serve as evaluators of the node (progress tests) and as criteria for selecting an operator (operator relevant to the differences). Reaching a node that differs less from the goal state than nodes visited previously is progress; and selecting an operator that is relevant to a particular difference between current node and goal is a technique for (possibly) reducing that difference.

The particular heuristic search system that finds differences between current and desired situations, finds an operator relevant to each difference, and applies the operator to reduce the difference is usually called means-ends analysis. Its common occurrence in human problem-solving behavior has been observed and discussed frequently since Duncker (1945). Our own data analyses reveal means-ends analysis to be a prominent form of heuristic organization in some tasks—proving theorems, for example. The procedure is captured in the General Problem Solver (GPS) program which has now been described several times in the psychological literature. The GPS find-and-reduce-difference heuristic played the central role in our theory of problem solving for a decade beginning with its discovery in 1957, but more extensive data from a wider range of tasks have now shown it to be a special case of the more general information-extracting processes we are describing here.

Search strategies. Information obtained by finding differences between already-attained nodes and the goal can be used for both kinds of choices the problem solver must make—the choice of node to proceed from, and the choice of operator to apply. Examining how this information can be used to organize search has led to an explanation of an important phenomenon observed by de Groot (1965) in his studies of choice in chess. De Groot found that the tree of move sequences explored by players did not originate as a bushy growth, but was generated, instead, as a bundle of spindly explorations, each of them very little branched. After each branch had been explored to a position that could be evaluated, the player returned to the base position to pick up a new branch for exploration. De Groot dubbed this particular kind of exploration, which was universal among the chessplayers he studied, "progressive deepening."

5 Brief descriptions of GPS can be found in Hilgard and Bower (1966)
The progressive deepening strategy is not imposed on the player by the structure of the chess task environment. Indeed, one can show that a different organization would permit more efficient search. This alternative method is called the scan-and-search strategy, and works somewhat as follows: Search proceeds by alternation of two phases: (a) in the first phase, the node that is more promising (by some evaluation) is selected for continuation; (b) in the second phase, a few continuations are pursued from that node a short distance forward, and the new nodes thus generated are evaluated and placed on a list for Phase I. The scan-search organization avoids stereotypy. If search has been pursued in a particular direction because it has gone well, the direction is reviewed repeatedly against other possibilities, in case its promise begins to wane.

A powerful computer program for finding checkmate combinations, called MATER, constructed with the help of the scan-search strategy, appears a good deal more efficient than the progressive deepening strategy (Baylor & Simon, 1966). Nevertheless, in chess and the other task environments we have studied, humans do not use the scan-search procedure to organize their efforts. In those problems where information about the current node is preserved in an external memory, they tend to proceed almost always from the current knowledge state, and back up to an earlier node only when they find themselves in serious trouble (Newell & Simon, 1971, Ch. 12 and 13). In task environments where the information about the current node is not preserved externally (e.g., the chessboard under rules of touch-move), and especially if actions are not reversible, humans tend to preserve information (externally or internally) about a base node to which they return when evaluation rejects the current node. This is essentially the progressive deepening strategy.

We can see now that the progressive deepening strategy is a response to limits of short-term memory, hence provides additional evidence for the validity of our description of the human information-processing system. When we write a problem-solving program without concern for human limitations, we can allow it as much memory of nodes on the search tree as necessary—hence we can use a scan-search strategy. To the human problem solver, with his limited short-term memory, this strategy is simply not available. To use it, he would have to consume large amounts of time storing in his long-term memory information about the nodes he had visited.

That, in sum, is what human heuristic search in a problem space amounts to. A serial information processor with limited short-term memory uses the information extractable from the structure of the space to evaluate the nodes it reaches and the operators that might be applied at the nodes. Most often, the evaluation involves finding differences between characteristics of the current node and those of the desired node (the goal). The evaluations are used to select a node and an operator for the next step of the search. Operators are usually applied to the current node, but if progress is not being made, the solver may return to a prior node that has been retained in memory—the limits of the choice of prior node being set mostly by short-term memory limits. These properties have been discussed in Newell and Simon (1971, Ch. 9).
shown to account for most of the human problem-solving behaviors that have been observed in the three task environments that have been studied intensively: chess playing, discovering proofs in logic, and cryptarithmetic; and programs have been written to implement problem-solving systems with these same properties.

2.5 Alternative Problem Spaces

Critics of the problem-solving theory we have sketched above complain that it explains too little. It has been tested in detail against behavior in only three task environments—and these all involving highly structured symbolic tasks. More serious, it explains behavior only after the problem space has been postulated—it does not show how the problem solver constructs his problem space in a given task environment. How, when he is faced with a cryptarithmetic problem, does he enter a problem space in which the nodes are defined as different possible assignments of letters to numbers? How does he become aware of the relevance of arithmetic operations for solving the problem? What suggests the "most-constrained-column-first" heuristic to him?

Although we have been careful to distinguish between the task environment and the problem space, we have not emphasized how radical can be the differences among alternative problem spaces for representing the same problem. Consider the following example: An elimination tournament, with 109 entries, has been organized by the local tennis club. Players are paired, the losers eliminated, and the survivors re-paired until a single player emerges victorious. How should the pairings be arranged to minimize the total number of individual matches that will have to be played? An obvious representation is the space of all possible "trees" of matchings of 109 players—an entirely infeasible space to search. Consider an alternative space in which each node is a possible sequence of matches constituting the tournament. This is, again, an enormous space, but there is a very simple way to solve the problem without searching it. Take an arbitrary sequence in the space, and note the number of surviving players after each match. Since the tournament begins with 109 players, and since each match eliminates one player, there must be exactly 108 matches to eliminate all but one player—no matter which sequence we have chosen. Hence, the minimum number of matches is 103, and any tree we select will contain exactly this number.

There are many "trick" problems of this kind where selection of the correct problem space permits the problem to be solved without any search whatsoever. In the more usual case, matters are not so extreme, but judicious selection of the problem space makes available information that reduces search by orders of magnitude in comparison with what is required if a less sophisticated space is used.

We cannot claim to have more than fragmentary and conjectural answers to the questions of representation. The initial question we asked in our research was: "What process do people use to solve problems?" The answer we have proposed is: "They carry out selective search in a problem space."

The empirical findings, only some of which have been published to date, are collected in Parts II, III, and IV, of Newell and Simon (1971).
space that incorporates some of the structural information of the task environment." Our answer now leads to the new question: "How do people generate a problem space when confronted with a new task?" Thus, our research, like all scientific efforts, has answered some questions at the cost of generating some new ones.

By way of parenthesis, however, we should like to refute one argument that seems to us exaggerated. It is sometimes alleged that search in a well-defined problem space is not problem solving at all--that the real problem solving is over as soon as the problem space has been selected. This proposition is easily tested and shown false. Pick a task environment and a particular task from it. To do the task, a person will first have to construct a problem space, then search for a solution in that space. Now give him a second task from the same environment. Since he can work in the problem space he already has available, all he needs to do this time is to search for a solution. Hence, the second task--if we are to accept the argument--is no problem at all. Observation of subjects' behavior over a sequence of chess problems, cryptarithmetic puzzles, or theorem-finding problems shows the argument to be empirically false. For the subjects do not find that all the problems become trivial as soon as they have solved the first one. On the contrary, the set of human behaviors we call "problem solving" encompasses both the activities required to construct a problem space in the face of a new task environment, and the activities required to solve a particular problem in some problem space, new or old.

3. WHERE IS THE THEORY GOING?

Only the narrow seam of the present divides past from future. The theory of problem solving in 1970--and especially the part of it that is empirically validated--is primarily a theory that describes the problem spaces and problem-solving programs, and shows how these adapt the information-processing system to its task environment. At the same time that it has answered some basic questions about problem-solving processes, the research has raised new ones: how do problem solvers generate problem spaces; what is the neurological substrate for the serial, limited-memory information processor; how can our knowledge of problem-solving processes be used to improve human problem solving and learning? In the remaining pages of this article, we should like to leave past and present and look briefly--using Milton's words--into "the never-ending flight of future days."

3.1 Constructing Problem Spaces

We can use our considerable knowledge about the problem spaces subjects employ to solve problems in particular task environments as our take-off place for exploring how the problem spaces come into being, how the subjects construct them.

Information for construction. There are at least six sources of information that can be used to help construct a problem space in the face of a task environment:

1. The task instructions themselves, which describe the elements of the environment more or less completely, and which may also provide some external memory--say, in the form of a chessboard.
2. Previous experience with the same task or a nearly identical one. (A problem space available from past experience may simply be evoked by mention of the task.)

3. Previous experience with analogous tasks, or with components of the whole task.

4. Programs stored in long-term memory that generalize over a range of tasks.

5. Programs stored in long-term memory for combining task instructions with other information in memory to construct new problem spaces and problem-solving programs.

6. Information accumulated while solving a problem, which may suggest changing the problem space. (In particular, it may suggest moving to a more abstract and simplified planning space.)

The experience in the laboratory with subjects confronting a new task, and forced, thereby, to generate within a few minutes a problem space for tackling the task, suggests that the first source—task instructions and illustrative examples accompanying them—plays a central role in generation of the problem space. The array presented with the cryptarithmetic problem, for example, suggests immediately the form of the knowledge state (or at least the main part of it); namely, that it consists of the same array modified by the substitution in it of one or more digits for letters.

The second source—previous experience with the same task—is not evident, of course, in the behavior of naive subjects, but the third source—analogous and component tasks—plays an important role in cryptarithmetic. Again, the form of the external array in this task is sufficient to evoke in most subjects the possible relevance of arithmetic processes and arithmetic properties (odd, even, and so on).

The fourth source—general purpose programs in long-term memory—is a bit more elusive. But, as we have already noted, subjects quite frequently use means-ends programs in their problem-solving endeavors, and certainly bring these programs to the task from previous experience. We have already mentioned the General Problem Solver, which demonstrates how this generality can be achieved by factoring the specific descriptions of individual tasks from the task-independent means-ends analysis process.

The fifth and sixth sources on the list above are mentioned because common sense tells us that they must sometimes play a role in the generation of problem spaces. We have no direct evidence for their use.

What evidence we have for the various kinds of information that are drawn on in constructing problem spaces is derived largely from comparing the problem spaces that subjects are observably working in with the information they are known to have access to. No one has, as yet, really observed the process of generation of the space—a research task that deserves high priority on the agenda.

Some simulation programs. Some progress has been made, however, in specifying for computers several programs that might be regarded as candidate theories as to how it is done by humans. Two of these programs were
constructed by Tom Williams (1965) and Donald Williams (1969), respectively, in the course of their doctoral research. A General Game Playing Program (GGPP), designed by Tom Williams, when given the instructions for a card or board game (somewhat as these are written in Hoyle, but with the language simplified and smoothed), is able, by interpreting these instructions, to play the game—at least legally if not well. GGPP relies primarily on the first, fourth, and fifth sources of information from the list above. It has stored in memory general information about such objects as "cards," "hands," "boards," "moves," and is capable of combining this general information with information derived from the specific instructions of the game.

The Aptitude Test Taker, designed by Donald Williams, derives its information from worked-out examples of items on various kinds of aptitude tests (letter series, letter analogies, number series and analogies, and so on) in order to construct its own programs capable of taking the corresponding tests.

These programs put us into somewhat the same position with respect to the generation of problem spaces that LT did with respect to problem solving in a defined problem space: that is to say, they demonstrate that certain sets of information-processing mechanisms are sufficient to do the job over some range of interesting tasks. They do not prove that humans do the same job in the same way, using essentially the same processes, or that these processes would suffice for all tasks. It should be noted that the programs written by the two Williamses employ the same kind of basic information-processing system that was used for earlier cognitive simulations. They do not call for any new magic to be put in the hat.

Planning and abstracting processes. The processes for generating problem spaces are not unrelated to some other processes about which we do have empirical data—planning processes. In several of the tasks that have been studied, and especially in the logic task, subjects are often observed to be working in terms more abstract than those that characterize the problem space they began with. They neglect certain details of the expressions they are manipulating (e.g., the operations or connectives), and focus on features they regard as essential.

One way of describing what they are doing is to say that they are abstracting from the concrete detail of the initial problem space in order to construct a plan for a problem solution in a simpler abstract planning space. Programs have been written, in the context of GPS, that are also capable of such abstracting and planning, hence are capable of constructing a problem space different from the one in which the problem solving begins.

The evidence from the thinking-aloud protocols in the logic task suggests, however, that the human planning activities did not maintain as sharp a boundary between task space and abstract planning space as the simulation program did. The human subjects appeared able to move back and forth between concrete and abstract objects without treating the latter as belonging to a separate problem space. In spite of this difference, the data on planning behavior give us additional clues as to how problem spaces can be generated and modified.
3.2 Production Systems

A hypothesis about the structure of a complex system—like a human problem-solving program—becomes more plausible if we can conceive how a step-by-step development could have brought about the finished structure. Minerva sprang full-grown from the brow of Zeus, but we expect terrestrial systems to evolve in a more gradual and lawful fashion—our distrust of the magician again.

Anyone who has written and debugged a large computer program has probably acquired, in the process, a healthy skepticism that such an entangled, interconnected structure could have evolved by small, self-adapting steps. In an evolving system, a limited, partial capability should grow almost continuously into a more powerful capability. But most computer programs have an all-or-none character: disable one subroutine and a program will probably do nothing useful at all.

A development of the past few years in computer language construction has created an interesting possible solution to this difficulty. We refer to the languages known as production systems. In a production system, each routine has a bipartite form, consisting of a condition and an action. The condition defines some test or set of tests to be performed on the knowledge state. (E.g., "Test if it is Black's move.") If the test is satisfied, the action is executed; if the test is not satisfied, no action is taken, and control is transferred to some other production. In a pure production system, the individual productions are simply listed in some order, and considered for execution in turn.

The attraction of a production system for our present concerns—of how a complex program could develop step by step—is that the individual productions are independent of each other's structures, and hence productions can be added to the system one by one. In a new task environment, a subject learns to notice conditions and make discriminations of which he was previously unaware (a chessplayer learns to recognize an open file, a passed pawn, and so on). Each of these new discriminations can become the condition part of a production, whose action is relevant to that condition.

We cannot pursue this idea here beyond noting its affinity to some classical stimulus-response notions. We do not wish to push the analogy too far, for productions have some complexities and subtleties of structure that go beyond stimulus-response ideas, but we do observe that linking a condition and action together in a new production has many similarities to linking a stimulus together with its response. One important difference is that, in the production, it is the condition—that is, the tests—and not the stimulus itself that is linked to the response. In this way, the production system illuminates the problem of defining the effective stimulus, an old bugaboo of stimulus-response theory.

3.3 Perception and Language

We have seen that research on problem solving has begun to shift from asking how searches are conducted in problem spaces, a subject on which we have gained a considerable understanding, to asking how problem spaces—internal representations of problems—are built up in human minds. But the subject of internal representation links problem-solving research with two other important areas of psychology:
perception and psycholinguistics. The further extension of this linkage (see Step 8 in the strategy outlined in our introductory section) appears to be one of the principal tasks for the next decade.

Elsewhere, one of us has described briefly the main connections between problem-solving theory and the theories of perception and psycholinguistics (Simon, 1969, pp. 42-52). We will simply indicate these connections even more briefly here.

Information comes to the human problem solver principally in the form of statements in natural language and visual displays. For information to be exchanged between these external sources and the mind, it must be encoded and decoded. The information as represented externally must be transformed to match the representations in which it is held inside. It is very difficult to imagine what these transformations might be as long as we have access only to the external representations, and not to the internal. It is a little like building a program to translate from English to Language X, where no one will tell us anything about Language X.

The research on problem solving has given us some strong hypotheses about the nature of the internal representations that humans use when they are solving problems. These hypotheses define for us, therefore, the endpoint of the translation process--they tell us something about Language X. The hypotheses should provide strong clues to the researcher in perception and to the psycholinguist in guiding their search for the translation process. Indeed, we believe that these cues have already been used to good advantage in both areas, and we anticipate a great burgeoning of research along these lines over the coming decade.

3.4 Links to Neurophysiology

The ninth step in the strategy set forth in our introduction was to seek the neurophysiological counterparts of the information processes and data structures that the theory postulates. In this respect, we are in the position of nineteenth-century chemistry which postulated atoms on the basis of observations of chemical reactions among molecules, and without any direct evidence for their existence; or in the position of classical genetics, which postulated the gene before it could be identified with any observed microscopic structures in the cell.

Explanation in psychology will not rest indefinitely at the information-processing level. But the explanations that we can provide at that level will narrow the search of the neurophysiologist, for they will tell him a great deal about the properties of the structures and processes he is seeking. They will put him on the lookout for memory fixation processes with times of the order of five seconds, for the "bottlenecks" of attention that account for the serial nature of the processing, for memory structures of small capacity capable of storing a few symbols in a matter of a couple of hundred milliseconds.

All of this is a prospect for the future. We cannot claim to see in today's literature any firm bridges between the components of the central nervous system as it is described by neurophysiologists and the components of the information-processing system we have been discussing here. But bridges there must be, and we need not pause in expanding and improving our knowledge at the information-processing level while we wait for them to be built.
3.5 The Practice of Education

The professions always live in an uneasy relation with the basic sciences that should nourish and be nourished by them. It is really only within the present century that medicine can be said to rest solidly on the foundation of deep knowledge in the biological sciences, or the practice of engineering on modern physics and chemistry. Perhaps we should plead the recency of the dependence in those fields in mitigation of the scandal of psychology's meager contribution to education.

It is, of course, incorrect to say that there has been no contribution. Psychology has provided to the practice of education a constant reminder of the importance of reinforcement and knowledge of results for effective learning. And particularly under the influence of the Skinnerians, these principles have seen increasingly systematic and conscious application in a variety of educational settings.

Until recently, however, psychology has shown both a reluctance and an inability to address itself to the question of "what is learned." At a common sense level, we know perfectly well that rote learning does not provide the same basis for lasting and transferable skills that is provided by "meaningful" learning. We have even a substantial body of laboratory evidence—for example, the research by Katona (1940), now 30 years old—that shows clearly the existence and significance of such differences in kinds of learning. But we have largely been unable to go beyond common sense in characterizing what is rote and what is meaningful. We have been unable because we have not described what is learned in these two different modes of learning—what representation of information or process has been stored in memory. And we have not described how that stored information and those stored programs are evoked to perform new tasks.

The theory of problem solving described here gives us a new basis for attacking the psychology of education and the learning process. It allows us to describe in detail the information and programs that the skilled performer possesses, and to show how they permit him to perform successfully. But the greatest opportunities for bringing the theory to bear upon the practice of education will come as we move from a theory that explains the structure of human problem-solving programs to a theory that explains how these programs develop in the face of task requirements—the kind of theory we have been discussing in the previous sections of this article.

It does not seem premature at the present stage of our knowledge of human problem solving to undertake large-scale development work that will seek to bring that theory to bear upon education. Some of the tasks that have been studied in the basic research programs—proving theorems in logic and geometry, playing chess, doing cryptarithmetic problems, solving word problems in algebra, solving letter-series completion problems from intelligence tests—are of a level of complexity comparable to the tasks that face students in our schools and colleges.7

7For the first three of these tasks, see Newell and Simon (1971); for algebra, Paige and Simon (1966); for letter series, Simon and Kotovsky (1963) and Klahr and Wallace (1970).
The experience of other fields of knowledge teaches us that serious attempts at practical application of basic science invariably contribute to the advance of the basic science as well as the area of application. Unsuspected phenomena are discovered that can then be carried back to the laboratory; new questions are raised that become topics for basic research. Both psychology and education stand to benefit in major ways if we make an earnest effort over the next decade to draw out the practical lessons from our understanding of human information processes.

4. IN CONCLUSION

We have tried to describe some of the main things that are known about how the magician produces the rabbit from the hat. We hope we have dispelled the illusion, but we hope also that you are not disappointed by the relative simplicity of the phenomena once explained. Those who have the instincts and esthetic tastes of scientists presumably will not be disappointed. There is much beauty in the superficial complexity of nature. But there is a deeper beauty in the simplicity of underlying process that accounts for the external complexity. There is beauty in the intricacy of human thinking when an intelligent person is confronted with a difficult problem. But there is a deeper beauty in the basic information processes and their organization into simple schemes of heuristic search that make that intricate human thinking possible. It is a sense of this latter beauty—the beauty of simplicity—that we have tried to convey to you.
Over a decade ago McCulloch (1955) posed the following critical challenge to all life scientists: "To the theoretical question, Can you design a machine to do what a brain can do?, the answer is this: If you specify in a finite and unambiguous way what you think a brain does do with information, then we can design a machine to do it. Pitts and I proved this constructively. But can you say what you think brains do (my italics)?"

The important contribution of McCulloch and Pitts (1943) was to prove, to the chagrin of the vitalists, that any perceptual or cognitive process believed to be carried out by a living nervous system which could be precisely defined, could be logically simulated by a network of abstract neural modules of considerably simpler structure than living neurons. In later developments they illustrated this possibility by showing how such networks had the ability to abstract universal properties of simple geometrical figures, to form general functional schemata for recognizing certain patterns, and to store abstract specifications of the patterns presented for future reference. Their work made precise what others had just hinted at and, thus, opened wide the door to mathematical simulation theory. However, they were unable then, as we are now, to say precisely what "brains" do.

The late John von Neumann, one of the great mathematicians of our day, glimpsed a basic difficulty that he thought explained why simulation models for complex psychological phenomena have failed to develop very far beyond their promising beginnings. He said: "The insight that a formal neuron network can do anything which you can describe in words is a very important insight and simplifies matters enormously at low complication levels. It is by no means certain that it is a simplification on high complication levels (von Neumann, 1966)."

He then illustrates this point with respect to the fact that about one-fifth of the brain is a visual brain, consisting of a network of about two billion neurons. Apparently, this complicated portion of the brain is required to organize and interpret visual analogies, i.e., similarity among patterns.

It is absolutely not clear a priori that there is any simpler description of what constitutes a visual analogy than a descrip-
tion of the visual brain. . . Normally a literary description of what an automaton is supposed to do is simpler than the complete diagram of the automaton. It is not true a priori that this will always be so. There is a good deal in formal logics to indicate that the description of the functions of an automaton is simpler than the automaton itself, as long as the automaton is not very complicated, but that when you get to high complications, the actual object is simpler than the literary description (von Neumann, 1966).

In my opinion, McCulloch on the one hand, and von Neumann on the other, put their fingers on the two most important problems facing psychological simulation theory: (1) deciding what the most general psychological function computed by the human brain is, and (2) constructing models which adequately simulate this surely very complex function that are themselves not too complicated to be understood.

For sake of argument, although I believe it to be true, let us assume that logico-mechanical models (perhaps, as computer programs) can be designed which adequately simulate simple, or even moderately complex, psychological phenomena. It is tempting to assume that if such models can be successfully achieved, then it automatically follows that mutatis mutandis models can be designed to simulate extremely complex phenomena. Unfortunately, this argument is mere handwaving, for as we shall see, there are numerous arguments to the contrary.

In what follows, I will attempt to show that the most characteristic psychological function computed by higher organisms is quite complicated indeed and, thus, requires of any reasonably adequate simulation model a corresponding degree of structural complication. Moreover, I will also argue that although there may be reason for optimism in assaying the difficulties encountered in constructing simulation models for phenomena of low to moderate structural complexity, there is, on the other hand, much reason for pessimism regarding the possibility of achieving similar successes with respect to extremely complex psychological phenomena.

It should be understood, however, that the arguments are aimed at the logic of complex phenomena, in general, rather than at the logic of simulation models per se; that is, complexity not simulation is the logical culprit since it is reasonable to believe simulation of processes below the critical level of structural complexity may indeed be possible.

1. WHAT THE BRAIN DOES: A HYPOTHESIS

Intuitively, it is so easy to demonstrate what I believe the most characteristic psychological function computed by our brains is, that I hesitate to do so, for fear of being thought naive. But it is such an obvious datum that many theorists may have overlooked it. Perhaps, this is because the function is too general to be tested, or its validity too obvious to require testing. It can be demonstrated by the following exercise.

Look around you. Now, close your eyes and describe aloud what you saw. Open your eyes and check to see how accurate your recall was. Although your recall of detail was by no means total, you probably ceased your description of your immediate environment out of boredom, rather than from lack of recall.
Again close your eyes and imagine you are a map-maker who must pinpoint exactly where in the world, in the country, in the state, ..., in the room you are located. What is your present body posture? What is the weather like? The temperature of your room?

Now play autobiographer. Who are you? Where and when were you born? How did you get from there and then to the here and now? What events made you most happy, most sad? Bored you? Are you bored now?

Notice how naturally and quickly you are cognitively geared to answer such diverse and complex questions. We psychologists, who have agonized over theories of rote memory just to explain serial list recall, are understandably annoyed by anyone who reminds us of the immensity of our theoretical problems. But as objective scientists, let us ask ourselves if there are any experiments one needs to run in addition to this simple phenomenological demonstration, to be convinced that the basic psychological function is indeed involved in this exercise. The crucial question which must be answered, however, if this exercise is not to remain trivial, is whether anything precise can be said about the nature of the psychological function implicitly demonstrated.

Minsky (1968) defines one of the main properties of a model as follows: "To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A." The significant question for cognitive theory concerns the nature and the organization of the "object" that is neurologically instantiated and that allows man to answer questions which interest him about the current state or history of his environment, of himself and the relation between the two.

Whatever such an object, or model, is, this ultimately, in whole or part, is what must be simulated if we are to make any significant gains toward a theory of perceptual organization, cognition or memory. Given this interpretation, it follows that the task of the cognitive theorist is essentially that of the simulation theorist, namely, to construct a theoretical model which reproduces the important features of what I will call the "psychological ecosystem," i.e., the joint system consisting of the historical interplay of the environment and organism as connected subsystems.

Several other theorists have suggested ways in which higher organisms cognitively reproduce the significant properties of the psychological ecosystem available to them. For instance, Kenneth Craik (1943), one of the first psychologists to clearly enunciate a cybernetic hypothesis regarding the logico-mechanistic nature of the cognitive reproduction function, stated it this way:

My hypothesis then is that thought models, or parallels, reality—that its essential feature is not "the mind," "the self," "sense data," nor propositions, but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices, which aid thought and calculation... human thought has a definite function, it provides a convenient small scale model of a natural process...

Some theorists, in essential agreement with Craik, have been bold enough to postulate rather specific cognitive structures by which the organism instantiates its knowledge of the world and itself: For Miller, Galanter
and Pribram (1960), the model is called "the image" and is instantiated by a hierarchical organization of TOTE units; according to Hebb (1949) the organism builds up its model by means of "phase sequences" and "cell assemblies;" for Koffka (1935) and other Gestaltists, the model is manifested by a "field of memory traces" which is related to the "environmental field" by a principle of isomorphism; Lashley (1942) postulated "interference wave patterns," and, more recently in the tradition of the field theoretic approach, Pribram (1966) suggested a model based on standing wave patterns governed by holographic principles; in a classical attempt, Tolman (1948), echoed by Bohm (1965) a theoretical physicist, declared for "cognitive maps" and "conceptual maps," respectively; Piaget and Inhelder (1967) sees the child's knowledge as being instantiated in a complex semi-lattice of logically coordinated ideational schemata; and McCulloch and Pitts (1943), as stated earlier, suggest a model consisting of patterns of excitation in a logically simplified neural net. The common core of agreement among these theorists is that higher organisms instantiate their knowledge of themselves and their environments by means of a cognitive reproduction function.

By Neisser's (1967) definition of a cognitive structure as "... a non-specific but organized representation of prior experiences" (including, of course, some experiences which are genetically based), all the above theorists can be said to be wrestling with the central problem of cognitive psychology—that of constructing a logico-mechanistic model which is, in principle, capable of simulating the basic cognitive achievement of higher organisms.

Since the basic achievement of the cognitive function is to reproduce the essential features of the psychological ecosystem by representing them in terms of cognitive structures, then an adequate simulation must provide an accurate reproduction of these cognitive structures. But how complex are these structures and how complex must a model be which adequately simulates their functioning?

2. THE PROBLEM OF COMPLEXITY

In this section, I want to survey some interesting mathematical results which shed light on the following statement by von Neumann (1966):

"It is characteristic of objects of low complexity that it is easier to talk about the object than produce it and easier to predict its properties than to build it. But in complicated parts of formal logic it is always one order magnitude harder to tell what an object can do than to produce the object. The domain of the validity of the question is of a higher type than the question itself."

In just what way von Neumann's remarks are to be taken is somewhat open to question. Unfortunately, he died without clarifying this issue. But it is much too important an issue to ignore, and it would be a mistake to underestimate the importance of von Neumann's thinking along these or any other lines. For this reason it would be well if we seriously pursue the line of argument he suggests. The most important clue has to do with his assertion that "...it is always of a magnitude harder to tell what an object can do than to produce the object. The domain of validity of the question is of a higher type than the question itself."
The crucial concept here is that of type. To those with training in symbolic logic, or set theory, the term "type" most likely evokes the connotation given by Bertrand Russell in his so-called "ramified theory of types," by which he showed mathematicians one way that the paradoxes of Cantor's set theory might be avoided. However, we must look elsewhere to understand what von Neumann apparently had in mind, since the following quote makes it unlikely that this notion of complexity type was what he meant:

There is a concept which will be quite useful here, of which we have a certain intuitive idea, but which is vague, unscientific and imperfect. This concept clearly belongs to the subject of information, and quasithermodynamical considerations are relevant to it. I know no adequate name for it, but it is best described by calling it "complication." It is effectivity in complication, or the potentiality to do things. I am not thinking about how involved the object is, but how involved its purposive operations are. In this sense, an object is of the highest degree of complexity if it can do very difficult and involved things (von Neumann, 1966).

Not only does this quote serve to show that von Neumann, who understood Russell's type scheme as well as anyone, had something else in mind, but it clearly indicates that von Neumann meant his conjecture to apply generally to the concept of effectivity and not just to biological, or mechanical, self-reproduction as special cases of purposive functioning. In this regard, cognitive reproduction qualifies as another degree of effectivity in complication, that is, as another purposive operation a system can perform if it is sufficiently complicated.

Thus, let us construe the notion of "complexity type" very broadly. Von Neumann surely believed his conjecture that complication is degenerative in effectivity below a certain minimum level, to be robust enough to stand almost any interpretation of the concept of complication. He asserts: "We do not know what complication is, or how to measure it, but I think that something like this conclusion is true if one measures complication by the crudest possible standard, the number of elementary parts (von Neumann, 1966)."

What we need is a scheme by which functions or behaviors, such as self-reproduction, can be categorized in terms of how hard they are to compute. Hartmanis and Stearns (1965) investigated such schemes where the computational complexity of a function is measured by the number of operations a multi-tape Turing machine took to produce a sequence of outputs defining the function. They showed, moreover, that no single class contained all computable sequences (functions) on the one hand and, on the other hand, that every computable sequence is contained in some complexity class. In this way, a hierarchy of complexity classes must exist.

What makes their proofs so important is the somewhat surprising fact that they were able to show that these complexity classes are independent of the time scale or of the speed of the components from which the machines are constructed. Thus, they were able to demonstrate that some computable functions possess an inherent complexity that makes them difficult to compute.

In another paper, Hartmanis (1968) shows the existence of a sharp boundary between the time or number of steps it takes for the recognition
of regular versus non-regular sequences. Since computation is used to define complexity of sets, it follows that there must be a jump in the structural complexity of machines which are able to compute functions from lower to higher complexity classes. This suggests the interesting possibility that an investigator, who has classified a large number of machines in terms of some additive measure of structural complexity such as number of parts, will see an apparent sharp bound between machines in terms of the functions they can compute. This would appear as the emergence of qualitatively distinct levels of functioning from their increase along a quantitative dimension of complexity. Perhaps this is what is now being observed by biochemists who catalogue those molecular systems too simple to reproduce from those complex enough to do so.

Thus, there is evidence that some functions are inherently more complex than others. If self-reproduction can be shown to be a function which is of a higher complexity "type" than others then it would follow that it can only be computed by systems which possess structures of high complications. Still additional support has been found for von Neumann's conjecture from what may prove to be legitimate interpretation of existing mathematical results.

Manuel Blum (1967) successfully demonstrated the existence of a class of functions which require an enormous number of steps to be computed, yet has a "nearly quickest" program. Thus, by the Hartmanis and Stearn's metric of complexity based on time-to-compute, such functions must be considered to occupy a superior position on the hierarchy of complexity classes. And just as they found, Blum was able to show that any machine program that can compute such complex functions must take practically as many steps as the "quickest program." In other words, if a machine (program) is able to do the complex functions of this type at all, they must all be beyond some sharp bound on the dimension of structural complications. Consequently, if self-reproduction (as well as the dependent concept of evolution) can be shown to belong to such a class for which no simpler program than the nearly quickest program can compute it, then von Neumann's conjecture must be reckoned with. But can we show that reproductive functions belong to such classes of high complexity? The following argument demonstrating the formal analogy between the concepts of simulation of complicated systems and self-reproduction by complicated systems lends support to this claim.

Results by both Tarski (1956) and Gödel (1965) suggest that it may be quite reasonable to conclude that the problems of self-reproduction and the simulation of highly complex systems involve functions which belong to the same order-type complexity class. This conclusion can be stated precisely as a corollary conjecture to von Neumann's conjecture.

Self-reproduction, reproduction and simulation of living systems (i.e., systems of great finite complexity) are semantic predicates of the same logical type since they are special cases of the truth functional property of predicates. For instance, the following statements require test procedures which are essentially of equivalent complexity:

(a) System X is a "strong" simulation of the behavior of system Y;
(b) System X is a structural reproduction of system Y;
(c) System X is self-reproductive;
(d) System X is a "true" model of system Y;
(e) System X is as complex as system Y.
We will briefly sketch the argument in support of this additional conjecture.

Tarski (1956) was able to show that the truth of a system as a model for another system depends upon the establishment of an isomorphism between the two systems at the appropriate level of analysis called for by the model. In so much as truth is a relationship which subsists by virtue of the correspondence between the state of affairs of one system with another, then truth is not a local property of either system. Rather, truth must be considered a metaproperty of the system of which it is specifically predicated. I.e., truth is a global property of the pair of systems.

By using Minsky's (1968) definition of a model, it is possible to show that a strong formal analogy exists between the theoretical process of constructing simulation models of natural phenomena and the cognitive process by which humans instantiate their knowledge of their environments and their place in them. If we use Minsky's definition as a formula and substitute the appropriate terms the analogy can be expressed quite precisely:

(A) "To an organism B, the set of cognitive structures A* is a model of an ecosystem A to the extent that B can use A* to answer questions that interest him about A." (i.e., B can use A* to react adaptively to its environment.)

(B) "To a theorist B, the machine A* is a model of a psychological phenomenon A to the extent that B can use A* to answer questions that interest him about A."

The definitions (A) and (B) suggest that the process by which the theorist attempts to explain psychological phenomena is logically the same process used by an organism in attaining knowledge of his world, that is, they both instantiate their knowledge in the form of a model by a process of reproducing the significant properties of the phenomenon in question.

Given the above, then reproduction, as well as simulation, belong to the same complexity class since, like the truth function, they are realized only by procedures which establish an isomorphism between two systems.

Another way of putting this argument, to use the earlier results, is as follows:

Let p be a program executed by some organism o in computing its behavior to perform a given psychological task. Assume that f is the behavioral function computed by p and that f belongs to the class of functions which are enormously complex and are known to have a "nearly quickest" program. Now assume that p' is a simulation program designed by the theorist to model o's computation of f. By the Hartmanis and Stearn's and Blum's results it follows that p' must be nearly as complex as p.

This conclusion, if valid, portends dire problems for theory construction in psychology. For if f is a function of sufficient complexity, say reproduction, so that von Neumann's conjecture holds, then so must the conjecture hold for p and the would-be-explanatory model p'. The model p' would by necessity be nearly as complex as p and therefore offer no aid to your understanding of the psychological function being computed by the organism. If this is the case then no simulation model for complex organisms can meet the most fundamental criterion of theory evaluation, namely,
that the theory be in some significant way simpler to understand than the phenomenon it purports to explain. We would then have two equally complex phenomena to explain, e.g., a human and an android.

Still another line of argument suggests that the above conclusion is indeed valid, i.e., that simulation and cognitive reproduction are very complex functions. Minsky and Papert (1967) suggest another interesting measure for the complexity of structures. They introduce the concept of the "order-type" complexity of a structure. The order-type complexity index of a geometric property is essentially determined by the relative degree of complexity of the structures required to define the property. For instance, consider the algorithm for determining whether a figure is convex. Figures la and lb offer examples of forms which do and do not have the property of convexity, respectively.

![Figure 1](image)

The test for convexity requires the determination of whether a pair of points $x$ and $z$, which define a line segment, can be selected such that if they are both contained in the same form then so must be their midpoint $y$. A form for which it is the case that the midpoint lies outside the form while the end points are contained by the form is said to lack convexity. A simple test of all point triples on Figure la fails to find a case in which the midpoint of a line segment falls outside the form unless one of the end points does also. However, the same test applied to Figure lb, shows, in fact, that a line segment can be placed across the pie-shaped slice so that the midpoint of the segment lies outside the boundaries of the form containing the endpoints.

The order-type index $k$ of the property of convexity is therefore at most $k = 3$. This illustrates how the order type index is determined by the cardinal number associated with the minimal structure required to define the property. It is especially interesting to note that some properties of structures are of an indeterminate order type, i.e., are defined over minimal structures whose associated cardinal number is indeterminately large. For instance the property of connectivity, i.e., of a form consisting only of connected parts, is an example of such a property of indeterminate order-type. Minsky and Papert (1969) contrasts the two properties by introducing the distinction between "local" and "global" properties.

It can be decided whether or not a figure has the property of convexity by simply testing to see if all point-triples meet the requirements defined earlier. Since a single test of a triple can decide the negative case then the tests are independent, but since a sum of all tests is required to determine the positive case then the test must agree unanimously. Thus a point triple as a local property is sufficient for determining whether the property of convexity holds for the figure or not. By contrast, however, connectivity is such a property that no procedure
Cognition, Simulation, and the Problem of Complexity

which tests less than all the points in the figure can determine whether the property holds. Thus connectivity requires the test of the whole figure or a global property while convexity only requires a test of a local property.

In other words, there are "global" properties like connectivity which are inherently more complex to define than other properties, local ones like convexity. Moreover, any test program which determines the global property must be of an indeterminate order of complexity in the sense that as the finite size of the figure grows the test increments proportionately. What happens when the object to be tested becomes astronomically complex, say in case of determining the connectivity of a neural net with several billion synaptic junctures?

This result poses a plethora of problems for theoretical psychology since it suggests that there may be "global" structural properties of organisms which the psychologist must determine if he is ever to understand complex human behavior. But unfortunately such properties would require an indeterminate number of parametric experimental tests. Since reproduction is a function (property) which by definition involves an operation that duplicates every significant point of a given system then, like connectivity, it too must be global and therefore, of an indeterminate order type. In order for a test to decide whether one system (the model) is a true reproduction of another system (the phenomenon), an isomorphism must be established between the sets of elements supporting the structures of the two systems. Thus, it would seem that any science like psychology, which desires formal models of highly complex systems, like organisms, will have to consider von Neumann's conjecture a threat to the fulfillment of its explanatory goals.

2.1 A Summary Example

The import of the foregoing argument can be dramatized by considering, for instance, the problem of an alien archaeologist who uncovers a clock and attempts to discover what it is. His investigation of its local properties, i.e., the intrinsic relationships among its gears, springs, etc., regardless of his advanced knowledge of engineering mechanics would fail to uncover the primary function of the device, since this can only be characterized in terms of an extrinsic global relationship of isomorphism that holds between this small mechanical system and the larger celestial one whose schedule it keeps. Clearly, this latter relationship can only be discovered by a study of conjoint systems consisting of the clock and various other sub-systems in its environment. The task of discovering from among all possible natural systems in its environment that exact one from which a high degree of isomorphism holds would be impossible by any bruteforce techniques of discovery. However, no discovery algorithms can possibly be devised to facilitate the solution due to the immense complexity of the domain of possible relationships.

Consequently, if we assume, as so many cognitive theorists have, that a significant function of man's cognitive machinery is the conceptual reproduction of figural relationships among objects in the environment, should we expect our task to be any simpler than that of the alien scientist? It is for this reason, of course, that many theorists seriously question the fruitfulness of theoretical constructs such as "images," "cognitive maps" and other psychological instantiations of figural
knowledge and turn rather to a more abstract conception of cognitive structures in which such knowledge might be instantiated (the "symbolic" mode). These alternative formulations, as attractive as they seem, have not yet been worked out with sufficient precision so that one can be certain they avoid the same difficulties inherent in correspondence models.

Note carefully that the difficulty is not merely that complex systems may be difficult to describe but that discovering significant correspondences, simple or otherwise, may be nearly impossible. It is here that theory is a proper guide for empirical research. Without his incredible theoretical sense Kepler would hardly have sifted for years through the chaos of astronomical data Tycho Brahe had gathered from years of observation until he discovered support for his formulation of the three laws of planetary motion. Similarly, only a deep theoretical belief in the symmetry of nature could have motivated Mendeleev's tireless construction of innumerable tabular arrangements which eventuated in his discovery of the periodic table of chemical elements which established the foundations of his science that empirical alchemy had failed to uncover. But we must admit, if begrudgingly, that the conjoint model systems of planets and ellipses, or chemical elements and tabular forms, which permitted Kepler and Mendeleev's scientific successes are far simpler than the conjoint systems of neural events and psychological constructs we now require.

3. CRITICAL EPILOGUE

The critical reader will no doubt have raised many serious objections to von Neumann's conjecture and, especially, to the interpretation and extension of it as developed in this paper. To attempt a detailed rebuttal of all the philosophical and technical objections that can be levied against the conjecture or against the manner in which the Tarski-Gödel theorem was used in support of it, would require separate treatment in another paper. Instead, I will briefly sketch what seem to me to be the major points of contention.

Two fundamentally different interpretations of the conjecture can be given, one expressing a serious degree of epistemological pessimism while the other merely expresses a pessimism regarding the usefulness of modelling techniques currently in vogue. The strong interpretation of the conjecture maintains that a formal understanding of truly complex natural phenomena is inaccessible to us due to the logical impossibility of codifying significant properties of complex systems in terms of simpler systems of abstract principles (e.g., a model). This interpretation fosters a cynical attitude toward psychology and other sciences aimed at explaining organized complexities; consequently, it seems quite unwise on pragmatic grounds and unnecessary on logical grounds to hold to it with any degree of rational belief.

A weaker, and I believe an interpretation more in keeping with von Neumann's intent, is the version that asserts that some (but not necessarily all) approaches to modelling are inherently inadequate for providing explanations of complex phenomena which pass muster under the criterion of conceptual economy, although the same approaches may be quite adequate when directed toward explaining simple to moderately complex phenomena. Indeed, it is for just this last reason that they have misled us.

In spite of the danger that brief arguments may be less than cogent, let me sketch the manner in which several popular modelling approaches may be
jeopardized by this weaker version of the conjecture. Let me make it perfectly clear, however, that my intent is not to weigh fairly the merits and demerits of these approaches, but rather to place in relief their major points of vulnerability to the conjecture if it should prove valid.

(1) Understanding through synthesis: The approach most vulnerable (and one that seems patently false in any case) is the view that you truly understand any object or phenomenon that you can construct, reconstruct or reproduce. Consider the expression of this view by Miller, Galanter and Pribram (1960): "It seems to the present authors, that attempts to simulate psychological processes with machines are motivated in large measure by the desire to test—or to demonstrate the designer's understanding of the theory he espouses. History suggests that man can create almost anything he can visualize clearly. The creation of a model is proof of the clarity of the vision. If you understand how a thing works well enough to build your own, then your understanding must be nearly perfect."

But does the procedure for constructing or reproducing a phenomenon really imply the principles by which the phenomenon can be adequately explained? Does the fact that an artist produces a masterpiece insure the fact that he knows in anyway but a tacit intuitive sense what the general aesthetic principles are on which his success is based? Nor, as we have argued, does the fact that a master watchmaker is able to construct a mechanical system (the clock) which reproduces accurately the temporal schedule of celestial spheres imply that he also understands the principles of celestial mechanics governing that schedule? In what sense are the astronomical principles, or even mechanical ones, involved in the construction of an accurate time piece, logically related to the engineering principles by which it can be constructed or simulated?

Moreover, contrary to what Miller et al claim, history shows that the mere knowledge of how to construct something, whether it be fire, a tabletop planetarium, a picture, a cake, or an internal combustion engine, is far from being sufficient for a scientific understanding of the object created. To be able to do something implies only that you know how to do it, not that you understand what it is you did. This is analogous to the fact that the learning of a skill may be quite different from the execution of a skill or that the programming of a computer may not involve the principles necessary to understand the most significant functioning of that program. The skill, like the program, may serve as nothing more than a link in an immensely large nomological net of facts and principles which provide the full context required to understand the real significance of either skill or program.

The logical relationship between a model and the natural phenomenon simulated is nothing more than the relationship between an example and that class of phenomena exemplified. To explain what is significant about a model, one must already be able to explain what is significant about the class that makes the model a good example of the phenomenon in question. To simply generalize from the synthesis of an instance to the nature of the whole class is to commit the inductive fallacy. A psychology founded on this attitude toward explanation has the same relationship to a psychology of true scientific understanding as alchemy did to chemistry.

(2) Understanding through analysis: Presumably, the major theoretical motivation for constructing simulation models is the belief that this technique affords a valid test for whether or not some subtle struc-
tural (or functional) analysis of a phenomenon is in fact adequate. Hence it is the analysis that is ultimately desired rather than the simulation per se. This analytic description might be achieved and evaluated by methods other than actual synthesis, say by formal procedures which show either that simple abstract descriptions of the elementary aspects of the phenomenon can be composed to yield the properties of the complex whole, or else by procedures which show how the complex whole can be decomposed into its elementary parts (i.e., by theorem proving). Clearly, a logical composition or decomposition need not be actually realizable by engineering principles as a physical surrogate of the phenomenon in order to possess explanatory value. The material medium of the model need not be the same as for the phenomenon. The question then is not the desirability of an analytic treatment of complex phenomena but whether or not either of these two particular approaches can provide an adequate one, given the conjecture holds. The two approaches rest on a common assumption regarding the nature of the organization of complex phenomena.

There seems to be developing a broad consensus that psychological or other complex natural phenomena possess "hierarchical" rather than "aggregate" structural organizations (Simon, 1969; Arbib, 1969b; Miller et al., 1960). The adequacy of the composability and decomposability approaches to providing an analytic theory of complex phenomena hinge on the degree to which the structure of the phenomena is indeed hierarchical.

a) The composability approach: This approach clearly fails if the higher levels of structural organization are in any sense "richer" than the lower ones so that the dictum "the whole is greater than the sum of its parts" holds. For Simon (1969) this is precisely the defining attribute of what he calls a "complex system":

Roughly by a complex system I mean one made up of a large number of parts that interact in a non-simple way. In such systems, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole. In the fact of complexity, an in-principle reductionist may be at the same time a pragmatic holist.

Evolutionary or developmental models represent special cases of the composability approach. In these, later more complex stages of development of a system are explained in terms of earlier simpler stages of organization. The crucial assumption, of course, is whether the laws of development are continuous across stages so that compositions of structures over time do not result in the emergence of intrinsically creative properties. Again this need in no way be considered a metaphysical emergence, but might simply mean that the surrounding environment of the developing system provides a different set of constraints on the activities of larger, more heterogeneous structures than it does on simpler ones.

For an example of how discontinuities may be manifested when one attempts to apply those laws developed for simpler phenomena to more complex ones, consider the following case: Wigner (1967) following suggestions by Elsasser (1958) has shown that systems capable of self-reproduction and hence evolution can not be adequately explained by the laws of quantum physics. The difficulty is apparently due to measurement or observation problems inherent to quantum physics rather than to any basic inconsistency in either quantum principles or our intuitive conception of self-reproduction.
Nevertheless, here seems to be an explicit case where a qualitative break between simpler and complex phenomena may have been found. Just as new laws had to be defined in order to transcend the molecular model of solids, liquids and gases when plasma was discovered to be the dominant phase of matter in the universe, so Wigner suggests we may need "biotonic laws" to account for the "living" phase of matter which manifests properties for which current physics can give no precise account. (Moreover, given the apparently unique properties of consciousness, rationalization, etc. one might even speculate if still another set of laws beyond the biotonic ones will not be ultimately required—laws to explain the "neuro-psychical" phase of matter.) (Shaw, 1971).

Counter to Wigner's claim for the need for the so-called "biotonic" laws to explain reproduction, Arbib (1969b) claims to have designed an automata model capable of self-reproduction and self-regeneration, which seems without doubt to be physically realizable in terms of electrical circuitry. If this is true, then doesn't Arbib's model refute Wigner's claim, since electrical theory is subject to quantum physical laws?

I suggest that it does not for the reason that Arbib's model exhibits just enough logical "slippage" to still allow for the validity of Wigner's claim. Arbib, in contradistinction to Wigner, is an avowed physical reductionist (op. cit.). He has developed a very interesting theory of self-reproducing automata, although his theory, like that of von Neumann, does not succeed in showing that in principle reduction of biological reproduction to known physical processes is possible because no such model has yet succeeded as an evolutionary mechanism which satisfies the rather stringent "cost" requirements of limited time and locales for change set by nature (Wistar Institute, 1967).

Arbib's model, however, succeeds in simulating several interesting properties, such as individual control on each cell, cell differentiation, and growth. A chief failure of earlier versions of the model was its inability to respond appropriately to damage by regenerating cells. To illustrate how his model can also accommodate this important property of biological systems, he describes a program for a hypothetical three-segment worm which regenerates when cut into pieces. The mechanism for regeneration is quite dynamic having to do with the coordinated interaction of discrete "pulses."

An initial problem with this model is that the head segment A, body segment B, and tail segment C, in the course of regeneration may pass through combinations that fail to preserve the natural order, e.g., instead of just an ABC order segments sometimes combine in ACA orders. This presumably unnatural and therefore undesirable property is quite easily corrected by defining a set of "context-sensitive" rules that specify possible environmental constraints which "filter" out the unwanted combinations. Arbib simply postulates that these pre-filtered states the worm might enter be considered potential rather than actual states.

It is interesting to note that Wigner anticipated such "patchwork" on theories of reproduction using mechanisms governed by quantum physical laws. Such tailoring or attuning of models to avoid local difficulties vitiates any serious claim that successful simulation of "biotonic" properties has been achieved by reductionistic models.
An axiomatic theory of self-reproduction has been formulated by Myhill (1964) which like Arbib's theory, utilizes von Neumann's proof of the existence of a universal constructor—a logical device that recognizes, selects, manipulates and joins components from a stockpile of parts according to a stored "blue-print" so as to reproduce itself. Wigner explains why proof of the existence or physical realizability of self-reproducing systems is insufficient to show that the laws of quantum physics govern such systems and that "biotonic" law are not needed for explaining biological systems. In reference to his proof he says:

The preceding result seems to be in conflict with von Neumann's well known construction of self-duplicating machines. If one tries to confront the evidence of the preceding section with von Neumann's explicit construction, one finds that such a confrontation is not possible because the model used by von Neumann (based on Turing's universal automaton) can assume only a discrete set of states, whereas all our variables are continuous. This permits the postulation of an ideal behavior of the system and the "tailoring" of the substitutes for the equations of motion in such a way that they permit reproduction. (Wigner, 1967).

Wigner's final conclusion is that classical laws of physics may be quite appropriate for giving approximate explanations of reproducing systems but that important details entailing additional principles will be missed. Quantum physics will not clear up such details, hence "biotonic" laws are needed. Moreover, if in the final analysis it turns out that all such simulation models also require post hoc adjustments by context sensitive rules to eliminate annoying details, as in Arbib's model, then one wonders how the theorist will be able to systematically correct similar undesirable features arising in the truly complex models required for simulating the symphony of cognitive processes carried out by the several billion cell colony of the human brain acting in concert so as to adapt to a complex changing environment.

b) The Decomposability approach: This approach, most eloquently defended by Simon (1969) in his essay on "The Architecture of Complexity," rests on the assumption that all natural systems tend toward hierarchical organization as they develop. The case that can be made is persuasive but by no means conclusive. Simon's argument may suffer from selective perception; just because some phenomena are amenable to hierarchical description does not obviate the possibility that most other phenomena do not, furthermore, those that do may still be essentially non-hierarchical under most other circumstances. For instance, the syntax of natural language, as Chomsky and others have argued, seems nearly hierarchical and, therefore, at least weakly decomposable, but its semantics does not. Similarly, resonance and field model descriptions for mnemonic and perceptual processes, following principles analogous to holography as offered by Westlake (1968), Gabor (1968), Longuett-Higgins (1968), and Pribram (1966) also seem essentially non-hierarchical. In what way can we conceive of the periodic table of chemical elements as hierarchical?

We must consider the possibility that human psychological processing when under study may be "shaped" by our laboratory conditions so as to force it to appear spuriously hierarchical. We may even be prone to construct experimental designs which are biased in this way. There is an incipient danger that our intellectual heritage, due to Plato, has indoctrinated us in the view that concepts form a general hierarchical system of classes.
which are related by an ever diminishing intersection of content as they become more abstract and general. Hence, no wonder we find it difficult to construct adequate semantic theories. Cassirer (1923) has to my mind presented cogent arguments that our platonic heritage has in fact perpetuated just such a view of concepts on us. Instead, he suggests, a more accurate view of concepts is that they consist of classes of relationships which form an integrated relational structure such that the whole is superadditive in the sense that it can be neither decomposed into isolable simpler levels nor adequately composed from simples.

Many mathematical groups, unlike either statistical aggregates or most formal rule systems, are non-linearly superadditive in just this sense. Consider, for instance, the successively ordered sequence of perspectives of a three dimensional object (the projective group) say a cube. In isolation, each perspective specifies only a two-dimensional object; randomly ordered they specify either an aggregate of disconnected two-dimensional fragments or a geometrically non-rigid object with quite different figural properties from the geometrically rigid cube. (This demonstration can be easily verified by stroboscopic illumination of a rotating cube). Such systems as these which are too tightly organized to be reductively analyzed von Foerster has termed "coalitions." Such structures offer a more highly organized alternative for modelling complex psychological phenomena than either statistical aggregates or formal hierarchies heretofore exclusively used. Unfortunately, at present there is no rigorous theory of the nature of such structures, although their general properties have been hinted at since Parmenides by Leibnitz, Gestaltists and many current systems theorists.

The requirement that simulation models incorporate context sensitive rules of functions that carry out filtering of undesirable states according to a post hoc selection of environmental constraints, is tantamount to an admission that the laws for composing or decomposing complex phenomena count as only approximate explanations. Precise explanations may require dramatically different principles than those based on hierarchical descriptions of phenomena.

The existence of "cross talk" among different levels, or different stages of development, of complex organic systems may be most clearly evidenced by the "plasticity" commonly exhibited by such systems. Adaptability, regeneration, reproductivity, evolution, etc., are just such properties that seem to require immense cross-referencing and continual interaction among the elements of complex living systems and their environments. The inability of hierarchical schemes to account for the details for context-sensitivity exhibited by the required "cross talk" among different components of complex systems may be just the consequences one might expect if von Neumann's conjecture is true. If so, the moral of this story is that theories which treat complex phenomena as hierarchical will, at best, prove pluralistic rather than unitary, and specialized rather than general.

Hence the conclusions by von Neumann and Wigner regarding the difficulties of studying complex living systems by existing principles seem to be closely in line. The positive import derived from a serious consideration of von Neumann's conjecture leads us then to reopen some old doors and admonishes us to peer more deeply into the nature of complexity and, perhaps, to admit seriously to the need of developing a logic of coalitions to supplant that of hierarchies if it too, like the logic of aggregate structures, continues to prove unsatisfactory.
COMPETENCE MODELS IN MATHEMATICS
First of all I would like to say that I am a professional mathematician and a very amateur psychologist. What I want to present is a class of mathematical systems which I think would be a valuable tool for research in the learning of mathematics. If psychologists would become interested enough to use these systems in their research, then they may find it worthwhile to study mathematics learning.

I have often found that although what the psychologists write about is quite suggestive, it very rarely answers the kind of questions that are of interest to me. In trying to devise ways of teaching mathematics I usually have to operate as an artist. I get some ideas from psychologists; but I usually rely on my own intuition and experience, and then work as any good artist would. If I can stimulate some psychologists to get answers to the types of questions that are of interest to mathematics educators, then maybe we will have a better basis from which to work.

It seems to me that one of the problems of psychologists is that the variety of behaviors involved in the learning of mathematical structures is so large that they are unable to get much of a hold on them. Furthermore, the tasks which they have selected do not provide enough information of the intermediate processes of the learner to be able to analyze his behavior. To overcome these difficulties, I propose using a specific class of mathematical systems which is varied enough to include mathematical learning problems at a great variety of levels and in a great variety of contexts and at the same time be specific enough to provide the researcher with some answers.

These mathematical systems are based on the canonical systems of Post.¹ In these systems there is an alphabet, consisting of a finite number of symbols and we work with strings of symbols in this alphabet. Certain strings are given to start with, and these are called axioms. Then, we have operations, called productions, which tell us how to produce new strings from existing strings.

¹These systems are described in the fourth chapter of Rosenbloom, P. C. The Elements of Mathematical Logic (New York, 1950) and are also described in Kleene, S. C. Mathematical Logic (New York, 1959).
In an experiment we can have a child start with a given string and through the use of productions derive a new string called a target, or we can present to the child the problem of getting to the target in the smallest number of steps. Let me give you an illustration. Let \(\{A, B\}\) be an alphabet and \(ABA\) be an axiom or starting string. Our rules of production are:

\[
\begin{align*}
I \\
1) & \ xBx \rightarrow AxBxA \\
II \\
2) & \ xBy \rightarrow xByx,
\end{align*}
\]

where \(x\) and \(y\) represent strings of \(A\)'s of arbitrary length.\(^2\) The first production applies only to a string which has exactly the same number of \(A\)'s in front of the \(B\) as after it and says if we have such a string, then we can put an \(A\) in front and an \(A\) at the end. The second production applies to any string within our system. It says that we can put at the end of the string an exact duplicate of what goes in front of the \(B\).

One of the first questions to investigate is whether the student can produce a given target from the axiom, \(ABA\). Later on one can ask questions such as "Can you produce a target that has \(AAA\) at the beginning or \(AAAA\) at the end?" To illustrate what is involved we can apply production \(I\) to the axiom and get

\[
ABA \rightarrow AABAA.
\]

If we apply production \(II\) to the axiom, then we get

\[
ABA \rightarrow ABAA.
\]

Observe that production \(I\) cannot be applied to the above result of production \(II\), since what is in front of the \(B\) does not match what comes after it.

What does knowing the rules to this system mean? It means, first of all, recognizing when a given rule applies. So in the case of the string, \(ABAA\), the child has to be able to recognize that rule \(II\) applies to this string but that rule \(I\) does not. Thus, a child trying to produce a given target using these rules gives us behavioral evidence as to whether he knows which rules apply and which rules don't. Secondly, it means that the child is able to apply the given rule. In other words if we give the child something that fits rule \(II\), he will be able to produce the result. Therefore, knowing the rule involves two things and we have means of getting behavioral evidence of both ways of knowing the rule. For example, if we give the child a desk calculator so that we put no burden on his own arithmetic, then the tape that he produces on the desk calculator will give us an exact record of what rules are applied and when they are applied.

If the child plays with this system for awhile as he would with a game, he should be able to construct the following array:

\[\text{In presenting this system to children I call the game "mathematical golf."}\]
From the array and additional productions the child should conclude that the only strings he can produce are of the form $A^m B A^n$ where $m$ divides $n$.

In the process of playing the game the child should also learn some facts about division, and that he is constructing a multiplication table. For example, if the student is asked to produce a string with 6 A's after the B he should be able to discover that he can produce only strings with 1, 2, 3 or 6 A's in front of the B. These numbers are precisely the divisors of 6.

This game can also be played with nursery school children, using blocks of two colors. We used white and black blocks and gave the child to start with the axiom white-black-white. For production I we told him that if whatever is in front of the black block matches what's on the other side of the black block, then he can put a white block in front and behind simultaneously. For production II if he has a row with a black block in it, then he can put at the end an exact match of what's in front of the black block. So, this system lends itself to the investigation of learning at quite a number of levels.

The same system can be presented in many different forms. If we take ordered pairs of integers together with the axiom (1, 1) and make the production rules:

1) $(x, x) \rightarrow (x + 1, x + 1)$

2) $(x, y) \rightarrow (x, y + x)$,

we will have a second game isomorphic to the first game. This form of the system naturally suggests a geometrical game which can be played on graph paper.
In this geometrical game the elements (1,1), (2,2), (3,3), ... form a diagonal which I call "Equality Boulevard." The N axis in Figure 1 could be 1st street going North and the E axis, 1st avenue going East. Thus, production I has the child move one block north and then one block east to get from one intersection on Equality Boulevard to another. Production II requires the student to move x blocks north along street x (e.g., (2,2 --II-- (2,4) along 2nd street). As in the alphabet game he can construct a multiplication table by locating the avenues he can reach along a specified street (e.g., along 2nd street he can only stop at 4th, 6th, 8th, etc. avenues). Also along any avenue he should be able to identify the divisors (e.g., along 4th ave. the divisors are 1, 2 and 4). He may also form the concept of a prime number by recognizing the avenues which have only two points (e.g., 5th ave.). Here you have a very simple game almost any child can learn to play that has both concept formation and strategy.

Let me illustrate another system. Suppose we have a four letter alphabet {A, N, B, C}. The axiom is NB and the productions are:

I
1) XB \longrightarrow AXC

II
2) XNY \longrightarrow XXNYB

The first production says that if we have something that ends in B, then we can put an A in front and change the B to C. The second production says that if we have something that has an N in it, then we can duplicate what goes in front of the N and at the same time put a B at the end.
This of course looks very mysterious but if we replace B and C by 0 and 1, the axiom NB by NO, and the rules of production by

I  
1) $X_0 \rightarrow AX_1$

II  
2) $XNY \rightarrow XXNYO$,

then we have a very simple language for producing binary numerals representing the number of A's which precede N (e.g., AAAAN101).

Since we are also concerned with practical problems of education, we tried a game based on this system with some nursery school children, using blocks. However, even with blocks, the game was too complicated for the children.\(^3\) So we took the productions and made them operations to be done by a team of two children, child A being responsible for what goes before a yellow block (N), and child B being responsible for the white blocks (0's) and the black blocks (1's) which go on the other side of the yellow block. If child B's row ends in a white block, then the rule is that he can change his white block to a black block and child A can add one more block to his row. The other rule was that no matter what these two children had, that child B could always put a white block at the end of his row, but then child A would have to duplicate whatever he had. In this form we found that nursery school children could learn the game and also could have fun with it.

Let us now turn to a set of games which can be played on graph paper. In these games, we start at a given point, P, having integral coordinates $(a, b)$. With the axes of the graph being east, E, and north, N, the production rules are:

I go east $d$ units, and north $c$ units

II go east $f$ units and north $e$ units

where $c$, $d$, $e$, and $f$ are fixed integral values.\(^4\) The target of the game is to arrive at an arbitrary point R with integral coordinates $(g, h)$.

For example, suppose the coordinate pair of our starting point $P$ is $(2, 3)$ and our rules of production are:

I go north 2 units and east 1 unit

II go north 1 unit and east 3 units

If we take the coordinates $(7, 8)$ to be our target, $R$, then we need to apply production I twice and production II once, as shown in Figure 2, to get from $P$ to $R$.

---

\(^3\)We had no problem in teaching this game to eighteen-year-old education majors and they had no trouble teaching it to children. However, we did encounter a great deal of difficulty in teaching this to experienced elementary teachers.

\(^4\)Since we are sophisticated, we can use positive and negative integers and two directions, north and east. With children, of course, we have north, south, east and west and use only positive numbers.
For those who want to use these tasks in learning research, an easy solution, as to whether a target point R may be reached through m productions I and n productions II, can be determined from the linear equations:

\[(h - b) = mc + ne\]
\[(g - a) = md + nf\]

If Cramer's rule is applied to these equations, then one sure solution as to whether the point R or, for that matter, any point can be reached is if the determinant, \(cf - de = \pm 1\). Any game, of course, can be adjusted so that the determinant is not \(\pm 1\) and different kinds of results will be obtained.

These games can also be set up so as to make a minimal demand on a child's verbal ability. The child could put a blue dot on the starting point P and every time he used production I or II he can place a red or green dot, respectively, on the new point. In this manner, by observing the number of reds corresponding to production I and the number of greens corresponding to production II, one can see if the child correctly solved the problem.

Using a different game which can also be played on graph paper, we ran a study with ten-year-olds to see how well children who learn one system transfer to other related systems. However, we hesitate to call this an experiment, since we were not rigorous in our design. The task involved two operations which could be performed on graph paper. One operation, A, was to rotate a ray through 120° (e.g., see Figure 3). The second operation, B, was to reflect the ray with respect to the x axis (e.g., see Figure 4). The students, then performed these operations A and B in all possible ways. They came to the conclusion that there were only six different operation combinations.
Any other combination of operations turned out to be equivalent to one of the above. That is, they found that nothing = AAA = BB, B = ABA, AB = BAA and AAB = BA. Then, they discovered without having to go through the physical process of doing the operations that any word in this two letter alphabet could be reduced to one of the above six combinations simply by applying the three discovered identities and the usual rules of equality (e.g., ABAABAAAABBA = AA). For them this was a big discovery.

We also gave them the same system in other forms. We gave them permutations on three objects and the corresponding Cayley graph. From these three different presentations they discovered that the same abstract system could be interpreted in three different ways.
As a transfer task we then asked the students what would happen if we used a 90° rotation instead of a 120° rotation. They were able to easily generalize to this new system and to another system using 60° rotations. From this point on they were not only able to work with any other system that we gave them involving the operations of rotation and reflection, but they were also able to make up systems of their own.

The games which I have presented are quite flexible. They include recursive systems and are as general as any mathematics described in a precise language.

Obviously, these games have an advantage over other learning tasks in that we can observe what tactics the child uses through the complete record of his performance. This provides us with a useful tool for investigating how his tactics change when we give him different kinds of experiences. (In using these games, one would usually give a series of tasks starting with the moderately simple and gradually leading to the more difficult. The child's performance, which could be recorded by a series of arrows, would show exactly what he has done.) Another advantage in many of these games is that we can precisely define the shortest way of getting from a given start to a given target. Therefore, we have a very clear measure of how well the child's tactics or strategies are working.

There are two other obvious differences between these games and the kind of tasks that Bruner, Austin, and Goodnow used. First of all, in their tasks, which are typical of what I have seen in other psychological research, the only way the child can find out whether he is right or not is if the experimenter tells him. In other words, the child has to guess what is in the experimenter's mind. There is no objective authority for the child. The second thing is that the games which I have presented lend themselves to investigation of the child's learning, under conditions where the learner has as much opportunity as he wants to record data, to meditate or investigate empirically and theoretically, whereas in the kind of tasks that Bruner et al gave, the learner had no opportunity to analyze the data or to think about what he is doing. So the experimenter really has an investigation of the effect of various kinds of mental stresses on the learner, but he does not have an investigation of the process of learning.

From the point of view of mathematical learning, the most interesting tasks are those in which the learner has plenty of opportunity to analyze his data in whatever way seems appropriate and where he has to record what he does in such a way that we can see what he has done. In other words, if he is trying to get from the start to the target and he makes some false starts then this will show up on his record, or if he starts monkeying around or he starts experimenting with some easy targets first, then we can also record this information.

For most elementary teaching and also for investigation purposes, I would concentrate on tasks of this type where to a certain extent most of the answers are known and where ultimately there would be some problems for deciding whether the target is attainable or for giving the best strategy to a target. Inasmuch as the general decision problem is not recursively solvable these problems require intelligence and cannot be done by machine.

MATHEMATICAL REASONING AND
THE STRUCTURE OF LANGUAGE

JOHN CORCORAN

The fundamental idea of my proof theory is none other than to de-
scribe the activity of our understanding, to make a protocol of the
rules according to which our thinking actually proceeds. —Hilbert

This is the first in a series of three articles concerning some of the
interrelations among three areas: first, linguistic work inspired by the
ideas of Zellig Harris; second, logical investigations concerning the
nature of mathematical reasoning; the third, mathematical education. The
main concern is to relate basic linguistic concepts and hypotheses to the
study of deductive reasoning and, then, to suggest applications to mathe-
matical education. Mathematical education is assumed to include not only
efforts to understand teaching and learning of mathematics but also
attempts to improve mathematical teaching in practice.

The global plan of the series is as follows. Part I presents a certain
view of mathematical reasoning together with a discussion of the inter-
relationship between the structure of mathematical discourse and the struc-
ture of normal English discourse. The second article outlines the nature
of a theory of proof and includes a discussion of the utility of such a
theory in mathematical education. The final article in the series develops
several basic ideas involved in the construction of a usable theory of
proof.

For readers with a practical interest in mathematical education, the
central core of the paper is the last section of Part II where the utility
of a theory of proof for mathematical education is treated. From this
point of view, the discussion following the central core is concerned with
the construction of such a theory.

Deductive reasoning has two common uses. First, when a set of known sen-
tences is considered, deductive reasoning is used to extend knowledge.
Second, when a set of hypotheses is considered, deductive reasoning can be
used to refute the hypotheses (jointly considered) by deducing from them
a known falsehood. Regarded in this way deductive reasoning is an instru-
ment of inquiry, absolutely essential not only in creatively advancing
knowledge but also in the course of the less creative process known as

6 An earlier version of this series has appeared in Journal of Struc-
tural Learning, v. 3 (1972), nos. 1, 2 and 3. (See list of references.)
acquiring an education. It should be clear therefore that theoretical and experimental investigation of deductive reasoning will play an enormously important role in advances in mathematical education. These articles represent an initial step in development of a two-fold program of incorporating present knowledge of deductive reasoning into mathematical education and of extending present knowledge in useful directions.

1. MATHEMATICAL EXPERIENCE (INTUITION) AND REASONING

A useful elementary cognitive distinction separates knowledge of (concepts, structures, systems, etc.) from knowledge that (certain propositions hold). Clearly it is possible to have knowledge of the natural numbers and the concepts "greater than," "even," "prime" and "sum" without having knowledge that every even number greater than two is the sum of two primes. The converse, of course, is false. The point is that while "propositional" knowledge (knowledge that) seems to presuppose knowledge of the constituents of the proposition known, knowledge of the constituents of a proposition does not presuppose knowledge that the proposition holds (or fails). Constituent knowledge is outside the scope of these articles.

From the time of the earliest axiomatic theories, evidently, there was a belief that each item of propositional knowledge in mathematics is obtained either from mathematical experience (intuition) or else by logical reasoning from intuitively known propositions. This plausible view is already clearly articulated in Aristotle (Beth and Piaget, p. 37) and emphasized throughout Descartes' writing on mathematical epistemology (Descartes, pp. 7, 8, 10, passim). Often axiomatic organization of theories seems to aim at concentration of the intuitively obtained propositions in the axioms and relegation of logical reasoning to the role of developing the consequences of the axioms. Thus it seems that ordinary use of the axiomatic method reflects a distinction between propositional knowledge obtained through mathematical experience and that obtained by reasoning from the latter.

Although the traditional view, which involves both intuition and logic in mathematics, is still dominant, it now has several rivals, some of which should be mentioned merely to avoid the false impression that the traditional view is the only view. In the first place there are extreme formalists who deny both intuition and logic (as traditionally conceived) any place in mathematics. These thinkers hold that mathematical laws are meaningless strings of symbols derived by arbitrary manipulations from arbitrarily chosen starting points. In the second place there are logicians who deny intuition any place in mathematics. These thinkers hold

---

7 In order to avoid excessive footnoting, items in the list of references are cited by means of parenthesized material including author and location. In case of authors of more than one item, publication date is added.

8 It is probably worth noting that there seems to be no intelligible or useful general theory of constituent knowledge--Plato was reduced to myth-making when faced with this problem and Descartes held that such knowledge is innate (two non-theories to my way of thinking). Later, empiricistic philosophers have held that such knowledge is "derived from sense impressions" (without saying how). Piaget's suggestions (cf. Beth and Piaget, 201ff) that such knowledge is related to operational skills seems promising.
that mathematical laws are derived by logic alone. Finally, there are intuitionists who deny that ordinary logic is applicable in mathematics. These thinkers hold that all mathematical laws are derived by means of intuitive constructions. Only the last of these rival views has significant and growing support among professional mathematicians today, although the other two still enjoy wide popularity. For purposes of argument we assume the traditional view in these articles.

The basic problem of mathematical epistemology is, of course, the problem of giving a correct account of how knowledge in mathematics is obtained. This problem divides into two: one problem at the level of constituent knowledge and one at the level of propositional knowledge. Constituent knowledge, as already mentioned, is eschewed in these articles. According to the traditional distinction the problem of propositional knowledge also divides into two: first, to account for the mathematical experience (or intuition) which provides the "content source" of all knowledge of mathematical propositions, and second, to account for the reasoning which enables deduction of the consequences of the "intuitive knowledge." Intuitive knowledge is also outside of the scope of these articles.

As is already evident, these articles concentrate on the problems of logical reasoning especially in relation to mathematical education. The passage quoted above from Hilbert clearly indicates that he shared our interests to a great extent. Hilbert, however, did not aim at results germane to education and, partly as a consequence, his formulations are all idealized to such an extent that they are largely useless for our purposes. Mathematical logic has not focused on reasoning per se. Indeed, Kreisel has recently inquired concerning how mathematical logic can contribute "to an understanding of actual reasoning" (Kreisel, 1967, p. 206). Kreisel goes on to explain that he has in mind "a descriptive rather than primarily normative science of reasoning" and this is our interest in these articles.

2. LOGICAL CONSEQUENCE AND INFERENCE

Mathematical reasoning, or deductive reasoning, is a process through which a person comes to understand, to know, that if certain statements were true then a certain other statement would necessarily also be true. When a man thinks through the axioms and a certain theorem in geometry and comes to know that if the axioms are true then the theorem must also be true, then that man is engaging in mathematical reasoning. In giving a proof a man expresses his reasoning in writing or in speech. Thus, reasoning is a mental process whereas a proof is a linguistic entity, written or spoken. Apparently we have no access to another person's mathematical reasoning except through the proofs he offers. Thus proofs provide "the data" in the study of mathematical reasoning.

The above paragraph presupposes familiarity with three special concepts: logical consequence, deductive reasoning, and proofs. A brief review of

---

9 For example, in the summer of 1968 a major international conference supported by the National Science Foundation and the Graduate School of the University of Buffalo (SUNY at Buffalo) was devoted almost entirely to various aspects of intuitionism. See Myhill, Kino and Vesley.

10 At the present time there seems to be no satisfactory theory of mathematical intuition. For a brief survey of the field see Beth and Piaget.
these ideas may prove helpful because each plays a prominent role in subsequent developments.

Consider a given set of statements (premises) and a given single statement. It may happen that the given statement follows logically from the premises. That is, it may happen that if the premises were all true then the given statement would necessarily also be true. In this case the given statement is a logical consequence of the premises; the premises imply the given statement. Naturally, even if the given statement does follow there is no guarantee that anyone knows that it follows (otherwise there would be fewer open questions in mathematics).

In order to know that a given statement follows from given premises, it is necessary to reason deductively from the premises to the given statement, to trace out the logical steps whereby one comes to know that if the given premises were true, the given statement must also be true. Thus, deductive reasoning is human activity; it is a mental process subsumed under the broader heading of learning. In a particular case of deductive reasoning, a person learns something, comes to know something, viz., that a certain statement is a logical consequence of certain other statements.

Knowledge that premises imply a conclusion is gained by logical reasoning alone without the use of intuition. Accordingly, such knowledge is without mathematical content. In order for reasoning to lead to knowledge with mathematical content it must be applied to premises that have such content and even then its conclusions can contain nothing not already in the premises. (Some psychologists, including Piaget, have used the term "reasoning" in a broader sense which subsumes not only deductive reasoning which does not "produce content" but also all sorts of mental constructions, inventions and intuitions.)

There is a metaphor which neatly separates the three ideas. A logical consequence connection is a path in logical space; when we reason we trace the path; and a proof is a set of directions we give to ourselves and others for retracing the path.

It is important to distinguish deductive reasoning from the more creative process of discovering a logical connection. Imagine, for example, that you are given a set of premises. You read them carefully and understand them and then you "see" that something else follows. This "seeing" is actually a kind of guessing because you do not know that the additional statement follows until you verify it by step-by-step reasoning. The word "infer" is frequently used to cover three things: the initial guessing, the discovery of the "logical path" and the subsequent deductive reasoning. According to the way that I am using the words "deductive reasoning," one usually does not reason until he has already done at least two things: (1) guessed a "possible" consequence and (2) discovered a "possible" proof. In addition, I should emphasize that deductive reasoning is also involved in "following" a proof, i.e., in seeing for oneself that it shows why the conclusion follows. Naturally, the reasoning involved in following a proof is not nearly as creative as the thought involved in discovering the exact logical connection.

In terms of the usual distinction between the mode of discovery and the mode of justification, deductive reasoning falls squarely within the latter. The mode of discovery, which deserves serious study in itself, includes both discovery that there is a logical connection (that the theorem follows) as well as discovery of the particular logical connection to be
described in the proof. Apparently it is possible to do the former without the latter. In a few cases mathematicians have claimed to know that a certain theorem followed without claiming to know any proof of the fact. It is beside the point that such knowledge could not be recognized as such in any official way. Polya (1954) and Hadamard (1945) have written on the mode of discovery. In any case, deductive reasoning consists solely in verifying an already discovered path.

The reader deserves to be warned that the word "proof" is ambiguous in normal mathematical parlance. It is certainly used in the above sense to indicate an articulation of deductive reasoning, i.e., a deduction or logical derivation. But it is also used to indicate a particular logical connection or path in logical space or something expressed by a deduction. For example, we speak of trying to discover a new proof of a known theorem (from given axioms). Here we are not looking for a new way of describing the known path of reasoning, but we actually want a new path of reasoning—a new way of getting to the theorem from the axioms. (The reason that we would want a new path would be that we find the known one to be devious, round-about, overly intricate, unnatural—in a word, inelegant.) Other writers have pointed out that the difference between the two senses of "proof" is related to the difference between copying a proof and understanding a proof.

It is interesting to notice that the distinction between the relation of logical consequence and the acts of inferring is already implicit in non-technical English. It is grammatically acceptable to say that one statement implies another statement, but it is not acceptable to say that one statement infers another statement. On the other hand, one can say that a person infers one statement from another statement.

At this point we have distinguished logical consequence, an objective logical relation, from deductive reasoning, a human activity. Deductive reasoning is one of the primary activities of mathematicians because mathematicians are concerned to establish logical connections between axioms and other statements. Since reasoning is a human activity it should be expected that some people are more skilled in it than others. It is almost by definition that a good mathematician is more skilled in deductive reasoning than a poor one. I say "almost" because I have heard of mathematicians who are unskilled in following proofs, but who have gained reputations for being able to guess new theorems with uncanny accuracy. In any case, after a statement has been guessed to follow and after a possible path of reasoning to it has been guessed, one may then go through the process of reasoning step-by-step why (or how) it follows. The verbal or written articulation of the reasoning is a proof. Given a written proof, one can retrace the steps of reasoning expressed in it and thereby rediscover (or see for himself) that the conclusion follows.

In view of the fact, alluded to above, that written proofs (linguistic elements) are the main data available for a descriptive study of reasoning, the balance of this article is devoted to a discussion of the relevant linguistic material. Reasoning itself will be more prominent in

Kreisel (1970) uses the term "derivation" to indicate the linguistic object while using the term "proof" to indicate the abstract object expressed by the derivation. Just as a sentence may be said to express a proposition, a derivation may be said to express a proof. Also see my review of Kreisel (1970) in Mathematical Reviews (forthcoming).
the next two articles in the series.

3. LEVELS OF LANGUAGE

Syntactic Stratification.

The notion that language is stratified into increasingly complex levels of organization is clearly reflected in our writing system (cf. Lyons, pp. 53ff, 171ff, and 206ff). A written text is organized in paragraphs. The paragraphs are naturally segmented into sentences. Sentences are perceived as composed of phrases which in turn are composed of words, and the words are strings of letters.

The lowest level of the written language is the alphabet. Next, we have the level of words, then the level of phrases, then sentences and, finally (for the present) the level of discourses which contains paragraphs and "texts." For illustrative purposes, let us assume that English is stratified according to the above scheme. It seems obvious that English is stratified in some way or other, but other schemes may be found more satisfactory than the above (cf. Lyons, p. 206ff and Harris, 1968, p. 9ff).

By "language" the linguist means the spoken language. Thus, we are assuming that the above stratification applies to the spoken language. For our purposes, however, this distinction is of no consequence.

Let us introduce terminology appropriate for a detailed discussion of the assumed stratification of English. The alphabet of (spoken) English is the set of basic, "indivisible" spoken symbols. The objects in the spoken alphabet are generally called phonemes and, in a written alphabet representing phonemes, the representations of e.g., "early" and "yearly" would not suggest that they rhymed, whereas the representations of "sax" and "sacks" would be the same. The lexicon of English is the set of words of English. For the following we do not introduce any special terminology: the set of phrases, the set of sentences, the set of discourses. The corresponding levels are called respectively: alphabetic, lexical, phraseological, sentential and discourse.

It is probably worth noting that we are respecting in large measure the usual distinction between proofs and computations. In a proof the answer (conclusion) is usually stated at the outset and indeed is always available in advance as a goal to be reached whereas, in a computation, the answer (resultant) is usually only stated at the end and is almost never available beforehand as a goal. Secondly, one can follow a proof without having conscious knowledge of the rules whereas "following a computation" is nothing more than checking that each line is written according to a rule of the logic then employed. Thirdly, each line in a proof is a sentence and must make sense whereas in a computation the lines are usually phrases (at best) and can contain uninterpretable parts. Fourth, there are always multiple proofs of a given statement from given premises so that subsequent lines are not determined uniquely given preceding lines whereas, generally speaking, algorithms are deterministic in the sense that the datum uniquely determines the computation based on it.

These distinctions may seem muted in some discussions of mathematical logic because one important problem treated in modern logic is the extent to which reasoning is replaceable by computation.
Reality of Language Structure.

An important theoretical question in linguistics concerns the "reality" of the stratification into levels. For example, could it not be the case that the stratification is only a structure which we find convenient to impose on English but which corresponds to nothing real in English? Many linguists do not regard this as a substantive question because, some reason, "Either there is a 'real' structure to English or there is not. If there is a 'real' structure, then, presumably, if we work hard enough and are flexible and imaginative, then the structure that we find most convenient will correspond exactly to the real structure. Thus, convenience is the important criterion. On the other hand, if there is no 'real' structure, then what else is there besides convenience?" (cf. Quine, 1961, pp. 78, 79) and Harris, 1954, pp. 36ff).

There is additional debate concerning which levels are "real" and what kinds of reasons are relevant to determining the "reality" of a level. Some linguistic work suggests that some relevant evidence can be gleaned from studies of the patterns of stress and intonation and of the co-occurrence restrictions in actual speech. Other linguists (Chomsky, 1957, pp. 85-88) have made interesting suggestions concerning ways of justifying an intermediate level, B, given the existence of two levels A and C.

My own inclination is to regard questions concerning "reality of levels" as substantive and certainly not a matter of the linguist's convenience. If one aims at describing language as it is then it makes sense to expect that one's success will be useful where language interacts with other phenomena--psychological, physiological, educational, social, etc. Regardless of how convenient one finds description of chemical phenomena in terms of atoms, would one take it seriously unless one believed that there are such things (cf. Kreisel, 1967, p. 208)?

In my opinion, the above questions are important and difficult. But they are outside the scope of this paper and are mentioned only because they may have started to bother the perceptive and critical reader. My brief remarks are intended as an acknowledgement of the difficulties.

Internal Structure of Levels.

Now, we wish to consider the internal structure of each of the given levels. What we mean by the internal structure of a level is merely how the elements of that level are interrelated. We disregard any possible internal structure on the alphabetic level because we are considering the phonemes (or letters) to be indivisible units.

Generalizing on Harris' ideas (1951, pp. vi, vii), we postulate a "kernel/compound" structure at each of the higher levels. This means that each element in a given level is either a simple element or a "combination" of simple elements. The set of simple elements of a level is called the kernel of that level, the non-kernel elements are called compounds. The kernel on the lexical level might contain 'black,' 'bird,' and 'like,' while "bird-like," "unbird-like," and 'blackbird' might be among the compounds. The kernel on the sentential level might include 'Birds fly' and 'Fish swim,' while 'Birds fly and fish swim' and 'Birds do not fly' would be compounds. On the discourse level the kernel would contain only sentences which can "stand alone." For example, 'Birds don't swim' would be in the kernel whereas 'They fly' would not.
Among the compound discourses would be things like the following: 'Birds don't swim. They fly.'

Notice two things. First, in each case, kernel elements were constructed directly only using elements on the lower levels. Second, compounds were composed of kernel elements, using only lower level elements as "connections." To say that this is true, in general, is most certainly wrong (cf. section 6 below). However, for the sake of our illustration, we are making this assumption.

The total set of assumptions that we have made about English amounts to something approaching the simplest non-trivial hypothesis about the structure of English—and as such it is certainly wrong. It is to be emphasized that the assumptions are made in a heuristic spirit only.

Let us briefly review the assumptions. First, we assumed that English is stratified into five levels roughly corresponding to letters, words, phrases, sentences and discourses. Second, we assumed that each level about the alphabetic has a kernel/compound structure. Third, we assumed that each of the higher levels is somehow "obtained" by combining only elements on the same or lower levels. Our third assumption is meant to imply that, given a description of the spoken alphabet of phonemes, the following hold: first, that the words can be described without reference to phrases, sentences, or discourses; second, that the phrases can be described without reference to sentences or discourses; and third, that the sentences can be described without reference to discourses. The last assumption is most implausible (see section 6 below) although it has been accepted by many linguists (Lyons, p. 171).

Semantic Stratification.

It is interesting to compare the levels of language with meanings as "perceived" through written language. On the lowest level, we have units which are perceived as written language (as opposed to mere marks), but which are not necessarily meaningful as such. For example, the letter 'p' is not meaningful, but the letter 'a' can be meaningful.

Next, we have words which are definitely meaningful. The kernel words can be thought of as words which do not have other words as parts. Actually one kernel word could have other words as parts, literally, but it would still be regarded as a kernel word if its meaning was not related in any ways to the meaning of its part. For example, 'dog' has 'do' for a part, but the meaning of 'dog' is in no way related to the meaning of 'do.' Naturally some words are composed of only one phoneme, e.g., 'a' and 'I.' Thus, the lexical level (or word level) is the primary level as far as meaning is concerned (cf. Hockett, p. 123ff). In regard to meaning, the alphabetic level is dispensible—we utter words and it is accidental that they are made up of phonemes (letters). (Theoretically we could have a language in which each word was a single phoneme so that the lexical and alphabetic levels would be the same. In this case after a certain number of words were introduced into the language we would have to have extraordinarily sharp ears (and "sharp" tongues) to communicate. Premack (1971) has used such a "language" in connection with experiments in animal communication.)

One interesting thing about the phraseological level is that the meanings of complex phrases are dependent in a regular way on the meanings of the words which are their parts, e.g., the meaning of 'the king of Iowa' is
Mathematical Reasoning and The Structure of Language

187

certainly dependent on the meanings of the words in it. At this level, however, the meanings are still in a sense word-meanings—they are not yet sentential in nature. They are not true or false, for example. We correctly speak of noun phrases, verb phrases, adjective phrases, and adverbial phrases—indicating that the meanings of phrases are "functionally" the same as the meanings of words.

On the sentential level a new kind of meaning emerges. Indeed, it might be said that communication begins on the sentential level because although the utterance of a word or phrase may permit us to know what the speaker is talking about, it will not tell us what he is saying about it. It is important to notice that the meaning we get from a sentence depends not only on the meanings we attach to the particular words in the sentence, but also on the way that we hear (see) the sentence composed of phrases. Consider the following:

You know how sincere freshmen are.

(Do you see 'sincere' grouped with 'how' or with 'freshmen'?)

On the level of discourse still another kind of meaning emerges. The point here is that the kind of thing communicated by a discourse is generally richer and more complex than the kind of thing communicated by a single sentence and qualitatively different as well. Reasoning is generally communicated in a discourse, rarely if ever in a sentence. In particular, a declarative sentence could be said to communicate a "proposition," whereas a paragraph composed of declarative sentences could communicate propositions together with an organization of them.

For example, some expository discourses begin with a general conclusion. Subsequent sentences clarify the import of the first sentence and give reasons for accepting it. Other examples of discourses also indicate that a discourse conveys an organization of the material expressed by its sentences. Recipes and other "how-to-do-it" discourses come to mind here.

It is even more instructive to consider proofs as examples of discourses (Harris, 1954, pp. 39, 44). What is communicated in a proof is the reasoning from its assumptions to its conclusion and generally none of the sentences in a proof are asserted as declarative sentences. We can reason from admittedly false assumptions to an admittedly false conclusion and communicate the reasoning in a correct proof—necessarily not asserting (as facts) any of the sentences in the proof (cf. section 1 above).

Some linguists have suggested that the meaning of a discourse is merely the meaning of the logical conjunction of all the sentences in the discourse. This is obviously not the case because imperatives and declaratives are not generally conjoinable. Moreover, even in the case of discourses including only declaratives it implies that the order of occurrence of sentences would be irrelevant. It is clear that the order of occurrence of sentences in a discourse is of crucial importance with regard to the meaning of a discourse. As an experiment one might

13 At risk of belaboring the obvious, consider the following point. In any proof by contradiction there is a contradiction. Does this mean that the proof is contradictory? Surely the conjunction of the lines of the proof is!
permute the sentences of a given paragraph and then see if the result means the same as the original. (For example, try interchanging the first and third sentences of this paragraph.)

A general characterization of the kind of thing expressible in an extended discourse over and above the kind of thing expressible in a single sentence has not yet been carried out. Harris has done some preliminary work on the problem of characterizing the distributional structure found in discourses. For example, he argues that there are sequences of sentences which can be conjoined to form a large sentence but which do not form a discourse. He concludes that the co-occurrence restrictions operative on the discourse level cannot be reduced to those associated with conjunctions at the sentential level (1968, p. 131). The general point that he wants to establish is that "... discourses are not merely sequences of sentences ... they show a certain additional structure." (1969, p. 146).

Most of the linguistics before the 50's was focused primarily below the sentential level and much of the work at the sentential level was piece-meal and (therefore?) unexciting to persons with a mathematical outlook. The interesting developments started in the 50's with Harris' investigations above the sentential level. As a result of Harris' work, the sentential level was investigated in a more systematic and mathematically interesting fashion. As mentioned above, Harris had become interested in the now obvious fact that certain stretches of speech composed of several sentences have a definite kind of structure not reducible to sentential structure (Harris, 1954, p. 44ff.). Harris (1952, section 1) was conscious even then that he was initiating a radical departure from previous limits of linguistics. He recognized that "... descriptive linguistics generally stops at sentence boundaries..." and that his was "the problem of continuing descriptive linguistics beyond the limits of a single sentence at a time."

One reason for the absence of linguistic work at the discourse level was widespread belief in what must now be regarded as a peculiar myth, viz., that the sentence is the highest-level grammatical form. This unfortunate doctrine is apparently due to the French linguist Meillet although it gained currency through the writings of the American linguist Bloomfield (Bloomfield, 1926, pp. 153, 158). As was argued above, there is clear linguistic structure above the sentential level—not every list of sentences forms a paragraph just as not every list of words forms a sentence. Indeed, grammatical analysis of (especially mathematical) discourses requires rejection of the Meillet-Bloomfield doctrine. This doctrine is still found in recent writings in linguistics (cf. Lyons, pp. 172ff and Hockett, pp. 199ff), although it has never been supported by any argumentation whatever!

4. MATHEMATICAL DISCOURSE

So far we have distinguished five levels of language: the alphabetic, the lexical, the phraseological, the sentential and the discourse levels. Each of the higher levels encompasses an increasingly complex structure which depends on the levels below it. When we apply the conceptual import of these distinctions to formalized mathematical communication, there are no new difficulties. The basic alphabet of primitive symbols provides the alphabetic level. For example, in a language designed for
Mathematical Reasoning and The Structure of Language

arithmetic we might have an alphabet containing: a symbol \( x \) and an "accent" ' for constructing variables (\( x, x', x'', \), etc.), the arithmetic symbols, 0, 1, +, -, <, and the logical symbols &, v, ~, \( \rightarrow \), =, etc.

The lexical level would include the so-called atomic terms, viz., 0, 1 and the variables (compound symbols obtained by "accenting" the symbol \( x \)). On the phraseological level we would have the atomic terms together with the molecular terms. More precisely the phraseological level includes all terms which enter equations, viz., 0, 1, \( x \), (0+1), ((0+1).1), (\( x ''' \)) \( x ''' \), etc. The sentential level would contain: the equations, ((0+1) = \( x \)), etc.; the inequalities (((0+1)+1) < 1), etc. and all of the compound formulas made up from equations and inequalities by use of quantifiers and connectives. Finally, the discourse level would include the proofs.

Thus we have an analogy between the levels of English and the levels of a language or arithmetic according to which the discourses of English correspond to the proofs in the arithmetic language.

5. GRAMMARS

One of the main projects of modern linguists is to give a complete and systematic description of the entire English language. Linguists and philosophers have suggested that English be regarded as two separate but interrelated systems (cf. Harris, 1959, p. 39), a syntactical system (or system of symbols regarded abstractly) and a semantical system (or system of meanings). A description of the syntactical system of a language is called a grammar of that language (Chomsky, 1957, pp. 11, 13ff).

Given the above assumptions it is reasonable to require that any adequate grammar of English must consist in adequate descriptions of each level in terms of the preceding level (or levels) together with whatever other concepts are needed (Chomsky, 1957, pp. 11, 18). More generally, a description of the syntactical system of English should be stratified in accordance with the way that English itself is stratified. In our case, a description of a given higher level must account for how each element of that level consists of elements of the lower level(s). Thus, a grammar of English will consist in: (1) a description of the alphabetic level, (2) a system of rules describing how the words are built up of phonemes and words, (3) a system of rules describing how the phrases are built up from phonemes, words and phrases, (4) a sentential grammar or system of rules describing how sentences are built up, (5) a discourse grammar describing how discourses are built up. In short, a grammar of English would consist in a description of the basic alphabet together with four systems of rules, each subsequent one of which depends on the lower level systems.

Without any loss of generality we can think of each of the four rule systems as including two types of rules: first, "initial-string" or "kernel" rules which describe the kernel; second, production rules which specify the compounds by indicating how the compounds may be constructed.

It is unfortunate that Chomsky himself has not continued to use the term grammar in this clear sense. Later, a "grammar" is a device for associating "structural descriptions" with sentences (1965, p. 3) and more recently a "grammar" relates meanings to sentences.
from the kernel elements. Rule systems of this sort are sometimes said
to be in "kernel/transformation" form. Let $A^0$ represent the alphabet, $A^1$
the set of words, $A^2$ the set of phrases, $A^3$ the set of sentences and $A^4$
the set of discourses. Let $P_1$, $P_2$, $P_3$ and $P_4$ be the rule sets which
describe (produce) $A^1$, $A^2$, $A^3$ and $A^4$ respectively. If we indicate
the levels that can be referred to in a rule set by writing the names of
those levels after the names of the rules, then, according to our third
assumption, a grammar of English can be represented as follows:

$$
\begin{align*}
A^0 & \quad \text{alphabet} \\
P_1 (A^0 + A^1) & \rightarrow A^1 \quad \text{lexicon} \\
P_2 (A^0 + A^1 + A^2) & \rightarrow A^2 \quad \text{phrases} \\
P_3 (A^0 + A^1 + A^2 + A^3) & \rightarrow A^3 \quad \text{sentences} \\
P_4 (A^0 + A^1 + A^2 + A^3 + A^4) & \rightarrow A^4 \quad \text{discourses}
\end{align*}
$$

The arrow means "produces" or "describes." The third line can be read:
a set of rules depending on elements of the alphabet, the lexicon, and
the set of phrases describes the set of phrases.

Below we give an example of a grammar of this sort to describe part of
the arithmetic language mentioned above.

**Partial Grammar of Arithmetic Language**

**Alphabet**

$A^0$: $x, ', 0, 1, +, - , <, =, \& , v, \Rightarrow , ), (, \Rightarrow , L$

**Lexical Grammar**

$P_1$: Initial Rules: (a) 0 and 1 are constant words, (b) $x$ is a variable word

Production: If $S$ is a variable word then $S'$ is a variable word

$A^1$: (sample words), 0, 1, $x$, $x'$, $x''$, $x'''$

**Phraseological Grammar**

$P_2$: Initial Rule: all words are phrases

Production: (a) If $S_1$ and $S_2$ are phrases then $(S_1 + S_2)$ is a phrase,
(b) If $S_1$ and $S_2$ are phrases then $(S_1 \cdot S_2)$ is a phrase

$A^2$: (sample phrases) 0, 1, $x$, ..., $(0+0)$, $(0+1)$, ... $(0+(0+0))$, 
$(0+(0+1))$, ...

**Sentential Grammar**

$P_3$: Initial Rule: If $S_1$ and $S_2$ are phrases then $(S_1 = S_2)$ and $(S_1 < S_2)$
are sentences
Productions: (a) If $S$ is a sentence then $\neg S$ is a sentence, (b) If $S^1$ and $S^2$ are sentences then $(S^1 \& S^2)$ is a sentence, (c) If $S^1$ and $S^2$ are sentences then $(S^1 \lor S^2)$ is a sentence, (d) If $S^1$ and $S^2$ are sentences then $(S^1 \Rightarrow S^2)$ is a sentence.

$A^3$: (sample sentences) $(0=0)$, $(0=1)$, ..., $(0<0)$, $(0<1)$, ..., $(0=0)$, $(0=1)$, ..., $(0<0)$, $(0<1)$, ..., $(0=0) \& (0=1))$, $(0=0) \lor (0=1))$, ..., $(0<0)$, $(0<1)$, ...

Discourse Grammar

$P^4$: ???

$A^4$: (sample proofs) ???

The above is a partial grammar of the part of the arithmetic language not involving the quantifiers. It is obvious that all such sentences are described or produced by $P^3$, i.e., are in $A^3$. $P^4$, which is presently left out, would generate or describe the correct proofs in this part of the arithmetic language. Before we give our opinion of what this would be like, we want to discuss the nature and value of a correct theory of proof. In the final article of the series we will contrast two possibilities for $P^4$.

6. STATUS OF ASSUMPTIONS

This section of the paper can be omitted in first reading by those who are not versed in modern linguistics and whose interests are more directed to mathematical education and epistemology. The present section is somewhat more technical than previous sections and is designed to clarify the status of the three assumptions about the structure of English which were made in a heuristic spirit for purposes of discussion.

Linguists take for granted not only the existence of levels but also their ordering. The problematic status of the existence assumption was alluded to above and it will become clear shortly that there may be problems in defining the order relation on levels. The immediate conclusion concerning our above assumptions is that the first assumption concerning the stratification of English into five ordered levels is vague to some extent. Although this vagueness is unsatisfactory, not only is the assumption useful in linguistics (Lyons, p. 170ff.; Chomsky, 1957, p. 85ff) but it is probably no more vague than certain crucial concepts of physics at a comparable stage in the development of that science (Kreisel, 1967, p. 222, 223). It is probably necessary to accept vague concepts in order to develop a precise science from scratch although many of us are annoyed when forced to make this admission. 

Reflection on this point will indicate the inadequacy of the view of the scientist as a kind of decision machine which assigns truth values to sentences. This view presupposes that the language of science exists antecedent to scientific activity whereas in fact part of scientific activity consists in the creation of scientific language.
Our second assumption was to the effect that each level had kernel/compound structure. Again there is a certain vagueness but this much is clear—having kernel/compound structure entails the existence of minimal elements not containing other elements of the same level and that all non-minimal elements are composed of minimal elements in a regular fashion. This assumption seems trivial, almost tautologous. One may note in passing the remarkable fact that the modern revolution in linguistics had its start in Harris' decision to exploit the kernel/compound assumption in the study of linguistics. Perhaps more remarkable is the fact that kernel/compound structures had been widely used in mathematical logic for about half century before linguists began using them (cf. Chomsky, 1965, p. 8).

Perhaps the simplest kernel/compound structures are those described by sets of productions which are called recursive by linguists. This term is used in a more restrictive sense by logicians for whom it is a synonym for "mechanically decidable" (Hatcher, p. 215). A set of rules is called recursive if some of its productions operate on its own resultants. The above grammar of the arithmetic language is recursive in every level. Chomsky has emphasized that the whole enterprise of constructing a grammar involves "generation" of an infinite set by means of a finite set of rules and it is obvious that the latter requires a recursive set of rules. The belief that every human language requires a recursive grammar now is wide-spread and certainly useful, but it may not be irrelevant to point out that the reasoning which supports it is circular. The usual reasoning begins by postulation of recursive rules to account for a finite number of constructions. Such rules necessarily go beyond the finite basis. It is then argued that all of the resultants of the rules are also part of the language and thus that the language is infinite. At this point it is concluded that since the language is infinite, a recursive set of rules is necessary to describe it. There is a tendency to overlook the fact that the original rules were postulated by the linguist and not somehow "given" in the data.

As was already noted, the above arithmetic language is stratified into levels in such a way that lower levels are independent of higher levels. For example no phrase contains a sentence as a part. Let us say that a language is super iterative if lower level elements depend on higher level elements. An adequate grammar for a super iterative language might require a lower level production to operate on resultants of higher level productions. Since, in the normal case, higher level productions involve lower level elements there is two-way coupling between lower and higher level grammars. English is obviously super iterative as the following two examples indicate.

1. a. Joe is an atheist.  
   b. That Joe is an atheist

2. a. Which is greater than two  
   b. The least number which is greater than two

In each of the examples (a) is a sentential level expression which is part of the phraseological level expression indicated by (b). Obviously the phraseological expressions in (b) can enter as parts of other sentence level expressions. Since our third assumption held that English was not super iterative it is obviously false. Not only is it false but it seems to be useless except for heuristic purposes despite the fact
that it has been accepted by linguists (cf. Lyons, pp. 171, 206, 207).

The above arithmetic language is of course not the only language suitable for development of arithmetic and, indeed, there are several others some of which are super iterative. In particular, some of them contain phrases which are constructed from sentences. For example consider the noun phrase

the least number greater than 2.

This would be expressed in some languages (cf. Hatcher, p. 64ff) by

ux((1+1)<x)

which contains the sentential expression ((1+1) < x).

In the case of super iterative languages it will not do to define one level to be lower than another when the latter has elements containing as parts elements of the former. In the case of the mathematical languages described in Hatcher (loc. cit.) the containment criteria can be used when restricted to the kernel, i.e., one can say that the phraseological level is lower than the sentential level because the kernel sentences contain kernel phrases as parts (and not vice versa).

The notion of super iterativity has some rather important ramifications for modern linguistics. Most modern linguists have followed Chomsky's idea (1957, p. 13) of considering a language to be a set of sentences. Consequently, development of sentential grammars is correctly taken as one class of major linguistic problems. However, as is to be expected, Harris (1968, p. 14) thinks of a language as a set of discourses and from a Harrisian standpoint linguists should be aiming at construction of discourse grammars instead of (or in addition to) sentential grammars. Moreover, if natural languages are super iterative at the discourse level then construction of adequate sentential grammars is impossible without simultaneous construction of discourse grammars. It may well be the case, then, that current efforts in linguistics will require substantial reformulation (cf. Chomsky, 1957, pp. 56-60).
There are few employments of life in which it is not sometimes advantageous to pause for a short time, and reflect upon the nature of the end proposed. —Boole

This is the second of a series of three articles dealing with application of linguistics and logic to the study of mathematical reasoning, especially in the setting of a concern for improvement of mathematical education. The present article presupposes the previous one. Herein we develop our ideas of the purposes of a theory of proof and the criterion of success to be applied to such theories. In addition we speculate at length concerning the specific kinds of uses to which a successful theory of proof may be put vis-a-vis improvement of various aspects of mathematical education. The final article will deal with the construction of such a theory.

1. PROOFS AND RULES OF INFERENCE

As we have been using the word above, a proof is an articulation of deductive reasoning from premises to conclusion. Thus, when a mathematician writes a proof he is primarily interested in communicating his reasoning to others. He is explaining to others his reasoning that if the premises are true, the conclusion must also be true. Secondarily, he is recording a mental process/event—viz., the particular process of reasoning from those particular premises to that particular conclusion during a particular time interval.

Regularity in Proofs. If we consider proofs that we have written or if we survey the proofs found in the literature of mathematics we find many repetitions of simple patterns. This is a clue to the fact that the writing of proofs is a rule-governed activity. However, if we recall our experiences we will notice that in writing proofs we do not think of ourselves as following rules. It is only after the fact that we see the patterns and postulate the existence of the rules to account for the regularity. This situation is analogous to the situation involving writing of sentences. After seeing many examples of sentences, we notice repeating patterns and postulate the existence of rules to account for

16. The nature of rule-governed activity is treated in several articles in this book.
the regularity. Sentences are constructed according to rules but we are not conscious of following rules in writing sentences. The same with proofs.\textsuperscript{17}

When you write a proof you are generally doing (or redoing) the reasoning that you are expressing in the proof. Moreover, when you are reasoning in a particular branch of mathematics (e.g., geometry or arithmetic) you are generally thinking about the subject matter of that branch—although, as Hilbert, Boole and others point out, if your reasoning is correct, the subject matter is irrelevant and the reasoning would apply equally well to any other subject matter.\textsuperscript{18} The point that I am making is that when you are writing a proof you are too busy to think of any rules even if you knew which ones to think of. This is exactly analogous to speech: when you utter a sentence you are generally thinking about what the sentence is about and thus are too busy to bother with rules. Indeed, for example, as you begin to learn a foreign language in a classroom situation, as long as you have to think of the rules you generally make rather dull conversation because you are too busy to give much thought to what you are talking about. Thus, carrying this over to reasoning, if you knew the rules explicitly and actually thought of them while you reasoned you would likely not get very far in your mathematics.

Rules of Inference. Let us use the term "rule of inference" to refer to the rules according to which proofs are constructed. The rules of inference are rules for constructing proofs in the same way that the rules in a sentential grammar are rules for constructing sentences. Because of our hypothesis that the discourse level, which includes the proofs, must have kernel/compound structure there will be two types of rules: initial string rules asserting that certain strings are proofs \textit{ab initio} and production rules which build up compound proofs from simpler ones. As a result of my own experience in formulation of rules of inference it seems that each production rule can be written in the following form: if such-and-such is a proof then the result of adding so-and-so to the end of it

\textsuperscript{17}The question of the reality of rules of either sort is in many respects analogous to the question of the reality of language structure briefly mentioned above in Section 3 of the first article in this series.\textsuperscript{18} The formal nature of reasoning was clearly presupposed if not explicitly recognized even by Aristotle. This is shown in my as yet unpublished article "A Mathematical Model of Aristotle's Syllogistic." It was explicitly recognized probably as early as 1851 by Boole (pp. 235ff). Hilbert's remarks quoted by Reid (pp. 57ff) show that he also was well aware of this fact very early in his career. However, despite the long history of this idea and despite widely published warnings by prominent mathematicians concerning misconstruals (e.g., Poincaré, pp. 5ff) it has nevertheless been taken to imply that reasoning itself consists in a mindless application of computational techniques. The important point to realize in connection with present purposes is that, although subject matter or content is irrelevant to soundness of reasoning in the sense that sound reasoning about one subject when reinterpreted correctly is equally sound when applied to another, it is still the case that reasoning divorced from all subject matter rarely, if ever, occurs in practice. Even Hilbert's heralded formal treatment of geometry was, by Hilbert's own admission (p. 3), a codification of the fundamental facts of our spatial intuition. Indeed, were Hilbert's proofs not understood in this way they would scarcely be understandable.
is also a proof. This implies that each production-type rule of inference has the effect of lengthening an already existent proof.

Since proofs frequently begin with assumptions laid down without proof, we may suppose that one initial string rule says that any finite list of sentences may be written down to start a proof provided that each such sentence is clearly marked as an assumption. Thus we might state the premise rule as follows: any finite list of sentences of the form 'Assume p' (for p a sentence) is a proof. Examples of production-type rules of inference are easy to think of. The rule of detachment (or modus ponens) can be stated: Any proof containing both p and 'if p then q' may be lengthened by adding q onto the end. Many other rules will come to mind.

Knowledge of Rules of Inference. It is important to distinguish a stronger and a weaker sense in which one may know a rule of inference. Let us say that a person has weak knowledge of a rule of inference if he reasons in accord with that rule. Thus weak knowledge of a rule of inference is a non-self-conscious kind of knowledge. All mathematicians and most people, I imagine, have weak knowledge of quite a few rules of inference although few people are self-conscious about the rules according to which they reason. On the other hand, let us say that a person has strong knowledge of a rule of inference if he can explain the details of the rule, point-out places where it is used, etc. Strong knowledge of a rule of inference is a very self-conscious kind of knowledge. Mathematicians generally have weak knowledge of many rules of inference and strong knowledge of very few. A logician who is poor at reasoning may have strong knowledge of many rules of inference and weak knowledge of very few, although most logicians, it seems, have weak knowledge and strong knowledge of many rules of inference.

The same distinction carries over to knowledge of rules of sentence construction. All speakers of English have weak knowledge of many sentential rules whereas only linguists can be expected to have strong knowledge of more than a few such rules. Linguists make it their business to have strong knowledge of rules of sentence construction whereas other speakers are content to be able to use the rules, i.e., to have weak knowledge of the rules.

Naturally, it is not to be expected that everyone has even weak knowledge of all rules of inference. Certainly the high school freshman could not be expected to know all of the rules of inference used by the professional mathematician. In a sense, knowing a rule of inference involves an understanding of a type of logical connection. Of course, as people acquainted with mathematical education, we have all had the discouraging experience of seeing a student mimic a teacher's pattern of reasoning without understanding it. In such cases, I believe, we will always be able to ascertain that the student has not learned the rule, but only the superficial aspects of a few applications of it. Nevertheless, I must acknowledge the theoretical possibility of a student who knows how to use an impressively large class of rules without understanding any of them. Such a student could verify a correct proof of a conclusion from some assumptions

---

19 For purely heuristic reasons we are using the term "proof" in such a way that a partial proof (initial segment) is counted as a proof. Thus, a finished proof will be a "proof" which satisfies certain additional conditions. This issue will be dealt with in the third article.
without believing that the conclusion actually followed from them—i.e., he would not be willing to risk anything to defend the thesis that, if the assumptions were true, then the conclusion would necessarily also be true. (Cf. fn. 20 below.)

Even though a given person may not know all of the rules of inference (as the skills of mathematical reasoning evolve new rules may come into use), it is most likely the case that most normal high school freshmen know several of the simpler rules. Moreover, it is my view that some of the more complex rules are learned by developing skill in the use of the simpler rules and, then, seeing how steps may be skipped. This is certainly not to suggest the obviously wrong conclusion that "quantum jumps" do not occur. For example, it was probably not until the late 19th century that mathematicians began using the choice rule \[\text{infer } (\exists x)Rxf(x) \text{ from } (x)(Ey)Rxy\] and it is difficult to see how this rule could be broken down into a deduction using significantly simpler rules. Indeed, "quantum jumps" must have occurred—otherwise we would have no rules at all.

The opinion concerning acquisition of knowledge of some of the more complex rules means that after a student has gone through a certain fixed pattern of detailed reasoning several times he may develop a feel for the upshot of the pattern and begin to omit the details in future proofs—thus, in effect, gaining weak knowledge of a more complex rule. We may imagine that the professional mathematician, after years of experience in deductive reasoning, has developed weak knowledge of very complex rules well beyond the comprehension of beginning students. From this point of view, it is natural to expect that as mathematical reasoning becomes increasingly sophisticated, more and more complex rules of inference will evolve.

If we wish we may even speculate that the mathematics student has two kinds of "vocabularies" of rules—an active vocabulary that he can actually use in doing proofs and a passive vocabulary of rules which he can "follow" but not use. This sort of hypothesis may partially account for inability of students to recreate reasoning that they have followed in class.

Correctness of Rules of Inference. We may wonder about correctness and incorrectness of rules of inference—is it conceivable that a small group of persons or even a whole society writes proofs according to incorrect rules? Indeed, suppose that everyone wrote proofs according to a certain rule, would not the universal acceptance of a rule make it correct? On a certain level, these are very easy questions once we recall that a proof is designed to show that a certain conclusion follows from certain premises. If a conclusion follows from some premises then it is impossible that the premises are true and the conclusion false. Thus if a system of rules could be used to prove a false sentence from a set of true sentences then certainly at least one of the rules is incorrect or, in the terminology of logic, unsound. Thus, it is possible that a small group or even a whole society writes proofs according to incorrect rules. (It is possible but I have never seen it happen—although I have seen people make mistakes in proofs.) Moreover, concerning this second question we can say that the universal acceptance of a rule of inference would not make it sound.

---

20It is instructive as well as amusing to imagine a "country" in which the system of reasoning devised by Copi (1954) were adopted as
Incidentally, it follows from what has been said above that if a certain society writes proofs incorrectly then possibly someone could discover that fact—however, if a society writes proofs correctly then there seems to be no way of finding out for sure that it does.

Parenthetically, I might add here that if I were an Intuitionist, I would have said that I had seen examples of the use of unsound rules. The Intuitionist, e.g., Heyting (1956), would say that most mathematicians use unsound rules and that much of the literature of mathematics contains incorrect proofs. In particular, Intuitionists regard one of the forms of indirect proof as unsound. Let us consider this in a little more detail. The kind of indirect (or reductio ad absurdum) reasoning involved in the standard proof of the irrationality of $\sqrt{2}$ from the axioms of arithmetic proceeds, after the (tacit) assumption of the axioms, by assuming that $\sqrt{2} = n/m$ for some integers $n$ and $m$ and deducing a contradiction. This sort of reasoning is regarded as sound by the Intuitionists because what the Intuitionist means by "not $p$" is that the assumption of $p$ leads to a contradiction. However the Intuitionist does not regard as sound the other reductio rule which allows one to prove $p$ from some assumptions by assuming "not - $p$" and deriving a contradiction. For him this would only prove "not-not-$p$" from original assumptions. "Not-not-$p$" means that it is absurd to assume that $p$ is absurd and, for the Intuitionist, this does not in turn mean that $p$ itself is true. This view leads to the rejection of one rule of double negation (any proof containing "not-not-$p$" may be lengthened by adding $p$) and to the rejection of the rule of excluded middle (any proof may be lengthened by adding "$p$ or not-$p$").

2. THEORIES OF PROOF

By a theory of proof for English, say, I mean a discourse grammar (1) which is intended to describe some or all of the proofs expressible in English and (2) whose rules are intended to be rules of inference known by persons who express their reasoning in English. If we are given such a theory, we may want to inquire concerning its correctness and its comprehensiveness. It would be natural to call it correct if each of its rules were used by some speakers of English. (There are, of course, other possibilities but this one will suffice in this context.) Furthermore, it would be natural to call it comprehensive if every rule used by any speaker of English were included among its rules. Of course, the correctness and the comprehensiveness of a given theory of proof would be relative to a given time in order to leave open both the possibility of "old" rules being abandoned and also the possibility of "new" rules being "devised."

The hope of ever getting a correct and comprehensive theory of proof is dim. But it is certainly possible to contribute toward such a theory. This would be done first by considering one's own reasoning and trying to formulate the rules implicit therein. The next step would be to survey the mathematical literature in an attempt to find correct proofs that are not constructible by means of one's own rules and which, therefore, may be presumed to be constructed according to "new" rules. After some of these were formulated the continuation of the project would involve

"official reasoning." Parry (1965) has discovered several invalid "arguments" whose respective conclusions are deducible from their respective premise sets in Copi's system.

21Cauman (1966) gives an interesting discussion of this proof.
getting other workers to formulate their own rules and to help in the survey of the literature. It is hard to imagine how one could ever determine whether a particular theory were comprehensive and, of course, if a theory were comprehensive relative to a fixed time it may very well not be comprehensive relative to a later time.

To many readers, the above will sound at least utopian if not far-fetched. It may very well be utopian but, given the Chomsky-Harris idea of trying to develop a sentential grammar of English, the above can easily be seen as an application of the same core idea to a part of the totality of English discourses. Thus, the idea of a comprehensive discourse grammar for all of English is even more utopian. Now, as for being far-fetched, I would simply reply that it is no more far-fetched than the ideal of a comprehensive sentential grammar of English, and a considerable body of researchers are developing this today.

As soon as one seriously considers the project of working toward a correct and comprehensive theory of proof in English, he is quickly faced with a crucial consideration. Since a discourse grammar takes as a starting point a sentential grammar, and since a sentential grammar for English does not exist in anything like a complete form, it becomes clear that the project cannot be begun in a systematic fashion. This objection is well-taken but fortunately a reasonable substitute for a sentential grammar is available at least for the part of English used in mathematical proofs. As a result of centuries of logical analysis of mathematical discourse we now have formally defined symbolic languages which are sufficiently rich so that all of mathematical discourse can be symbolically stated. Thus, we may choose a formal language into which to translate proofs and use the grammar of this formal language as the sentential grammar needed for the theory of proof. Taking this path our resultant theory of proof will necessarily be an idealization of an actual theory of proof in the same sense that, say, a formal language for arithmetic is an idealization of the part of English used in discourse about arithmetic. If it so happened that a group of mathematicians actually used a formal language in their investigations and they wrote their proofs in the formal language then we could investigate the body of proofs as such without translating and without regarding ourselves as developing an idealization. (Cf. Church (1956), pp. 2, 3, 47, fn. 108).

22Current symbolic languages can express all mathematical statements only in the sense that to each mathematical statement there corresponds a symbolic sentence having the same truth conditions. This is not to say that for every mathematical statement there corresponds an equivalent symbolic sentence which makes the same statement in the same way. For example, "No even number is odd" would be glossed as '¬ x(Ex&0x)' because in none of the current languages do we find a "nothing quantifier." Moreover, the phrase "a, b, and c are distinct objects," which occurs repeatedly in mathematics, must be glossed in current languages by a tortured construction involving a conjunction of three inequalities. Problems of this sort, once noticed, are easily solved. Indeed, Lewis and Langford (pp. 306ff, 385ff) have solved the above two problems. However, all such problems must be solved before a comprehensive theory of proof can be constructed. The reason is that the variety of regular reasoning possible in a language depends on the linguistic devices available.
Moreover, the use of the symbolic language may in the end be seen as a distinct advantage as it may enable the theory to transcend English and provide a theory of proof for other languages as well. However, one should not overlook the possibility that the idiosyncrasies of the various languages will also make themselves known on the discourse level and, in particular, in the proofs expressible in the various languages. This is not to suggest that a conclusion may be provable from certain premises in one language but not in another, though this may be true. Our suggestion was that even if exactly the "same" conclusions are provable from the "same" premises in two different languages it may turn out that there are means of doing it in one language not available to the other. Both of these hypotheses are likely—and perhaps interesting to investigate.

3. THE VALUE OF A THEORY OF PROOF

Before we can consider the possible value of a theory of proof, we should try to determine specifications for a theory which could actually be developed. Otherwise, our speculations would be too hypothetical to be very interesting.

In the first place we postulate the existence of a managably small set of simple rules of inference which must be known in order, for example, to be able to prove the main theorems of plane geometry and arithmetic. It is immaterial whether these rules, which we will call the basic rules, are redundant. [A set of, say, three rules is redundant if everything that can be proved using all three can also be proved using only two.] We can easily imagine that the basic rules can be discovered. It is my opinion that the basic rules could be discovered and formulated within a short time by several logicians working with several high school mathematics teachers—provided that the mathematics teachers (1) had been in the habit of making up new proofs and encouraging their students to make up new proofs and (2) had been developing geometry in different ways from year to year. In other words, the mathematics teachers working on the project must have some wide experience to refer to in these matters. What I have in mind as a model is the situation wherein several linguists work with several native informants in developing a sentential grammar of an exotic language.

In order to discuss the value (utility) of a theory of proof then let us imagine that we have the basic rules neatly formulated. Now, when we are asking about the value of this theory of proof what we are really concerned with is the possible answers to the following question: how could a mathematical educator use this theory to improve mathematical education?

23 Instead of regarding symbolic languages as idealizations of natural languages some linguists and logicians prefer to distinguish "the logical form" of a sentence from its "grammatical form" and to regard symbolization of a sentence as an attempt to express its logical form. From this point of view a discourse grammar based on a symbolic language would generate the logical forms of discourses (or discourse deep structures). Grammatical forms or surface structures of sentences and discourses are thought of as obtained from their logical forms or deep structures by means of encoding functions called transformations (cf. Keenan, 1969).
A theory of proof which included the basic rules would provide strong (self-conscious) knowledge of the rules of inference commonly used in elementary mathematics. It seems to me that there are four areas within mathematical education in which such knowledge would be of use, viz., in teaching, in testing and guidance counseling, in curriculum design, and in attempts to understand the psychology of mathematical learning.

Teaching. One important part of a mathematical education is learning to reason deductively and developing skill at it. There may be much more to learning to reason than merely acquiring knowledge and skill in the use of the rules—but certainly these are part of it. Imagine a teacher who has knowledge of the rules in both the weak and the strong senses, i.e., he not only knew how to use them, but he also could refer to them explicitly, formulate them, etc. Such a teacher would be in a very advantageous position vis-a-vis trying to teach mathematical reasoning. Firstly, he would be better able to detect ignorance of specific rules. Now, when a teacher sees a student having difficulty with a proof he is left to his own ad hoc devices concerning diagnosis of the difficulty. Secondly, he would be able to be much more clear in his own writing of proofs because he could be self-consciously critical of his own proofs. Thirdly, he would have a guide in choosing exercises and examples. When the class is having difficulty seeing a proof which involved a complicated application of a rule, the teacher would be able to choose another theorem which involves a simpler application of the same rule, and then, in presenting it to the class he could point out that the reasoning in the complicated case is similar to the reasoning in the simple case. All three of these points hinge on the advantage that an articulate teacher has over one who is merely expert in the subject matter. Consider, for example, the excellent tennis player who is not articulate about what is involved in playing tennis. In trying to teach a beginner to play tennis, the expert player is reduced to showing. If he sees the student doing something wrong he cannot say exactly what is wrong. Even in showing the student what the motions are like, the teacher will not know what to exaggerate and he will not be able to distinguish his own idiosyncrasies from what is essential about tennis. Finally, he will be poor at developing drills, etc.

Testing and Guidance Counseling. It seems to me that a student's ability in deductive reasoning is an important index of his mathematical aptitude, his ability to learn mathematics. This means that a student who is skilled in understanding and producing mathematical proofs will be much more likely to benefit from mathematics courses than one who does not have such skills. It is obvious that a man who has a characterization of what he wants to test is in a better position to design a test than a man who does not have such a characterization. A theory of proof is a characterization of the abstract structure underlying reasoning ability and it should provide a very useful framework for designing tests of reasoning ability. At the very least, a theory of proof would provide a better knowledge of what is being measured in tests of reasoning ability and, therefore, also in mathematical aptitude tests.

In order to get an idea of how such tests may be helpful in guidance counseling we must speculate concerning the kinds of things that might be discovered by use of the tests. For example, one might be able to show experimentally that unless a student had acquired weak knowledge of the basic rules by a certain age the chances of his ever being competent in mathematics are very slim. This would enable counselors to advise students concerning careers in mathematics and related areas. Moreover, it
is not unreasonable to suppose that normal mathematical development could be characterized in terms of the number and kind of rules learned at various ages (or at various testable stages). This would permit objective identification of unusually able and unusually backward students, again leading to more efficient and more scientific counseling. The professional mathematical educator can certainly conceive of other applications in this vein.

Curriculum Design. One of the aims of curriculum design is to trace a sequence of topics in mathematics which parallels the optimal development of the student's interests and abilities. The reason for this is the desire to give the student the maximum benefit from his formal educational experience. The idea is that the student is best educated by presenting to him at each stage in his education those concepts and proofs which he is best able to respond to. It is absurd either to present things which are too trivial or to present things that are beyond the student's ability. It seems to me then that a characterization of the development of mathematical skill in terms of the number and kind of rules acquired at various ages would provide a valuable framework for use in the design of an efficient curriculum. It would at least permit the knowledge of what would be very difficult and what would be very easy, as far as reasoning is concerned, and this, in turn, would permit more rational choices among alternative theorems to be presented or between alternative developments of a particular topic.

In addition, one can easily imagine a battery of specific remedial programs each designed to teach a specific rule or cluster of rules. Such remedial programs used in conjunction with the diagnostic tests mentioned above might very well form a formidable weapon in trying to overcome inadequate preparation.

In the discussion of knowledge of rules of inference we suggested that complex rules are sometimes learned through experience with simpler ones. If this turns out to be true then the details of the interrelation of knowledge of complex and simple rules will be very important in the choice of alternative developments of a subject as well as in the design of drills and so on.

Finally, we return to the hypothesis of active and passive vocabularies of rules. The truth of this hypothesis would lend additional justification to the suggestions of Professor J. J. LeTourneau (personal communication) to the effect that there should be two separate but parallel mathematics programs—one aimed at developing skill and concrete experience in creating theorems and proofs, the other aimed at acquainting the student with the body of existent mathematical knowledge. Naturally, a theory of the active vocabulary would be applied in the former, whereas the latter would use the passive theory.

Psychology. It is already clear enough that a theory of proof would provide a fruitful source of ideas for hypotheses and experiments in the psychology of mathematical learning. Moreover, one might wish to consider a more comprehensive theory of proof as an idealized description of the more-or-less behavioral aspects of the psychological processes of reasoning. We have already pointed out that the written (or spoken) proof is our only access to another person's reasoning processes. The written proof is a permanent record of the reasoning and, moreover, it is a "trace" of the behavioral aspect of the reasoning. The rules of inference in accordance with which the proofs are written are thus more-
or-less behavioral "norms." Given all this, it is easy to speculate that a theory of proof could lead to a psychological theory of deductive reasoning--perhaps analogous to the way that Kepler's Laws describing the orbits of planets lead to a kinetic theory explaining the orbits in terms of the effects of forces.

Finally, on the subject of applications of a theory of proof, I would like to suggest that the quality of writing of mathematics texts could be greatly improved if the writers would take the trouble to learn the rules of inference used by their prospective audiences. A mature mathematician must learn how to reason in a fashion understandable to a freshman if he wants freshmen to learn the mathematics (and not just memorize). Frequently, the mature mathematician encounters (in teaching) theorems which he sees "immediately" and he finds himself at a loss as to what to say to prove them. If he knew the rules of inference used by his class then he would know exactly what to say. If mathematics texts (and mathematics teaching) are improved in this way then one can expect that capable but non-genius students will be more able both to appreciate the beauty of mathematics and also to keep from "getting turned-off by the chicken scratching." Quite possibly all this could lead to the kind of improvement in the field of mathematics that we have seen after the rediscovery of the axiomatic method. In the axiomatic method we find the ideal of the deductive/definitional organization of branches of mathematics: a theory of proof provides a partial answer to the question of what deduction is.

Following all of these hopeful speculations I want to emphasize two negative points. In the first place, none of the above applications will be easily or mechanically achieved despite the fact that much of the groundwork is done. A tremendous amount of very detailed creative thought, dialogue and experimentation is needed. There is even cause to wonder whether there is a natural place to begin. And, there are pitfalls, one of which is the gap between the precision and simplicity of the symbolic languages, on the one hand, and the vagueness, ambiguity and complexity of natural language on the other. Anyone seriously desiring to pursue any of the above applications must become extremely sensitive to the nuances of normal English--and very few mathematicians have the patience for this. A pilot experiment in deductive reasoning recently conducted in a Philadelphia school ended distressingly because the subjects were diverted by too many linguistic red herrings in the test questions. Something can be perfectly clear in the symbolic language and perfectly confusing when translated mechanically into English.

Paradoxically, the second negative point issues from the exhilarating feeling of power and self-confidence that a mathematically competent person derives from learning to be articulate about what he is good at, i.e., from learning a clearly presented and apparently comprehensive theory of proof. Such a person naturally wants to teach the theory to his students--but if the students are not yet good at reasoning they cannot appreciate the significance of what they are learning. They may learn the rules and they may learn how to follow the rules. The disaster is that they come to believe that mathematical reasoning is nothing but following rules. As we pointed out in the beginning of this article, if a person has his mind occupied with the rules then the chances are slim that he will have any attention left for the subject matter or for the deeper parts of reasoning. If a person learns the rules as external rules (as prescriptions) and not as descriptions of what he already does (or would do naturally), the result is stultifying. If pressure is
put on a student to accept a rule self-consciously before he knows the rule non-self-consciously (i.e., if a rule is imposed on a student), he will either rebel or lose his intellectual integrity, or adopt the view that it's all a silly game. Another equally undesirable but less disastrous effect of teaching an uncomprehensive theory of proof even to students who can appreciate it derives from the fact that they may reason according to rules not in the theory. In this case, the students will tend not to use the rules absent from this theory thus weakening their powers of reasoning. The upshot is that they will be poorer at reasoning after learning the theory than they were before learning it.\textsuperscript{24}

\textsuperscript{24}Dr. Albert Hammond, late professor of philosophy at Johns Hopkins University, reported to the author in a personal communication the results of tests administered to logic students before and after his course. The tests involved making elementary inferences from material presented in the form of imaginary newspaper articles and narrations of fictional events. His report was to the effect that almost every subject was significantly worse at elementary reasoning after the course.
TWO THEORIES OF PROOF

JOHN CORCORAN

There was, until very lately, a special difficulty in the principles of mathematics. It seemed plain that mathematics consists of deductions, and yet the orthodox accounts of deduction were largely or wholly inapplicable to existing mathematics. Not only the Aristotelian syllogistic theory, but also the modern doctrines of Symbolic Logic, were either theoretically inadequate to mathematical reasoning, or at any rate required such artificial forms of statement that they could not be practically applied. —Russell

This part of the series has a dual purpose. In the first place we will discuss two kinds of theories of proof. The first kind will be called a theory of linear proof. The second has been called a theory of suppositional proof. The term "natural deduction" has often and correctly been used to refer to the second kind of theory, but I shall not do so here because many of the theories so-called are not of the second kind—they must be thought of either as disguised linear theories or theories of a third kind (see postscript below). The second purpose of this part is to develop some of the main ideas needed in constructing a comprehensive theory of proof. The reason for choosing the linear and suppositional theories for this purpose is because the linear theory includes only rules of a very simple nature, and the suppositional theory can be seen as the result of making the linear theory more comprehensive.

1. THEORIES OF LINEAR PROOF

A theory of linear proof is a theory of proof which holds that proofs have a certain simple structure which can be metaphorically called linear.\(^\text{25}\) As will be obvious shortly, such theories can be quite plausible a priori but, of course, the comprehensiveness of a theory of proof is an empirical

\(^\text{25}\) See par. 2 of the second article in this series.

\(^\text{26}\) A system of logic need not contain a theory of proof. There are other purposes for constructing such systems. For example, a logician may be concerned to codify consequences of sets of premises without even considering the problem of describing proofs per se. A system designed for such a purpose was called a consequence system in Corcoran (1969). Many consequence systems are systems of formal deductions which would be theories of linear proof were they put forth as theories of proof—which they usually are not.
matter. As usual, what is plausible a priori turns out to be a gross oversimplification of reality.

Because theories of linear proof are simple and (often) plausible it is perhaps remarkable that the first theory of proof was actually suppositional in nature (cf. my "A Mathematical Model of Aristotle's Logic"). However, theories of linear proof are quite old, tracing their common history at least as far back as Boole. Indeed, Boole (pp. 142, 143) was so clear about his own theory that his description of it can still serve as a concise introduction to the general topic.

All demonstration essentially consists of the deduction of conclusions from premises, a conclusion once deduced being itself admissible as a premise. And it is in this order that reasoning usually proceeds. Certain premises are laid down, either from experience or from testimony, or from some other extralogical source; from these are deduced conclusions which simply or combined with other premises derived from the same class of sources as those first given, serve as bases for further inference, until the chain of argument is completed. At any stage of the process we may find ourselves dealing with two sorts of data, viz., such as have been deduced in the previous course of argument from given data, and such as have not before appeared. A very slight examination of any actual specimen of demonstrative reasoning will show that such are the materials of its composition and such the order of its progress.

In more modern terminology we can say that a linear proof of a conclusion c from a set P of premises is a sequence of lines, beginning with a list of all or some of the premises and such that each subsequent line is derived immediately from premises and/or previously proved lines and, finally, ending with c. In other words a linear proof of c from P is written linearly in a column, say, beginning with the premises P at the top and proceeding step-by-step through intermediate conclusions all derived from P and ending with c the final conclusion. This is the idea; in practice things are a little more complicated, but the following general statement always holds—in a linear proof from premises P to conclusion c each sentence in the proof is a logical consequence of P. (The reader should note that the concept of logical consequence as defined above is not relative to any system of proof.)

There are three minor modifications to be made to the above loose account of linear proofs. The first is that for clarity the premises shall be marked as such to make it clear that they are not asserted to follow from any sentences which they may happen to follow. The second is that in some systems premises may be written at any place in the proof, not just at the top. Finally, in addition to assumptions and inferences,
Two Theories of Proof

properly so-called, all linear systems of proof permit the writing of
so-called logical axioms at any point in a proof. For example, in writ-
ing proofs in algebra we often have occasion to write logical identities,
t = t, in proofs. Corresponding to this we would have a logical axiom
rule which permits any proof to be lengthened by addition of a logical
identity. For another example, in setting up a "proof by cases" we often
write in proof lines of the form 'p or not-p' where p is a sentential
formula. Corresponding to this we would have a logical axiom rule which
permits any proof to be lengthened by addition of "excluded middle for-
mulas." These two are probably the most prominent logical axioms rules.28

Below we will mark premises with a plus sign. Thus '+p' will be read
"assume p as a premise" or simply "assume p."

As our example of a theory of linear proof we will give a (non-comprehen-
sive) theory of the proofs found in the abstract algebra of equations—
the so-called equational algebras wherein all sentences are either equa-
tions or universal generalizations of equations. Following the statement
of the rules we will give a proof of the theorem \( (x)(x=x^{-1}) \) [every ele-
ment is identical to the inverse of its own inverse] from the group axioms.

Rule Set A

Initial String Rule (Kernel Rule)

(1) Premises: A finite sequence of sentences each affixed by + is
a proof.29

Production Rules

(2) Identity Law: any proof may be lengthened by addition of any
logical identity, \( (t = t) \) where \( t \) is a constant term.

(3) Substitution of "Equals": any proof containing \( (t = s) \) and
also \( p \) may be lengthened by adding \( p' \) where \( p' \) is the result
of replacing occurrences of \( t \) in \( p \) by \( s \) and/or vice versa.

(4) Instances: any proof containing \( (v)p(v) \) may be lengthened by
adding \( p(t) \)--where \( v \) is a variable and \( t \) is a term composed of
constants.

(5) Generalizations: any proof containing \( p(d) \), \( d \) a dummy30 con-
stant, may be lengthened by addition of \( (v)p(v) \) provided that

fact that in actual proofs where premises are made explicit at all they
are put down at the beginning. If this is so then any theory which fails
to take account of it is, strictly speaking, incorrect regardless of how
valid it may be on other grounds.

A one way of characterizing the difference between normal reasoning
and the so-called Hilbert-type systems of deduction is to note that in
the former there are few logical axioms rules but many inference rules
whereas in the latter there are commonly many logical axioms rules but
few (usually one or two) inference rules. (Cf. Thomason, chapters III,

Note that for purely heuristic reasons we have tacitly been using
the term "proof" in such a way that a partial proof is counted as a proof,
thus a finished proof will be a "proof" which satisfies some additional
conditions. This issue will be discussed in more detail below. See es-
specially, the discussion of "developments" of axiomatic theories in Sec-
tion 5.
no assumptions concern \( d \) (i.e., provided \( d \) is "arbitrary").

(6) Repetition: any proof may be lengthened by repeating any previous line dropping a '4' if it occurs.

Obviously, each of the above rules corresponds exactly to a rule commonly used in proofs in algebra. Notice however that there are commonly used rules which do not appear in the list. For example, the only way of instantiating here is by rule 4 and this permits the elimination of quantifiers only one per application. This will be an annoying deficiency. Similarly for generalizations. Another deficiency is that substitutions can be done using only one equation at a time. In the proof below we have starred the lines that would remain were the deficiencies eliminated.

\[
\begin{align*}
+ (x)(y)(z)(x.(y.z) & = (x.y).z) \\
+ (x)(x.l & = x) \\
+ (x)(l.x & = x) \\
+ (x)((x.x^{-1}) & = 1) \\
+ (x)((x^{-1}.x) & = 1) \\
(y)(z)(a.(y.z) & = (a.y).z) \\
(z)(a.(a^{-1}.z) & = (a.a^{-1}).z) \\
(a.(a^{-1}.a^{-1}-1) & = (a.a^{-1}).a^{-1}-1) \\
(a^{-1}.a^{-1}-1) & = 1 \\
a.1 & = (a.a^{-1}).a^{-1}-1 \\
a.a^{-1} & = 1 \\
a.1 & = 1.a^{-1}-1 \\
a.1 & = a \\
a & = 1.a^{-1}-1 \\
1.a^{-1}-1 & = a^{-1}-1 \\
a & = a^{-1}-1 \\
(x)(x & = x^{-1}-1) \\
\end{align*}
\]

Having a more powerful instantiating rule would permit going from the associative law directly to the first unquantified line—skipping two lines. The other two unstarred lines would be skipped by doing two substitutions at a time.

Incidentally, the above rule set (or discourse grammar) describes proofs—

---

Use of the term "dummy" is redundant here; a dummy constant is simply one which does not occur in the premises. Usually the constants
but it does not make explicit what "a proof of \( c \) from \( P \)" is. Naturally, we define a proof to be a proof of \( c \) from \( P \) if \( c \) is the last line of the proof and all premises in the proof are in \( P \). The above example is a proof of \( (x)(x = x^{-1}l) \) from the group axioms.

As the rule set is being used here, the (metalinguistic) symbols \( p, p(t), p(v), \) and \( p(d) \) refer to formulas in the language of groups. Thus this set of rules presupposed a sentential grammar for the language of groups. However, if we interpreted the symbols as referring to formulas in the arithmetic language, then we could use Rule Set A for the theory of proof needed to complete the Partial Grammar of the Arithmetic Language given at the end of the first article. This would actually be a bit silly for two reasons: first, the Partial Grammar has no quantifiers so rules 6 and 7 would never apply; second, the Partial Grammar does have the logical connectives whereas none of the rules permit any inferences involving connectives. The point, therefore, is not that the Partial Grammar would be finished but rather that the reader can now see what a finished grammar would be like. The respective natures of an alphabet, a rule set for words, a rule set for phrases, and a rule set for sentences are already clear from the Partial Grammar. Now we have also seen a discourse grammar which describes or produces a certain set of proofs. This discourse grammar, Rule Set A, is a theory of proof.

Rule Set A is obviously a correct theory of proof--each of its rules corresponds exactly to (or is) an actual rule of inference that we have all used when doing proofs in elementary group theory. Rule Set A is obviously not comprehensive in the sense that I have defined the term because, e.g., it lacks the complex rules alluded to above which permit the unstarred lines to be omitted. However, it is complete in a certain sense.

occurring in premises are given special symbolization: '0,' '1,' 'n,' 'e,' etc.; whereas dummies are indicated by 'a,' 'b,' 'c,' 'd,' or by variables subscripted with a '0,' e.g., \( x_0 \). Incidentally, Thomason (chapter IX, esp. p. 183) does not class his rule of generalization with the immediate inference rules. His rule of generalization is sound but, in my opinion, it does not correspond to actual reasoning as closely as does the present rule.

The possibility of obtaining a correct theory of (symbolic) proof depends on having a "correct" symbolic sentential grammar to begin with. Indeed, finding "natural reasoning" blocked by restrictions dictated by peculiarities of the sentential grammar can indicate need for revision of the latter. For example, in the otherwise correct theory of symbolic proof given by Resnik (1970), every proof of \( F \) from \( (x)(y)Fxy \) involves getting a generalization of \( F \) as an intermediate step because of the need to avoid "capturing." Similar situations are common. However, it is possible to design the symbolic language in such a way as to make "capturing" grammatically impossible. This makes it unnecessary to add special restrictions on the rules. Once the symbolic language is thus revised, as in Lemmon (1965), as an unexpected advantage one finds that intrinsically awkward symbolic sentences are eliminated without loss of expressive power.

A theory of proof for a particular language is called equationally complete when the following holds: given any set of equational sentences (either equations properly so-called or universal generalizations thereof) and any single equational sentence \( c \), if \( c \) is a logical consequence of \( P \), then there is a proof of \( c \) from \( P \) constructible by the rules of the theory. Rule Set A is equationally complete. This fact will be plausible to any reader who understands it. To the other readers the following remarks are addressed. Let \( P \) be the axioms for groups. Let
In any theory of proof which describes or produces only linear proofs, it is possible to give a very simple description of all proofs from a particular set $P$ of premises to a particular conclusion, $c$. Given a definition of the logical axioms and the rules one can then say: a proof of $P$ from $c$ is a finite sequence of lines ending with $c$, each subsequent line of which either is an assumption in $P$ or is a logical axiom or is obtained from previous lines by a rule.

The underlined expression (or rather an even simpler version of it) has become a slogan and, sometimes, a battlecry. One eminent logician related to me that when he first heard this slogan presented he was struck by its simplicity and truth and was moved to say to himself, "By God, that is what proofs are!"

If one takes the slogan as a rough description of all proofs, then one is led (1) to distinguish three kinds of rules of inference and (2) to believe that all rules of inference must be of one of the three kinds. The first kind contains only the rule of assumption—essentially to the effect that an assumption may be written to start (or to lengthen) any proof provided that it is marked as an assumption. The second kind contains all logical axiom rules—to the effect that a logical axiom may be written to lengthen any proof. The third kind contains all immediate inference rules; rules which state that any proof containing one or two (or some fixed finite number of) sentences of certain specified forms may be lengthened by adding a sentence in another form.

2. IMMEDIATE RULES AND SUBSIDIARY PROOF RULES

It so happens that by surveying the proofs in the mathematical literature (or by looking at our own proofs) we find many rules that are not of any of the above three kinds. Indeed, if all rules were of the three above kinds then there would be no room in mathematical reasoning for making subsidiary assumptions. Much of the most elegant and enlightening reasoning in mathematics turns on the ability to imagine good subsidiary assumptions. Below are some examples. (1) In proving that the square root of two is not rational, we assume, in addition to the axioms of arithmetic, the subsidiary assumption that the square root of two is rational. (2) In proving the right cancellation law $[(x)(y)(z)((x.z = y.z) \Rightarrow x = y)]$ from the group axioms, we assume, in addition to the group axioms, that $a.d = b.d$ where $a$, $b$ and $d$ are arbitrarily chosen but fixed elements of the group. (3) Whenever we give proofs by cases after we have proved that there are two cases, say, we assume that the first case holds and then prove our theorem in that case, then we assume the second case and prove our theorem in that case—finally we conclude that the theorem holds in general.... In each of these three examples the proof involves making subsidiary assumptions, assumptions other than those from which the conclusion is shown to follow.

$c$ be any equational sentence written in the language of groups and which is true in all groups. $c$, then, is a logical consequence of $P$; since (1) a group is by definition any mathematical system in which the axioms of groups are true and (2) to say that $c$ is a logical consequence of $P$ is to say that $c$ is true in any mathematical system which makes all of the sentences in $P$ true. The above-mentioned completeness condition implies, then, that by using Rule Set A one can construct a proof starting with $P$ as assumptions (as in the example) and ending with $c$. In fact, such a proof can be gotten by lengthening the one given as a sample.
At some point in each of these examples an inference is made not from certain previous lines in a proof but rather from (or on the basis of) a certain part of the proof. In other words, there are rules which can be stated as follows: any proof containing a subsidiary proof of a certain form may be extended by adding \( p \). For example, in reductio reasoning we are following the rule: any proof containing a subsidiary proof beginning with \( p \) and containing a contradiction may be extended by adding \( \neg p \) (not-\( p \)).

A subsidiary proof begins with a subsidiary assumption, a "new" assumption made for purposes of reasoning. The subsidiary assumption is marked with a "beginning" corner bracket ' \( \Rightarrow \)'. Thus ' \( \Rightarrow \ p \)' may be read "for purposes of reasoning suppose \( p \)" or simply "suppose \( p \)." When the subsidiary reasoning is completed one adds a "closing" or "ending" corner bracket ' \( \Leftarrow \)' to the last line. Each time an ending bracket is added it is matched with the last beginning bracket not yet matched. The latter is always on the line containing the supposition which begins the subsidiary proof in question. Thus a subsidiary proof may be defined as a section of a proof enclosed in matching brackets. The details, if not already clear, will be so after considering a couple of examples.

Two paragraphs back we stated the reductio rule. We now give as an example an indirect (reductio) proof of \( \neg(x) \neg(x = x^{-1}) \) [not every element is different from its own inverse] from the group axioms.

\[
\begin{align*}
\Rightarrow (x) \neg(x = x^{-1}) \\
\neg(1 = 1^{-1}) \\
1.1^{-1} = 1 \\
1.1^{-1} = 1^{-1} \\
\neg(x) \neg(x = x^{-1})
\end{align*}
\]

The subsidiary proof is enclosed in matching brackets. The contradiction in question is "between" the starred lines. Notice that the conclusion is inferred to follow from the group axioms (not from all assumptions) on

The equational completeness of Rule Set A was proved several years ago by Jan Kalicki and Dana Scott (1955).

For a wide-ranging discussion of this particular proof in the general context of a concern with the history and the soundness of indirect reasoning see Cauman (1966).
the basis of the subsidiary proof. Once a subsidiary proof is marked off by an ending bracket (L), it must be regarded as an isolated, separate unit in the proof. In particular, one may no longer apply any of the immediate inference rules to lines inside of the subsidiary proof. For example, we could not write down as a next line ¬(1 = 1⁻¹) by repetition because this does not follow from only the group axioms.

Let us use the phrase 'subsidiary proof rule' to refer to rules which permit the lengthening of a proof on the basis of a subsidiary proof. Of course, the most notorious of subsidiary proof rules is the rule of conditionalization which permits inference of 'if p then q' on the basis of a subsidiary proof beginning with p and ending with q. We will give a proof of the right cancellation law from the group axioms to illustrate this. (In the proofs below we do not necessarily follow Rule Set A but use other commonly known rules as well.)

\[ + (x)(y)(z)((x.(y.z)) = ((x.y).z)) \]
\[ + (x)(x.1 = x) \]
\[ + (x)(1.x = x) \]
\[ + (x)(x.x⁻¹ = 1) \]
\[ + (x)(x⁻¹.x = 1) \]
\[ \{ \begin{align*}
  a.d &= b.d \\
  (a.d).d⁻¹ &= (b.d).d⁻¹ \\
  a.(d.d⁻¹) &= b.(d.d⁻¹) \\
  a &= b
\end{align*} \] \]

\[ (a.d = b.d) \Rightarrow (a = b) \]
\[ (x)(y)(z)((x.z = y.z) \Rightarrow x = y) \]

It will be valuable to notice that in proofs by cases more than one subsidiary proof is needed—one for each case. Actually, all proofs-by-cases rules are "combinations" of the two-case rule stated as follows: any proof containing 'c₁ or c₂', together with two subsidiary proofs, one beginning with c₁ the other beginning with c₂ both ending with c, can be extended by adding c. To illustrate this we will give a proof of the two-sided cancellation law. The proof will involve one application of the two-case rule inside of a subsidiary proof on which conditionalization is used.

\[ + (x)(y)(z)((x.(y.z)) = ((x.y).z)) \]
\[ + (x) (x.1 = x) \]
\[ + (x) (1.x = x) \]
\[ + (x) (x.x⁻¹ = 1) \]
\[ + (x) (x⁻¹.x = 1) \]
Two Theories of Proof

\[ (a \cdot d = b \cdot d) \lor (d \cdot a = d \cdot b) \]

\[ \Gamma \quad a \cdot d = b \cdot d \]

\[ (a \cdot d) \cdot d^{-1} = (b \cdot d) \cdot d^{-1} \]

\[ a \cdot (d \cdot d^{-1}) = b \cdot (d \cdot d^{-1}) \]

\[ a = b \]

\[ \Gamma \quad d \cdot a = d \cdot b \]

\[ d^{-1} \cdot (d \cdot a) = d^{-1} \cdot (d \cdot b) \]

\[ (d^{-1} \cdot d) \cdot a = (d^{-1} \cdot d) \cdot b \]

\[ a = b \]

\[ a = b \]

\[ ((a \cdot d = b \cdot d) \lor (d \cdot a = d \cdot b)) \Rightarrow a = b \]

\[ (x)(y)(z)(((x \cdot z = y \cdot z) \lor (z \cdot x = z \cdot y)) \Rightarrow x = y) \]

The notations on the right are designed to help the reader see exactly where and how the two subsidiary proof rules are applied (* and **).

Before we proceed to a discussion of theories of suppositional proof (theories involving subsidiary proof rules), the reader should note that the above three proofs are not linear because the subsidiary assumptions are not among the premises from which the proof proceeds and neither are they consequences of the premises. That is, for example, in the proof of the cancellation law from the group axioms there are sentences which are not logical consequences of the group axioms. Thus in these proofs we do not reason in a linear fashion—we take 'side trips'.

3. THEORIES OF SUPPOSITIONAL PROOF

The defining characteristic of a theory of suppositional proof is that the rules permit the use of subsidiary assumptions which are later "discharged" and are not among the assumptions from which the final conclusion is shown to follow. These rules are subsidiary proof rules which countenance an inference not from previous lines but rather on the basis of a subsidiary proof. Such rules are not unusual but rather they comprise the essence of clear, elegant mathematical reasoning. Indeed, I think the mathematically experienced reader will agree that linear proofs have a very computational flavor to them whereas suppositional proofs seem to embody more creative and enlightening reasoning.

There are a few questions concerning the formulation of suppositional rules which might have been annoying some readers. I will digress slightly at this point to take up some of them.

In the first place we must give an explicit rule for adding subsidiary
assumptions: **rule of supposition**—any proof can be lengthened by addition of a formula prefixed by a beginning bracket. Secondly the following rule explicitly accounts for introduction of closing brackets: **closing rule**—any proof containing more beginning brackets than ending brackets can be modified by affixing a closing bracket to its last line. The idea is that each supposition line, \( \text{rp} \), starts a subsidiary proof and that each subsidiary proof must start with the last supposition line not already a part of another subsidiary proof. Each time an ending bracket is put into a proof there is exactly one beginning bracket with which to match it—namely the last one not already matched.

A **subproof** of a proof is a sequence of lines beginning and ending with matched brackets. An occurrence of a sentence is **inactive** in a proof if it occurs within a subproof. An occurrence is active if not inactive. A sentence is active in a proof if it has an active occurrence therein. A subproof occurrence in a proof is inactive if it occurs within another subproof. An occurrence of a subproof is active if not inactive. A subproof is active in a proof if it has an active occurrence in the proof.

All rules must be stated so that they apply only to sentences and subproofs which are active.

A given proof is a proof of its last line (if active) from its set of active assumptions (premises plus active suppositions). Now we can state two important general principles for suppositional proofs.

Let \( p \) be a given line of a suppositional proof. (1) The sequence of lines up to and including \( p \) is itself a proof. Let us call this the **partial proof ending with** \( p \). (2) In any suppositional proof, each given line \( p \) is a logical consequence of the active assumptions of the partial proof ending with \( p \). (If \( p \) is prefixed by \( \text{ar} \), the \( \text{r} \) counts as in the partial proof—if by \( \text{l} \) the \( \text{l} \) does not count as in the subproof).

Now we can define a **finished proof** to be one which satisfies the following two conditions: first, it contains no active subsidiary assumptions; second, it ends with an active sentence. The first condition guarantees that any reasoning for purposes of which a supposition has been made is completed. The second condition allows a subsidiary proof rule to be applied after the last subsidiary proof has been completed. This definition includes every proof that one would want to count as finished and excludes most unfinished proofs, but it still counts as "finished" certain proofs which one would not wish to consider as such. A more adequate definition would involve intricacies undesirable in an article of this sort (but see below for an easy improvement).

It is obvious that the framework of a suppositional theory is much more adequate for characterizing mathematical proofs than is the framework of a linear theory—even though anything that can be proved in a given suppositional theory will also admit of proof in some linear theory. In other words, we are not contrasting the abstract power of such theories but rather their relative adequacies in characterizing the proofs which we actually write. Given the advantage of suppositional theories we can ask: *Are there other kinds of rules of proof which could be added and...*
which would constitute an even more adequate framework? Let us put this question another way. Besides the premises rule and the closing rule we have seen four kinds of rules of inference: (1) assumption rules, (2) logical axiom rules, (3) immediate inference rules and (4) subsidiary proof rules. Are there other kinds of rules which are actually employed in writing of proofs?

The most obvious kind of rule to suggest adding is a rule that permits the writing of "goals." Frequently when we are writing a proof, after some assumptions (premises and/or suppositions) have been entered, we indicate our goal by writing, for example, "we want to show p." This is actually a very handy device which helps convey the reasoning to be expressed in the proof. Since the purpose of proofs is to express reasoning we should certainly consider such a rule. We could state it: Any proof may be lengthened by adding ?p. The question mark in this context could be read "to prove," say. We would then have to define all occurrences of ?p as inactive because otherwise we would be applying immediate rules to what we were trying to prove—thus begging the question.

Now let us consider another important kind of rule. We have actually given an example of this kind of rule, but we did not classify it. Notice that all of the above kinds of rules apply only to a part of a proof to which they apply, i.e., it is usually unnecessary to look at each line in the whole proof in order to apply any of the above four kinds of rules—supposition does not require looking at any lines, the same for logical axiom rules, immediate inference rules involve only fixed finite numbers of lines, subsidiary proof rules involve perhaps a few subsidiary proofs plus perhaps a few active lines. The rule of generalization, however, requires looking at a particular line p(d) and then checking through the whole proof to determine that nothing has been assumed about d—i.e., that d is indeed arbitrary ('d' is dummy). Such rules we call global immediate rules. Thus, the classification of linear rules above was inadequate.

In addition there are subsidiary rules which involve reference to the entire proof to which they are applied. The most prominent example of a global subsidiary rule is the rule that is generally used in reasoning from an existentially quantified statement. For example, suppose that we have assumed the right cancellation law in a proof and we are aiming to prove (3x)(y)(y.x=x) = (x)(x=x -1). We assume the antecedent (g)(y)(y.x=x) and we say "let x0 be such an object." ("Let" is a sure sign of an assumption.) We are assuming that x0 is an "arbitrary object" satisfying the condition (y)(y.x0 = x0). We reason then of an (genuinely) arbitrary b that b.x0 = x0 and that b.b -1.x0 = x0. Then, using the cancellation law, infer b = b -1. Since b is arbitrary, (x)(x = x -1). Now we say: "Since x0 was arbitrary and (x)(x = x -1) does not depend on x0, the

36 This rule may profitably be compared with a similar device of Kalish and Montague (pp. 14ff) which involves writing 'show p' to indicate a goal and which requires the 'show' to be crossed out once "the goal has been reached." As useful and valid as this device surely is, it is not correct in our sense because it violates the principle that every sub-proof of a (partial) proof is itself a (partial) proof. The latter is a rough statement which corresponds to the apparent facts that we do not alter previously written (partial) proofs and that we read them "top to bottom" checking each line as encountered. The Kalish-Montague device may correspond better to a description of how proofs "emerge in thought" which, of course, is not our goal.
conclusion follows from the original assumption." This corresponds, in the below formalized version, to taking \((x)(x = x^{-1})\) out of the subsidiary proof and making it active [starred line].

\[+ (x)(y)(z)((x.z = y.z) \Rightarrow x = y)\]
\[?(x)(y)(y.x = x) \Rightarrow (x)(x = x^{-1})\]
\[\Gamma(x)(y)(y.x = x)\]
\[\Gamma(y)(y.x_0 = x_0) \quad "let x_0 be such an object"\]
\[b.x_0 = x_0\]
\[b^{-1}.x_0 = x_0\]
\[b.x_0 = b^{-1}.x_0\]
\[(b.x_0 = b^{-1}.x_0) \Rightarrow b = b^{-1}\quad (\text{cancellation law})\]
\[b = b^{-1}\]
\[\land (x)(x = x^{-1})\]
\[\land (x)(x = x^{-1}) \quad *\]
\[\forall x)(y)(y.x = x) \Rightarrow (x)(x = x^{-1})\]

It might be worthwhile to do another example using the above rule. We will prove \((y)((\exists x)(Dx \& Hyx) \Rightarrow (\exists z)(Az \& Hyz))\) from \((x)(Dx \Rightarrow Ax)\).

\[+ (x)(Dx \Rightarrow Ax)\]
\[\Gamma(x)(Dx \& Hbx)\]
\[\Gamma(Da \& Hba) \quad "let a be such an object"\]
\[Da\]
\[Da \Rightarrow Aa\]
\[Aa\]
\[Hba\]
\[Aa \& Hba\]
\[\land (z)(Az \& Hbz)\]
\[\land (z)(Az \& Hbz) \quad *\]
\[(\exists x)(Dx \& Hbx) \Rightarrow (\exists z)(Az \& Hbz)\]
\[(y)((\exists x)(Dx \& Hyx) \Rightarrow (\exists z)(Az \& Hyz))\]

The rule just exemplified could be called "existential instantiation" because it involves "instantiating" as existential statement to begin the
Two Theories of Proof

Often in writing a proof after a pair of contradictions have been proved (made active) we write 'a contradiction' and it is on the basis of that notation that we apply the reductio rule. Thus it is necessary (for comprehensiveness) to add a special symbol, say X, to the language of proofs. The rule of contradiction introduction is the following: any proof which contains active sentences p and not-p may be lengthened by addition of X. Given this we can now state two new reductio rules: any proof which ends in a subsidiary proof beginning p (respectively ~p) and ending with X can be lengthened by addition of ~p (respectively p).

The usual proof of Russell's theorem [no set contains exactly the sets not containing themselves] involves all three of the rules just mentioned together with the subsidiary proof rule of "existential instantiation." It should be mentioned that Russell's theorem is proved without the use of premises—it is proved using logic alone. For this reason it is often counted as a "law of logic"—indeed, its denial implies a contradiction.

\[ \neg(\forall x)(\exists y)(x \in y \equiv \neg(y \in y)) \]
\[ \Gamma(\forall x)(\exists y)(x \in y \equiv \neg(y \in y)) \]
\[ \Gamma(y)(x_0 \in y \equiv \neg(y \in y)) \quad \text{"let } x_0 \text{ be such an object"} \]
\[ x_0 \in x_0 \equiv \neg(x_0 \in x_0) \]
\[ \Gamma x_0 \in x_0 \]
\[ \neg x_0 \in x_0 \]
\[ L x \]
\[ \neg x_0 \in x_0 \]
\[ x_0 \in x_0 \]
\[ L x \quad \text{"but } x_0 \text{ was arbitrary"} \]
\[ \Gamma x \]
\[ \neg(\exists x)(\forall y)(x \in y \equiv \neg(y \in y)) \]

Because of limitations of space we merely mention a class of rules called definitional rules which actually form a subclass of the global subsidiary rules and which, as can be surmised from the name, countenance the use of nominal definitions within proofs.

As a final question we consider the nature of an axiomatic development of a mathematical theory. An axiomatic development of a theory begins with the axioms. Subsequently the first theorem is proved, then the

Linguistically, this may be a radical move. We are adding to the "sentences" used in discourses something that does not appear in the underlying language.
second, then the third, etc. However, after the first proof the axioms are not repeated. Moreover, in addition to the axioms, previously proved theorems are also used as new "axioms"—but these are generally not rewritten either. One way of characterizing such a development is to say that it is one long proof and that axioms and previously proved theorems can be used because they are already active above. There is something artificial about this characterization—we usually say that a development of a theory contains many proofs, here we say that it is just one long proof. It is obvious that there is a level above the level of proofs—a level containing "axiomatic developments" which, in a sense, are composed of proofs. This implies that in a development of a theory there is structure which is not reducible to the structure of proofs. Thus there are at least two levels of language above the sentential level.

4. SUMMARY OF SUPPOSITIONAL THEORIES

We have seen that linear theories contain four kinds of rules: premises, logical axiom, immediate inference, and global immediate inference. Next, we noticed that suppositional theories contain two additional kinds of rules: subsidiary proof rules and global subsidiary proof rules. It is important to realize that relative to linear systems both kinds of subsidiary rules are radical innovations because they countenance inferences not based on previously proved sentences but rather on the basis of previously performed patterns of reasoning. In addition, we pointed out that the definitional rules are merely a species of the global subsidiary proof rules.

We explained the concept of an active sentence in a proof and we asserted that the general principle behind suppositional proofs has two parts: (1) that given a proof and a sentence occurrence p in the proof, the part of the proof ending with p is also a proof (called the partial proof ending with p) and (2) each such p is a logical consequence of the active assumptions of the partial proof ending with p. Given this principle, the notation for subsidiary proofs, and the classification of the rules,

38 In a development of an axiomatic theory each theorem and each lemma is a "main goal" and within the course of deduction of a main goal one often chooses "intermediate goals" in order to focus on the local direction of the reasoning. Several things follow. The first is that one needs at least two "goal indicators," one for main goals and one for intermediate goals. One way of handling this is to use a single question mark to indicate a main goal, two question marks to indicate a conclusion to be reached in proving a main goal, (perhaps) three question marks to indicate a conclusion to be reached in proving a "level-two" goal, etc. The second is that the notion of a "finished proof" must be modified in order that a proof is counted as finished only if all of its goals have been reached in the required order. As each subsequent theorem or lemma has been reached the entire proof up to that point should be finished and it may be necessary to have a special symbol to indicate the end of a finished proof. Indeed many current authors use such symbols. Kelley (1955) uses a small shaded rectangle which he attributes to Halmos; Suppes (1960) uses the traditional 'Q.E.D.;' and Dean (1966) uses a triple asterisk. For further discussion of the structure of a development of an axiomatic theory see my "A Mathematical Model of Aristotle's Syllogistic."

39 For a more detailed discussion see my "Three Logical Theories."

40 See Section 3 above.
anyone having a background in mathematics is prepared to formulate his own theory of proof.  

5. SUMMARY OF THE SERIES

In the interest of accuracy we must admit that the obvious heuristic value of the notion of a partial proof probably refutes the hypothesis that the class of discourses has a kernel/transformations structure. The proof discourses are clearly the "finished proofs" and it does not seem to be the case that these have the requisite structure: one does not build up finished proofs by applying "natural" transformations to other finished proofs. Indeed, it seems to be generally the case with discourses that the beginning of a discourse is not itself a discourse but rather it seems that the beginning of a discourse makes "promises" which must be fulfilled later in order for the discourse to be "finished." When we put down some axioms and "a goal" (see above), that proof is not finished until the goal has been reached. Likewise with discourses, generally. For example, if someone were to say, "I have called this meeting to give you my views on the latest crisis," and then sat down, he would not have uttered a complete discourse. There are innumerable similar examples. The conclusion that the class of discourses fails to have a kernel/transformations structure seems inescapable.

In part I we discussed some fundamental concepts involved in the analysis of mathematical reasoning. In addition, we introduced the concept of levels of language and pointed out that a grammar of an entire language should be composed of several grammars, one at each level. We also made the point that a proof is a certain kind of discourse which, in turn, suggested the possibility of a theory of proof—a discourse grammar which describes the proofs of a language.

In part II we outlined what a theory of proof would be like. We noted that the grammatical rules used in describing proofs are the rules of inference according to which we write proofs. We discussed the nature of our knowledge of rules of inference distinguishing weak and strong varieties of such knowledge. Finally, we speculated concerning the utility of a theory of proof vis-a-vis improvements in mathematical education.

In the course of Part III, we contrasted what has become the traditional theory with a newer and more adequate theory whose essential features were discovered in the 1920's (Jaskowski). The older theory holds that mathematical reasoning proceeds from axioms step-by-step to conclusions.

\[\text{In mathematical logic one constructs a precisely defined mathematical analog (formal deductive system) of a system of proofs and a precise mathematical analog (formal semantic system) corresponding to the (actual or imagined) system of interpretations associated with the language. In this way the philosophical problem of the soundness of a system of proofs is replaced by a precise mathematical problem. The form of the main lemma in a soundness proof for a system of linear proofs is this: for every proof } \pi \text{ the assumptions of } \pi \text{ taken together imply each sentence in } \pi. \text{ In my opinion the form of the corresponding lemma for any correct theory of suppositional proofs is this: for every proof } \pi \text{ the active assumptions of } \pi \text{ taken together imply each active sentence of } \pi. \text{ This opinion, if correct, will account for the feeling of strangeness encountered in trying to construct proofs in the system of Quine's } \text{Methods (pp. 159-167).}\]
in a strictly linear fashion; i.e., each step in a proof must be a logical consequence of the axioms. Apparently this view was first systematized by Boole in the nineteenth century. It became the commonly accepted view until the 1920's when Łukasiewicz pointed out in his seminar that the theory did not agree with mathematical practice. Jaskowski, who was a student in the seminar, accepted the project of developing the exact details of a theory of proof which would take into account the salient features of mathematical reasoning not accounted for by Boole's theory. The newer theory is largely the result of Jaskowski's effort. The older theory we called linear, the newer suppositional.

We gave several examples of rules and proofs with the intention of supplying enough detail so that the basic ideas can be grasped in a useful way.

6. POSTSCRIPT

The linguist and the logician will doubtless disagree with many of the above assertions. Several serious oversimplifications have been made—mostly concerning linguistics. My hope has been to show the overlap and possible cross-fertilization between, on the linguistic side, the ideas of Harris and Chomsky and, on the logical side, the ideas of Jaskowski. I have tried to do this in a way that would be of benefit to persons of diverse backgrounds. I was trying to write to an audience of mathematics educators, linguists, mathematicians, psychologists, and logicians.

One final technical point: the so-called natural deduction systems found in books by Suppes, Lemmon, and Mates are not theories of suppositional proof. By looking carefully at each of them, one notices that the lines of their proofs are not sentences, but rather ordered pairs (P, c) where P is a set of "premises" and c is a single sentence. Moreover, a grammar to generate their proofs takes the form of a linear theory without any assumptions. In particular, in each of these systems each proof is a finite sequence of lines (P_1, c_1), (P_2, c_2), ..., (P_n, c_n) where each subsequent line is either (axiomatically) of the form ([c], c) or else is the result of applying an immediate rule to a fixed, finite number of preceding lines. An example of such a rule would be: if (P_i, d) and (P_j, d ⊃ c) are lines in a proof, then the proof can be lengthened by writing (P_i + P_j, c). The idea behind constructing a proof of c from P in these systems is not to try to deduce c from P, but rather to construct the ordered pair (P, c) starting initially from ordered pairs ([x], x) using rules which when applied to "valid arguments" produce "valid arguments." In a word, these systems stack-up valid arguments starting with the simple and building to the complex. As far as either the characterization of normal reasoning or utility in teaching is concerned, it seems to me that none of these systems fares well in comparison to a suppositional system as found in the following: Anderson and Johnstone (1963), Kalish and Montague (1964), Leblanc (1966), or Thomason (1970).

Acknowledgements

This work originated as a talk given at the Conference on Mathematical and Structural Learning held at the University of Pennsylvania in April of 1968. I wish to thank Dr. Joseph Scandura for inviting me to what developed into a valuable conference. If the final version is a material improvement over the talk, then Professors James Greeno, Paul Rosenbloom...
and Joseph Scandura deserve credit for their suggestions and criticism (not all of which I had wisdom to agree with). I also wish to acknowledge the fact that had I not been fortunate enough to receive a Summer Research Fellowship (NSF-1G-68-3) from the National Science Foundation through the auspices of the University of Pennsylvania, then I likely would not have written these pages. I gratefully acknowledge the helpful and sympathetic criticism that I have received from the students in my logic seminar. Especially significant in regard to this work were the ideas of John Herring, William Frank, Edward Keenan and George Weaver. Finally, I acknowledge ideas received in private communication separately from Mr. James Munz, Linguistics Project, University of Pennsylvania and Professor J. J. LeTourneau, Mathematics Department, Hampshire College. I wish to dedicate this work to the memory of Albert L. Hammond and Ludwig Edelstein; both late of Johns Hopkins University. These two men were instrumental not only in bringing me to appreciate the search for truth but also in making practical arrangements in order that I could pursue graduate studies. Both had tragic lives, one was harassed by philistines in American higher education, the other by Nazi's in Germany as well—but they both fought the good fight and neither lost faith in the ultimate value of truth and kindness.
A THEORY OF MATHEMATICAL KNOWLEDGE:
CAN RULES ACCOUNT FOR CREATIVE BEHAVIOR?1

JOSEPH M. SCANDURA

Mathematics is perhaps the most highly organized body of knowledge known to man. Yet, in spite of its clarity of structure, most of the research done on mathematics learning and behavior has been strictly empirical in nature. To be sure, there has been a fair amount of research in the area and the amount seems to be growing rapidly, but there has been no superstructure, no framework within which to view mathematical knowledge and mathematical behavior in a psychologically meaningful way.

A number of psychologists feel that the mechanisms involved in language, mathematical, and other subject-matter behavior may be accounted for within the confines of S-R mediation theory. This may be possible in principle (e.g., see Millenson, 1967; Suppes, 1969a), but the networks of S-R associations required to do the job would almost certainly be so complex as to provide little intuitive guidance in formulating research on complex mathematical learning. For arguments pro and con, see Arbib (1969a), Scandura (1968, 1970b, 1970c), and Suppes (1969b).

As a way around these problems, linguists, like Chomsky (1957, 1965), have introduced rules and other generative mechanisms to account for (idealized) language behavior. Although many details still need to be worked out, most generally agree that some sort of analysis in terms of rules will prove adequate to account for most language behavior.

During the past few years, the author has been attempting to develop a similar approach to mathematics learning (Scandura, 1966d, 1967d, 1968, 1969b). No comprehensive scheme for classifying mathematical behaviors has been proposed, however, and most (but not all) of the experimental research has been based on relatively simple mathematical tasks (cf. Scandura, 1969b). The basic supposition has been that an understanding of what is involved in such tasks will provide a better position for explaining more complex mathematical learning. While there has been increasing support for this contention among behavioral scientists (e.g., Bartlett, 1958; Gagné, 1965; Miller, Galanter, & Pribram, 1960), some mathematics educators have been skeptical. Presumably, the position is that any interpretation of complex mathematical learning in terms of simple rules will surely be inadequate.
In reaction, the author proposes and defends the rather strong thesis that rules are the basic building block of all mathematical knowledge and that, if looked at in the right way, all mathematical behavior is rule-governed. More specifically, it is proposed that the mathematical behavior any given individual is potentially capable of, under ideal conditions of performance, can be accounted for precisely in terms of a finite set of rules.

This statement is clearly meant to imply more than just a post hoc account of a given finite corpus of behaviors. If limited to this, the claim would be trivially true since any given subject during his lifetime is necessarily limited to a finite number of behaviors. (A finite number of behaviors can obviously be generated by a finite number of rules.)

Furthermore, this is not a thesis to be proved since it is basically empirical in nature. The problem is that there is no operational way of determining the behavior potential of a subject independently of the rules used to characterize his knowledge. Unfortunately, it would be extremely difficult and time-consuming to obtain an adequate sample of mathematical behaviors to work with under the ideal conditions envisioned—that is, where the subject is unencumbered by memory or his limited capacity to process information.

To compensate for this difficulty, the author suggests the proposal and evaluation of alternative characterizations of given finite corpora of behavior in terms of their relative powers and/or parsimony. That is, given a large class of behaviors, such as those associated with mastery of a given school curriculum, the idea is not only to come up with a finite set of rules which characterizes the curriculum but to come up with the best possible set. (Loosely speaking, power refers to the diversity of behaviors which the characterization accounts for; parsimony refers to the number and intuitive simplicity of the rules in the characterizing set.) Such criteria, of course, have been an essential part of formal linguistics ever since Chomsky's (1957) influential Syntactic Structures was published.

In order for a characterization to have maximal relevance to psychology, however, these criteria alone are not sufficient. It is also important that a theory of knowledge (i.e., a characterization) be compatible with the mechanisms which govern human learning and performance. Specifically, it is important, in addition to specifying finite rule sets, to also specify how the constituent rules may be combined to generate behavior. It is these "rules of combination" which must find parallels in the way learned rules are put to use in particular situations. This question of relationships between different levels of theorizing is an extremely important one. For further discussion, see Scandura (1971a).

The basis of the present argument is that, given suitable rules of combination, much of what normally goes under the rubric of creative behavior can be accounted for in terms of finite rule sets. In order to limit the scope, this paper will deal primarily with those kinds of rules which are more properly associated with mathematical or logical content--

If there was some way of knowing, then Church's thesis would provide a natural basis for deciding whether or not the behavior (potential) is rule-governed. Church's thesis (Rogers, 1967, 20-21) is that partial recursive functions (which can be defined formally) are precisely those which can be computed by algorithm (which is an informal notion). Thus
specifically, with mathematical systems and axiomatic theories. In each case, one begins with a mathematical characterization and then shows what it means to know the underlying mathematics in a behavioral sense.

Relatively little attention is given to so-called mathematical processes. Thus, for example, inference rules are discussed, but relatively little is said about heuristics and other higher order rules by which inference rules may be combined in constructing proofs. This does not imply, however, that such processing skills cannot be formulated in terms of rules. To the contrary, it is basically a simple matter to formulate such heuristics as "organize (arrange) the data" and "work backward from the unknown" (cf. Polya, 1962) as rules. What is hard is to show explicitly how these rules may be combined with other rules to solve problems. Even this problem is not insurmountable, however, and some illustrative analyses of this sort have been worked out (Scandura, 1970b).

1. WHAT IS A RULE?

Before continuing, it is necessary to define what is meant by a rule. In spite of an increasing amount of research on the subject, it is perhaps surprising that the term has no clearly defined meaning among behavioral scientists.

As a first step it is necessary to make a sharp distinction between underlying rules—or generative procedures composed of rules—and rule-governed (RG) behavior. Intuitively speaking, a class of behaviors is said to be RG if the behavior can be generated by a common algorithmic (generative) procedure of some sort. This means, in effect, that a person who has mastered any underlying procedure should, ideally speaking, be able to generate each and every response, given any particular stimulus in the class of stimuli.

More specifically, RG behavior involves the ability to give the appropriate response in a class of functionally distinct responses to each stimulus in a class of functionally distinct stimuli. (The term "functionally distinct" refers to the fact that each effective (i.e., functionally distinct) stimulus (response) corresponds to a class of overt and "functionally equivalent" stimuli (responses).) The class of S-R pairs, defined in this way, are called S-R instances. To see what this means, consider simple addition. The proposed definition says that the behavior is RG if each pair of numbers is attached to a unique number called the sum. Thus, for example, any overt representation of the number pair (5, 4) can be paired with any overt representation of the number 9 but not with any representation, say, for the number 6.

Ideally, then, RG behavior corresponds precisely to the notion of a function in the mathematical sense. That is, every stimulus is paired with a unique response. When looked at in this way it is clear that the proposal would be true or false depending on whether the class of potential behaviors is or is not partial recursive.


4 As indicated above, of course, the behavior of human beings is not always ideal. People make mistakes. There are two conceptually differ-
what psychologists call concepts and associations can be viewed as special cases of rules (Scandura, 1968, 1969a, 1969b). Concepts are simply rules in which each stimulus in a class is paired with a common response. Associations are further restricted to a single stimulus-response pair.

In its simplest form, a rule can be viewed as an ordered triple \((D, 0, R)\), where \(D\) is the set of \((n\)-tuples of) stimulus properties which determine the responses, and \(0\) is the operation or generative procedure by which the responses in \(R\) are derived from the critical properties in \(D\) (Scandura, 1966d, 1967d, 1968). More of the detail involved can be represented by adopting ideas taken from recursion theory. In particular, a generative procedure is a sequence consisting of at most four kinds of rules:

1. encoding rules by which essential properties of stimuli are put into store,
2. transforming rules by which things in store are transformed into something else in store,
3. decoding rules by which things are taken out of store and made observable,
4. rules for selecting other rules (These correspond to decision making capabilities.)

Church (1936) has proposed that any set of behaviors which mathematicians would be willing to classify as partial recursive can be generated by a procedure composed of just these four types of rules. In general, this would include just about all of the mathematical behaviors one normally expects of the school-age child, the ability to perform arithmetic computations, to construct geometric figures with ruler and compass, etc.

2. CHARACTERIZATION OF MATHEMATICAL KNOWLEDGE

The main purpose of this paper is to indicate how complex mathematical behavior might possibly be accounted for in terms of finite rule sets. Various ways in which errors may occur. First, the rule(s) learned by a subject may only apply to a subclass of S-R instances (of the given RG class). Thus, for example, young children are frequently unable to add numbers which involve "carrying" although they can perform perfectly well on those that do not. In this case, following the notion of partial function in recursion theory, one may refer to such behavior as partial RG behavior. Partial RG behavior is rule-governed but not (necessarily) by rules associated with the given RG class. The other way in which errors may arise is due to the limited capacity of human subjects to process information (Miller, 1956). There is, in effect, an important difference between knowing a rule and being able to use it (Chomsky & Miller, 1963). Thus, a person may know how to add any pair of numbers but be quite unable to perform the necessary operations mentally when the numbers are large. In the present discussion, the author assumes throughout that all rules can be used perfectly.

Note (parenthetically) that the abstract notion of a functor is sufficiently flexible to capture either or both senses of incompleteness. (Roughly, a functor is a structure preserving function between two categories, the categories being analogous to classes of functionally distinct stimuli and responses.) Whether there is any real significance to this fact or not, however, the author cannot say (cf. Scandura, Volume 1).
2.1 Mathematical systems.

Every mathematical system consists of one or more basic sets of elements, together with one or more operations and/or relations and/or distinguished elements of the basic sets. By capitalizing on certain logical equivalences it is possible to reduce the characterizing elements to one basic set and one or more relations. Consider a simple example—the system whose basic set consists of three "undefined" elements A, B, C, denoted \{A, B, C\}, with A being distinguished in the sense that it serves as an "identity," and whose defining relation is \(\circ = \{(A, A) \rightarrow A, (A, B) \rightarrow B, (B, A) \rightarrow B, (A, C) \rightarrow C, (C, A) \rightarrow C, (B, B) \rightarrow C, (C, C) \rightarrow B, (B, C) \rightarrow A, (C, B) \rightarrow A\}\). This is a system in which the distinguished element A "maps" every element it is paired with into itself. When B is combined with B, the result is C and when C is combined with C, the result is B. Finally, B combined with C in either order results in A. Notice that no meaning is specified for either the elements A, B, C, or the operation. They are "undefined terms."

What may be called an embodiment of a mathematical system results on assignment of meaning to the undefined elements. Thus, in the example just cited, the undefined terms might correspond to certain rotations with A corresponding to a rotation of 0°; B, to a rotation of 120°; and C to a rotation of 240°. In this case, the operation would simply be "followed by." That is, the result of combining two rotations is that single rotation which results in the same action as first doing one rotation and then the other. For example, a rotation of 120° followed by one of 240° results in the same action as a rotation of 0°.

These definitions of systems and embodiments say something about the nature of the objects we are studying and in that sense they are extremely important. They do not, however, tell very much about their psychological nature.

What kinds of behavior are implied by knowing systems and embodiments of this sort? And, how can such behaviors be accounted for in terms of rules?

First, knowing a system certainly implies the ability to compute within the system. Thus, for example, given the pair, A, B, the "knower" should be able to give the "sum," B. He should also be able to do more complex computations, like \((A \circ B) \circ A \rightarrow A, (B \circ A) \circ C \rightarrow C, (B \circ B) \rightarrow A\), which involve combining individual facts (i.e., associations). These facts correspond to non-degenerate rules in the various embodiments of the system. For example, in the illustrative embodiment, the fact, B \(\circ C = A\), corresponds to a rotation of 120° "followed by" one of 240°. The rule (operator) of doing one and then the other applies to all pairs of rotations, not just one (pair). In addition, knowing a concrete display corresponding to this embodiment involves being able to perform the various rotations on whatever concrete objects (e.g., an equilateral triangle) might be involved and whatever its position or orientation.

As anyone who has worked with young children knows, this is not something which can automatically be assumed. (One thing which can easily be overlooked in analyzing behaviors, for example, is that these "rotations" are actually equivalence classes of rotations, and that these equivalence classes may be different for child and observer.) While such things may not be important in mathematics, strictly speaking, they are relevant in science and, in the opinion of the author, ought to be dealt...
the knower should be able to give "differences," i.e., given the sum and one of the "addends," he should be able to generate the other addend.

If these were the only kinds of behavior to be accounted for one could simply list the facts (rules) involved. But clearly any reasonable interpretation of "knowing a system" must also deal with relationships as well. For example, mastery of a system would surely include the ability to generate the subtraction (difference) rule from the addition rule, and vice versa. Knowing that \( B + C = A \), for example, should be a sufficient basis for generating the corresponding subtraction fact, \( A - B = C \).

Relational rules of this sort provide a simple way to account for such behaviors. Thus, instead of listing all of the subtraction facts separately it would be sufficient to know the addition facts together with the relational rule. That is, assuming, as is traditional in formal linguistics, that individual rules can be composed--performed in succession.

The obvious way to account for such relationships--the way taken by curriculum developers of the operational objectives persuasion--is to simply add more rules to the characterization. There are, however, major problems with this approach (Scandura, 1970a). For one thing, listing a new rule for each kind of relationship would have a post hoc flavor not likely to add much in the way of understanding more creative behavior. For each new system (of the same type) considered, for example, there would be a new relational rule for each one in the original system. Even granting the economy obtained by eliminating inverses, and the like, the number of rules could grow large very fast. This would not be bad in itself assuming that this is the best one could do. The important question, however, is: Can one come up with a more efficient account which is at the same time more powerful--and which allows for some measure of creative behavior?

To answer this question, first note that knowing how one or more systems are related to a given one may provide a basis for knowing how to compute in the new systems given how to compute in the original. The relationships of interest will generally be mathematical in nature, but they need not be limited to morphisms. For example, one system may be a simple generalization of another, as with cyclic 5 and cyclic 3 groups. Because of the way particular relationships are defined, however, this advantage will generally be of a limited sort. With homomorphisms, for example, the ability to compute in the new system applies only to the defining operations themselves and not, say, to their inverses or to relationships between the operations. It is worth noting, nonetheless, that knowing even a relatively simple set of interrelated rules such as this would make possible a certain degree of creative behavior--what might be called "analogical reasoning." For example, suppose that a subject has learned how to add in system A and that he knows the homomorphism which connects A to system B (i.e., that he can generate the elements in B with as an integral part of the elementary school mathematics curriculum.

In an important sense, then, knowing a concrete embodiment (or a corresponding display) may involve a different type of knowledge than knowing the same amount about a corresponding system. This observation could have relevance to a number of recent results (Dienes & Jeeves, 1965; Scandura & Wells, 1967; Suppes, 1965) and should be taken explicitly into account in designing future studies.
which correspond to those in A). Then, the subject should be able to add in system B without ever being told how. Consider the homomorphism to be one-to-one (i.e., an isomorphism), system A to be the embodiment of the illustrative 3 group above, and system B to be the illustrative system itself, then one might generate a sum in the abstract system B by (a) using the isomorphism to determine the corresponding elements in A, (b) adding in A, and (c) using the isomorphism in reverse direction to determine the element in B corresponding to the sum (in A). Notice that this follows only if our rules of combination allow for combination (of rules).

A far more powerful and parsimonious characterization results by simply allowing rules to operate, not on just ordinary stimuli, but on other rules. Such rules may be said to be acting in a higher order capacity—or, in short, to be higher order rules. Although functions on functions are common in various branches of analysis, and their formalization is routine, the idea seems not to have pervaded formal linguistics. The closest linguists have come in this regard has been to introduce the notion of a grammatical transformation between phrase markers (Chomsky, 1957), which closely parallels what are here called relational rules (e.g., between addition and subtraction).

There are two reasons why this has probably not been done in the past. First, grammatical transformations have so far resisted mathematical treatment (Nelson, 1968) insofar as this relates to computer science and, second, no existing approach to psychology (known to the author) provides any real motivation for introducing them.

This is unfortunate since there is a very simple and intuitively sound reason for including higher order rules. The main one is just this: The idea of allowing rules (in rule sets) to operate on other rules is compatible with the following intuitively appealing hypothesis concerning performance. If a subject does not have a rule available for achieving a desired goal, then he typically will try to construct a rule which does work (cf. Scandura, 1971a). There is a good deal of introspective evidence in favor of this hypothesis, and some empirical support for it has been collected. In a recent study (Scandura, 1967b), it was found that the ability to "use parentheses" was a sufficient basis for combining learned rules so as to solve the given tasks which involved interpreting new statements of mathematical rules. Later analysis of these tasks showed that use of parenthesis may be viewed as a higher order rule (Scandura, 1970b). The author is currently involved in research in which success in generalizing this result to a number of different kinds of situations and populations has been achieved (Scandura, 1971a).

Allowing rule sets to act in this way makes it possible for them to "grow" in ways not possible by just forming simple compositions (of rules). Thus, (higher order) rules may generate completely new kinds of rules, and these rules, in turn, may be used to generate still other rules.

Consider what higher order rules might suggest in the present situation. Suppose that a subject has learned a higher order rule which connects each operator (rule) with its inverse. Such a rule would connect not only, say, addition of numbers with subtraction, but composition of all sorts (e.g., of permutations, rotations, rigid motions, etc.) with the corresponding inverse operations. The defining operation of each system and its inverse may be thought of as being distinct rules which are mapped one on the other by this higher order "inverse" rule. Assume, in addition, that the subject has learned how to add in system A, the rela-
tionship (e.g., a homomorphism) between system A and system B, and also how to form the composition of arbitrary rules (in the rule set).

In this case, there are all sorts of behaviors that the (idealized) subject would be capable of. For example, he would be able to subtract, not only in system A but in system B as well. To see this, one need only observe that the subject can form the composition of the rule between systems A and B and the higher order inverse rule. This composite (higher order) rule in turn allows the subject first to generate an addition rule in system B and then to generate a subtraction rule in system B. This subtraction rule, in turn, would allow the subject to subtract. Translated into more meaningful terms, a rule set of this sort would imply such abilities as finding inverses with rigid motions given only the ability to add numbers. But, then, isn't this just what is considered as creative behavior?

2.2. Axiomatic theories.

There is clearly more to knowing systems than simply knowing the rules and interrelationships within these systems. This amounts to internal knowledge of the systems but it says nothing about the systems in the descriptive sense.

Axiomatic theories are concerned with properties of systems. As an example of one such property, notice that in the illustrative system it does not make any difference in which order two elements are combined. The system satisfies the commutative property; in fact, it satisfies all of the axioms (i.e., properties) of a commutative group of order three.

In order to define precisely what is meant by an axiomatic theory, the next thing to observe is that a set of axioms or properties defines a family of systems, namely that family consisting of all, and only, those systems which have each of the given properties. Therefore, an axiomatic theory may be defined to be the set of properties which holds in the family of systems defined by a given set of axioms. The set of axioms, of course, belongs to the set of properties.

Paralleling the discussion of systems, consider the question: 'What kinds of behavior are involved in knowing axiomatic theories and what kinds of rules are needed to account for these kinds of behaviors?' Due to the complexities involved, the discussion will be restricted largely to lower order rules.

The sine qua non of mastering a theory is to know the axioms and theorems of that theory. In behavioral terms, this ability may be thought of as being able to give on demand the conclusions associated with each set of premises. Thus, as with knowing the particular "addition" facts of the illustrative system, one might be tempted to characterize knowledge of particular theories as sets of discrete associations. This would be wrong, however, on two counts. First, the number of theorems associated with any given theory (including trivial ones) is infinitely large, so that they could not all be learned in this way. (Of course, the number of important theorems is usually much smaller.) Second, and more basic, such a characterization, while feasible in part, would not be very parsimonious or powerful. Many more rules would be needed than might be desired and important relationships would simply be ignored.
One problem has to do with not knowing proofs of the theorems, but there is more to it than just that. Proofs can be learned in a strictly rote fashion and being able to generate one may signify little more than simply knowing the theorem itself.

The kind of rule in mind may act not only in any given theory, or even in any class of theories, but these rules may act in any theory whatsoever—indeed, in any situation at all. They are closely related to inference rules of formal logic but they do not act on strings of symbols nor do they generate strings of symbols. Neither do they all map properties of systems into properties of systems as one might suspect in view of the relationship between formal systems and axiomatic theories. (Strings of symbols of formal systems correspond to properties of mathematical systems.)

Some inference rules are of an entirely different sort. Instead of operating on properties of systems and generating new properties, what have been called suppositional inference rules map logical arguments into properties. Some work has been done in this area under the label "natural deductive systems," e.g., see Kalish & Montague (1964), Prawitz (1965), but little has been done with behavioral questions in mind. In present terminology, the suppositional inference rules correspond to rules which map instances of other inference rules, or combinations thereof, into properties. For example, from any specific argument, in which property B follows directly from property A, one can infer the property, A → B. In an important sense, then, suppositional inference rules correspond to what are referred to above as relational rules, and transformations, and not to higher order rules—since they do not operate on other rules, but on instances of other rules.

The stimuli of RG behavior may be viewed as families of systems and the responses as derived properties of these families, called theorems. Thus the RG behavior associated with any particular logical procedure involves a class of families of systems and a class of corresponding theorems of these various families. If the procedures are sufficiently unique, e.g., as in proving many non-trivial theorems, the class of RG behaviors may be quite small, indeed it could include only one instance.

In effect, a logical procedure may act on corresponding properties of different families of systems, and produce other properties of the respective families, called theorems. Some idea of the way complex logical procedures operate can be obtained by considering familiar rules of inference. Modus ponens provides a simple illustration. Suppose that the statements "If G is a finite group and S is a subgroup of G, then the order of S divides the order of G" and "G is a finite group and S is a subgroup of G" are properties of one family of systems (actually, of pairs of systems), and "If a function is continuous over a closed interval of the real line, then it is uniformly continuous" and "The function is continuous over a closed interval of the real line" are properties of another family. Then application of the logical rule (of inference) modus ponens tells us that "the order of S divides the order of G" and "the function is uniformly continuous" are also properties of the respective families. The corresponding premises and conclusions are quite different but the (logical) rule of inference by which they are related is identical.6

^Note, the proposed definition of RG behavior as a function, has been questioned. The comment has been made that "the futility of trying to think of rules of inference (even) as functions is already evident once one considers substitution of equals." However, careful thought
The same general idea may be extended to more complex logical procedures. In this case, encoding rules involve accepting, or rejecting, properties, axioms and theorems of families as appropriate to given goals the subject might have. Rules of inference correspond to transforming rules (type two rules) and stating theorems, to decoding (type three). Branching rules (type four) may also be involved in logical procedures, as, for example, when repeated application of a rule of inference is required. For example, the conclusion "D" can be inferred from the premises "A ⊃ (B ⊃ (C ⊃ D))", "A", "B," and "C" by repeated application of modus ponens.

Since inference rules and the generative procedures which may be constructed from them apply in all conceivable situations (i.e., to properties of situations), it may be that they might be discovered at an early age from instances—in the same way as many other rules. That is, (learning) deduction may be viewed as induction on a logical rule. If this is true, it could have important implications both for the study of mathematical reasoning and for teaching it.

Of course, no one individual has mastered, or ever will, all of the logical procedures that might be constructed. Such knowledge constitutes an ideal which can only be approached. The behavior involved in proving any non-trivial class of theorems is necessarily partial. According to Church (1936), there exist classes of theorems for which no generative procedure can possibly exist. This does not necessarily mean, however, that theorems which belong to such classes can never be proved. Some procedure might exist for deriving any particular theorem; Church's thesis is simply that no one procedure will do for the entire class.

Nonetheless, many logical procedures, even reasonably complex ones, are apt to be common to a number of different theories. The number of more or less unique procedures in any particular theory is likely, according to the present view, to be relatively small. Hence, assuming prior mastery of most "standard" logical procedures, a skilled mathematician may gain mastery of a new theory in relatively short order by concentrating on those procedures associated with some of the deeper theorems of the theory. Note that logical procedures correspond roughly to proof schemas—that is, to classes of proofs of the same general form.

In order to prove most theorems, indeed to successfully engage in complex deductive reasoning of any sort, a subject must know more than just rules of inference, or even a large number of relatively complex logical procedures. The subject must also have higher order rules available by which he can combine known inference rules and other logical procedures into new forms—that is, so that he can create. One type of higher order rule that is frequently used in constructing proofs is closely associated with the heuristic: "Work backward from the conclusion." In this case, the learner attempts to derive a procedure for generating the conclusion from the premises, i.e., to construct a proof, by first selecting an inference rule which yields the conclusion and then trying to derive a logical procedure, by using this or other higher order rules, which yields the input of the first rule selected. Presumably, the subject continues

should convince one that the input of such an inference rule maps pairs of the form y = b, P(b) = K into elements of the form P(y) = K where y is allowed to vary. Thus, the form corresponds to a class of functionally distinct stimuli (e.g., $a_1 = b$, $P(b) = K$; $a_2 = b$, $P(b) = K$; ...) and so it is not surprising that one can generate any number of different responses (e.g., $P(a_1) = K; P(a_2) = K; ...$).
in this way until he either succeeds or the whole approach breaks down. The widely used technique of proving theorems indirectly, by assuming that the conclusion is false, provides a particular example of a higher order rule generated by application of this heuristic (a still higher order rule). In this case, the problem reduces to one of constructing a proof of \( \neg A \) from \( \neg B \). The final step in constructing such a proof just amounts to selecting what might be called the contrapositive inference rule by which the theorem \( A \Rightarrow B \), can be inferred from the argument from \( \neg B \) to \( \neg A \).

More could be said about such things as formal systems and metamathematics but space does not permit. In the first case, it suffices to say that formal systems are easier to work with than axiomatic systems. Nothing new is required, except that the allowable inference rules are specified, and no decoding rules are needed. The axioms and theorems are themselves the stimuli and responses. Metamathematics turns out to be nothing more than an axiomatic type of theory in which only non-controversial rules of inference are allowed.

3. CONCLUDING COMMENTS

In conclusion, this paper has dealt primarily with what it means to know an existing body of mathematics. Relatively little has been said about intellectual skills of the sort that must inevitably be involved in doing real mathematics. Nonetheless, it has been shown that what appears to be creative behavior might well be accounted for in terms of growing rule sets. The key idea in making this a feasible and rather attractive possibility is that of the higher order rule. Although space limitations have made it necessary to ignore many details, and there obviously are still a good many important questions left unanswered, the author feels that enough has been said to convince the reader that the basic conjecture must be taken seriously: all mathematical behavior is a rule-governed activity and the basic underlying constructs are rules.
FORMULATING MATHEMATICAL MODELS
OF PSYCHOLOGICAL PHENOMENA

ZOLTAN DOMOTOR

Mathematical modeling is rapidly becoming one of the most powerful and important methods of studying the formal behavior and structure of empirical systems, processes, events, and phenomena. However, there are still many people who question the real value of model building technique in the methodology of social science. They argue, firstly, that contemporary mathematics no matter how sophisticated it may be, is poorly adapted to the needs of social science problems which are characterized by an extremely high number of interacting variables. Secondly, they object that even if one is successful in isolating a reasonable number of essential variables, still the domain of validity of the corresponding formal model is hopelessly small. Simply, in complex systems minor causes sum up to major effects, without knowing anything about the causal or stochastic mechanism involved.

It is the purpose of this paper to show that this criticism is wrong and based on a complete misunderstanding. In particular, we shall consider the general aspects of mathematical modeling, and argue its basic importance in the study of behavioral phenomena by presenting a concrete model of supervised learning. We have deliberately chosen the example from learning because it contains all the important patterns of model building relevant in social science. Let us also point out that the case study presented here has an independent importance in the theory of education.

More concretely, since the supervised learning process is a highly complex problem, it cannot be solved adequately and economically by a simple (say S—R) theory. It has to be torn apart into several subprocesses. Each subprocess is solved separately and then the component solutions are interconnected into the solution of the resultant process. The main components are as follows:

1. The technical part of this paper is in preparation.

The author profited greatly from conversations with Joseph M. Scandura (cf., (1970), in particular), concerning the method of applying rules in problem solving. Clyde Greeno's talk on the communication-theoretic aspects of learning and instruction also has certain aspects in common with the model proposed.
(i) **Concept algebra C**, together with a **complexity measure C**. The algebra structures a homogeneous class of concepts (rules, etc.) being taught by the teacher;

(ii) **Problem algebra P**, together with a **difficulty measure P**. This algebra organizes the problems being utilized for testing the achievement of the learning subject;

(iii) **Solution space S**, together with the **average achievement function A**. This space contains all the solutions of problems from the problem algebra P.

(iv) **Instruction space D**, together with the **average teaching efficiency function E**. The space contains all the teaching strategies corresponding to concepts from the concept algebra C.

(v) **Learning subject <K, a>**, being represented by a knowledge structure K and achievement function a.

(vi) **Teacher <T, e>**, being represented by a class of teaching strategies T and efficiency function e.

These components are studied separately and then put together into a complex system which then reveals useful information about supervised learning, and answers some of the basic questions of education, teaching, and learning (quality of teaching, efficiency, achievement, learning rate, and their mutual dependence and causal influence). The flow chart of the resulting model is presented in Figure 1.

Composition of subsystems (subprocesses) into bigger system (process) units requires additional mathematical entities, expressing the interaction of component systems (processes).

---

**Figure 1**

1. **MATHEMATICAL MODELING**

Mathematical model is a formal representation of an actual object, system, situation, process, event, or phenomenon from the point of view of those aspects which are relevant to the given task. It mimics the behavior and structure of its real counterpart in some respect. However, the subject matter of modeling is not to imitate for the sake of imitation, but to enable certain decisions to be taken; and the question is whether what we omit in a model is relevant to the decision, and not to
the accuracy of the representation! The criterion by which we should evaluate the success or failure of any modeling process is the extent to which we have constructed a conceptual tool enabling us to achieve the given specified objective.

Generally speaking, the scientific purpose of modeling is to study an unknown entity $\mathcal{U}$ by examining the collection of structure preserving maps $\mathcal{F}$ together with some simple, or at least more concrete mathematical entity, called a model $\mathcal{M} : \mathcal{U} = F(\mathcal{M})$, where $F \in \mathcal{F}$.

If the examination of the representing model $\mathcal{M}$ is to serve as an effective tool for obtaining information about the structure of the original unknown entity $\mathcal{U}$, we must choose (or discover) such a model $\mathcal{M}$ from the a priori given model space $\mathcal{M}$, which completely determines the structure of the unknown $\mathcal{U}$ up to the map $F \in \mathcal{F}$.

In cases, when the unknown entity $\mathcal{U}$ cannot be represented by a single model, we may still succeed in representing the entity by several partial models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m \in \mathcal{M}$, each of which represents a particular part of. Then the representation has the form:

$$\mathcal{U} = F(G_1(\mathcal{M}_1), \ldots, G_m(\mathcal{M}_m)),$$

where $F \in \mathcal{F}$, and $G_i \in \mathcal{G}_i$, $i = 1, 2, \ldots, m$. $\mathcal{G}_i$'s are groups containing various structure preserving maps. Actually this is the case of modeling supervised learning processes.

The process of modeling is based on the following five general principles:

- **P₁** Existence (Abstraction) Principle
  With every consistent class of statements (presenting experimental data, knowledge, describing specific aspects of reality, forming a theory of certain phenomena) there is associated an empirical (idealized, intended) system, process, or phenomenon $\mathcal{U}$.

- **P₂** Representation (Realization) Principle
  Every empirical system, process, or phenomenon $\mathcal{U}$ is representable by a minimal (simplest, optimal) set-theoretic model $\mathcal{M}$ (chosen from an a priori given model space $\mathcal{M}$), in which the irrelevant parts of $\mathcal{U}$ (which do not help to explain the given experimental facts) are filtered (factored) out and whose formal properties are invariant with respect to the specific intensional nature of the empirical entity $\mathcal{U}$.

- **P₃** Uniqueness (Invariance) Principle
  The set-theoretic model $\mathcal{M}$ is unique up to a group of (scaling, space-time, observation, discourse frame) structure preserving transforms $\mathcal{G}$.

- **P₄** Adequacy (Continuity) Principle
  The behavior and the properties of the set-theoretical model $\mathcal{M}$ (being nothing but a mathematical scheme of possible experimental data) are inherent in the data, description, and theory used in its construction. That is to say, the minimal mathematical model $\mathcal{M}$ does not introduce any constraints or limitations other than those which are stated by the structure of...
data, generating the model. Any change of data (like choice of the level of description, increase of the number of observed attributes, change of functional dependence) leads to a corresponding change of the mathematical model $M$.

**P5 Testing (Operationalization) Principle**

To match the model $M$ to the corresponding real world problem $\mathfrak{A}$ always leads to inherent subjective value judgments (loss, risk, cost, error) expressed in terms of various criteria of adequacy and performance $B$. The method of modeling is therefore limited and qualified by the modelbuilder's value judgment, his a priori knowledge and by his degree of realism he wishes to incorporate into his model.

A model, being nothing but an information-theoretic conceptual device (a channel), providing us with information about the real world situation, must be tested. Testing is a decision-theoretic process giving us information about the cost, certainty, quality, accuracy, and adequacy of the news received about the real world problem $\mathfrak{A}$, via the model $M$.

The task of modeling is to construct a least complex model $M$ with maximal amount of information about the real life problem $\mathfrak{A}$ satisfying the required level of accuracy and adequacy $B$. Hence, modeling is always a ternary relation: $M$ is a model of $\mathfrak{A}$ given the conditions $B$. $M$ is an adequate model of $\mathfrak{A}$ exactly when $M$'s answers agree with those of $\mathfrak{A}$'s with respect to conditions $B$ (see Minsky, 1968).

Instead of going into technical details of modeling principles, we shall give some comments on model formation.

To arrive at a mathematical model which will represent the system, phenomenon, etc., under consideration with a given level of adequacy, we have to:

(i) First, itemize all the components and the attributes that will contribute to the system's structure and operation. Once the list of component elements and attributes is completed, the next step is to decide which of these items will be actually used in the model. The choice depends on the required level of accuracy.

   It is convenient to partition the items into various classes (deterministic, stochastic, learning, etc. components, and controllable, observable, parametric, causal, etc. attributes) and specify what are the possible values the attributes can take;

(ii) Secondly, structure all the components and attributes. That is to say, give on the basis of prior knowledge and requirements the hierarchical or network-like composition of the components, and the explicit relationships among the attributes in order to represent the system's structure and behavior.

The set of relevant components forms a "dead" skeleton of the represented system on which a "live" collection of empirical attributes is superimposed. While the components are usually represented by some kind of algebraic structures (lattices, graphs, groups, vector spaces), the attributes are described by real or random functions, defined on the
algebraic structures. The behavior of the components is represented by equations or inequalities expressing the interaction of attributes.

It is important to realize that the study of a real life situation always reveals two basic aspects: (1) the formal which is encoded in the model and serves the purpose of information communication about the unknown real life entity, (2) the concrete which is encoded in the testing procedure of the model and serves the purpose of evaluation, decision, and control.

2. SUPERVISED LEARNING MODELS

The purpose of teaching or instruction is to induce the learning subject to behave in ways which are deemed desirable, such as to integrate complicated functions, solve differential equations, design physical devices according to stated specifications, etc.

Learning is a process of change in behavior which follows the previously shaped behavior. The learning subject externalizes his thinking so that the teacher can evaluate his progress and decide what instructional step to take next. There is a constant interaction between teacher and student.

Every teacher has a set of goals—a list of things that he wants to teach. These things can be almost anything. Suppose that we teach concepts or rules. For example, if one were teaching integral calculus, the list of rules to be taught might include such items like:

\[ r_1 \text{ Summation rule: } \int [f(x)+g(x)]dx = \int f(x)dx + \int g(x)dx; \]
\[ r_2 \text{ Integration by parts rule: } \int f(x)g(x)dx = f(x)\int g(x)dx - \int \left( \frac{d}{dx}f(x) \right) g(x)dx; \]
\[ r_3 \text{ Substitution rule: } \int f(u)du = \int f(\varphi(y))\varphi'(y)dy, \]

where \( \varphi \) is the substitution transform.

Besides the rules one will certainly include various definitions which specify the conceptual basis of calculus.

Furthermore, the instructor will arrange these concepts or rules in some kind of logical order so that he can teach the concepts to the student in sequence. The order will represent the degree of complexity of these concepts. Less complex concepts are taught first. Some of the concepts or rules will be composed of simpler concepts.

It is a matter of modeling to realize that the collection of concepts to be taught, forms an algebra \( \mathbb{C} = \langle \mathbb{C}, \cup, \rangle \), where the operations have the following meaning: If \( C_1, C_2 \in \mathbb{C} \), then

\[ C_1 \cup C_2 \text{ is the superposition of two concepts } C_1 \text{ and } C_2, \text{ involved in an application to a problem situation. For example, if one were to integrate a function, say } f(x), \text{ it may be necessary or desirable to use the rule } r_1 \text{ and then the rule } r_2. \]

This would amount to using the formal composition

\[ r_1 \cup r_2 (\int f(x)dx) = r_2 (r_1 (\int f(x)dx)); \]
(ii) \( c_1 + c_2 \) is the conjunction of two concepts \( c_1 \) and \( c_2 \). For example, if we would apply both rules \( r_1 \) and \( r_2 \) to a problem situation, e.g., finding the integral of \( f(x) \), then

\[
\int r_1 + r_2 (f(x)dx) = \int r_1 f(x)dx \quad \text{and} \quad \int r_2 f(x)dx.
\]

We shall not explore in this paper the algebraic properties of the concept algebra \( C \). As pointed out before, it is an ordered algebra, ordered by a binary (complexity) relation \( \leq \). To this ordering corresponds a complexity measure \( C \) which measures the degree of complexity of concepts:

\[
c_1 \leq c_2 \iff C(c_1) \leq C(c_2), \quad c_1, c_2 \in C
\]

Concepts \( c \), for which \( C(c) = 0 \), might be viewed as elementary or known. Concepts \( c \), for which \( C(c) = \infty \), are impossible to understand. Often it is convenient to consider the notion of dependence and/or independence of concepts. In this case we need a further binary relation \( \perp \), the independence relation. In addition, it is desirable to require the additivity of the complexity measure:

\[
c_1 \perp c_2 \Rightarrow C(c_1 \land c_2) = C(c_1) + C(c_2), \quad c_1, c_2 \in C
\]

By complexity of concepts we mean here a function of all elements of a complete set of invariants of concepts (see Domotor (1969, p. 110).

Now suppose the teacher has his organized algebra of concepts \( \langle C, C \rangle \) that he desires to teach. How does one find out whether or not the student has absorbed or understood the concept that he was taught? It is necessary to assume that there is a set of test questions for each concept. The test questions are presented to the student at the end of his instruction on the concept to find out how well he has absorbed the subject matter. Questions are chosen appropriately from a problem area related to concepts. So, in our example of integral calculus the instructor would present some problems like computing certain messy looking integrals. Again, the problems will have to be ordered, this time with respect to their difficulty. Moreover, problems are decomposable into subproblems, and these into subsubproblems, etc., until we arrive at trivial problems. The collection of all problems, corresponding to the concept algebra \( \langle C, \land \rangle \), will be a new structure \( \mathcal{P} = \langle \mathcal{P}, + \rangle \), where the operation \( + \) stands for composition of problems. Problems are ordered by a binary relation, called the difficulty relation, and measured by a difficulty measure \( F \):

\[
p_1 \leq p_2 \iff F(p_1) \leq F(p_2), \quad p_1, p_2 \in \mathcal{P}.
\]

Problems \( p \), for which \( F(p) = 0 \), are trivial and problems \( p \), for which \( F(p) = \infty \), are impossible to solve. It is useful to consider the notion of independence \( \perp \) in the set of problems. For independent problems the measure is additive:

\[
p_1 \perp p_2 \Rightarrow F(p_1 + p_2) = F(p_1) + F(p_2).
\]

By the difficulty measure we mean the amount of labor necessary for solving the problem.
The concept algebra \( \langle C, \mathcal{C} \rangle \) is mapped by a problem functor \( R \) into the problem algebra \( \langle P, \mathcal{P} \rangle \), so that certain homomorphic conditions are satisfied. These two algebras form the stimulus set of the instructor and the learning subject.

The response of the learner to problems is a collection of answers. As expected, answers to questions are closely related to problem solving matters. We are concerned with problems which can be described by conditional imperative sentences of the form

\[ \forall x \in S \Phi(x) / R, \]  

where the intended interpretation is: Find all \( x \) from the set \( S \) such that condition \( \Phi(x) \) is satisfied, given the data base \( R \). The variable \( x \) can run over various sets: sets of numbers, truth values, functions, models, proofs, theories, etc. \( \Phi(x) \) specifies the attributes of the desired solution, while \( R \) contains general descriptive information about the problem environment (system of concepts, rules, theory).

Take, for example, the self-explanatory problem of quadratic equations over the real field \( \mathbb{R} \):

\[ \forall x \in \mathbb{R} \left[ 3x^2 + 2x - 5 = 0 \right], \]  

where \( \mathbb{R}[x^2] \) stands for the field of quadratic polynomials over \( \mathbb{R} \).

A standard form of a problem from integral calculus is:

\[ \forall f \left[ f(t) = \int g(t) dt \right] / \mathcal{V}, \]  

where \( \mathcal{V} \) = linear space of real functions.

The class of problem sentences of the form (2) is governed by imperative logic. It is not the purpose of this paper to develop the syntax of these sentences. Our primary concern is the semantics of problem solving. It is convenient to describe the semantics of a problem by a triple \( \langle A, \mathcal{O}, B \rangle \), where \( A \) denotes the initial state of the problem, \( B \) is the final (goal) state of the problem and \( \mathcal{O} \subseteq \mathcal{C} \) is the set of admissible operators (concepts, rules), defined on the set of states of the problem. It is assumed that every problem has a state space, and to solve a problem means to find operators from \( \mathcal{O} \) in such a way that their subsequent application to the initial state will eventually give the final state.

For example, problem (3) is processed as follows:

\[
3x^2 + 2x - 5 = 0 \quad c_1 \quad (\sqrt{3}x + \sqrt{3}/3)^2 - 1/3 - 5 = 0 \quad c_2 \\
\sqrt{3}x + \sqrt{3}/3 \quad c_3 \quad 4/\sqrt{3} = \pm 4/\sqrt{3} - \sqrt{3}/3 \quad c_4 \\
x \in \left[ -5/3, 1 \right],
\]

where operators \( c_1, c_2, c_3, \) and \( c_4 \) from have the following definition:

\( c_1 \) = put the quadratic polynomial into square form;
\[ c_2 = \text{get rid of the square by computing the square root}; \]
\[ c_3 = \text{get rid of the absolute value by considering the + and - cases}; \]
\[ c_4 = \text{express } x \text{ in the membership form.} \]

In this example, \("3x^2 + 2x - 5 = 0\)" is the initial state of problem (3), and \("x \in \{-5/3, 1\}\)" is the goal state of the problem. We have
\[
x \in \{-5/3, 1\} = c_4(c_3(c_2(c_1(3x^2 + 2x - 5 = 0)))).
\]
The solution procedure of problem (3) is the sequence \(c_4 \land c_3 \land c_2 \land c_1\). Clearly, this is not the only way of solving a quadratic equation; one may use some substitution operators as well. This means, in particular, that from the initial state we can reach the final state via several different paths. We can rank the paths and consider a corresponding cost function.

The reader should take notice of the fact that to every state of a given problem corresponds in a unique way a problem sentence; \(\forall\) change a state into a problem by adding the imperative operator \(\forall x\) and the problem frame \(P\). The notion of state space is used only in the semantics of problems.

More often than not, the problem is reduced to several subproblems. Consider, for example, the problem of indefinite integral,
\[
\forall f \left[ f = \int g \, dh \right] / V.
\] (5)
Here we may use rule \(r_2\) from the list (1) and thus reduce our problem (5) into two subproblems:
\[
\forall f_1 \left[ f_1 = \int dh \right] / V, \quad (6)
\]
\[
\forall f_2 \left[ f_2 = \int h \, dg \right] / V. \quad (7)
\]
Then the state diagram of problem (5) will have the following form:

![State diagram](image)

where we have to use some additional rules:
\[
r_4 = \text{multiply } \int dh \text{ by } g; \]
\[
r_5 = \text{multiply } \int h \, dg \text{ by } (-1).\]
Clearly, "g dg" is the initial state of problem (5) and the final state will be obtained from

\[ r_2(r_4(\int dh) + r_5(\int h dg)). \] (8)

Operators \( r_2 \), \( r_4 \), and \( r_5 \) are assumed to be admissible in the frame of real functions \( \mathcal{F} \).

The tree structure of the term (3) reveals the solution method corresponding to problem (5). It is up to the student to find at least one solution method (procedure).

We have now enough motivation for realizing that the set of solution methods, called the solution space \( S \), can be viewed as an extension of the algebra of problems \( P \) by considering an external binary operation \( \ast : C \times S \to S \), interpreted as the application of a concept \( c \) to a problem state \( A \), \( c \ast A \). We will not give the axiomatic definition of the solution space. Notice, that conditions like

(i) \( A + B = B + A \);
(ii) \( A + (B + C) = (A + B) + C \);
(iii) \( c \ast (A + B) = c \ast A + c \ast B \);
(iv) \( (c_1 \land c_2) \ast A = c_1 \ast (c_2 \ast A) \);
(v) \( (c_1 + c_2) \ast A = c_1 \ast A + c_2 \ast A \),

will be valid in \( S \), when \( c, c_1, c_2 \in C \) and \( A, B, C \in S \). Terms of the form (3) are also elements of \( S \).

Let us mention without motivation that the teaching process leads to a similar, in some sense, dual space, called the instruction space \( D \). Here the external operation is an application of teaching strategies to concepts to be taught. \( S \) and \( D \) are the response sets of the learning subject and the teacher (see Fig. 1).

The subjective attribute of the learning subject representing his heuristic abilities, is characterized by the achievement function \( \alpha : S \to [0, 1] \) ([0, 1] stands for the real unit interval). Then the average achievement is given by

\[ A(p, \alpha) = \sum_{s \in \mathcal{J}(p)} P(p)\alpha(s), \] (9)

where \( p \in P \), and \( \mathcal{J}(p) \) is the set of possible solutions corresponding to problem \( p \).

Analogously, we characterize the subjective attribute of the teacher corresponding to his teaching heuristic by the efficiency function \( \varepsilon : D \to [0, 1] \) and the average efficiency is defined by

\[ E(c, \varepsilon) = \sum_{d \in \mathcal{J}(c)} C(c) \varepsilon(d), \] (10)

where \( c \in C \), and \( \mathcal{J}(c) \) is the set of possible teaching strategies corresponding to the concept \( c \).

By a test \( \tau \) we shall understand a sequence of problems \( \tau = \langle p_1, p_2, \)
After having introduced a suitable scaling function $F$, we are ready to define the performance of the student on a test $T$:

$$A(a, \tau) = F(A(a, p_1), \ldots, A(a, p_n)).$$

Similar measure can be used for evaluating the teaching ability:

$$E(\varepsilon, \gamma) = G(E(\varepsilon, c_1), \ldots, E(\varepsilon, c_m)),$$

where $\gamma = (c_1, c_2, \ldots, c_m)$ is an instruction, and $G$ is an appropriate scaling function.

By averaging out the formula (9) over all problems we get the estimate of the knowledge $K$ acquired by the learning subject. Similar averaging of the formula (10) over all concepts gives the estimate of the teaching capacity $T$ of the instructor.

There are several important details concerning the growth of knowledge of the learning subject and the decision making of the teacher which are not included in this paper.

Solving problems and accepting new concepts increases the knowledge of the learning subject. This process is modelled by a tree whose branches grow every time a new concept is accepted or a problem is solved. Teacher, on the other hand, chooses his instruction method on the basis of achievement and performance of the student. The instruction method is structured by a network of pedagogical rules.

Let $A$ be the set of all achievement functions and $E$ the set of all efficiency functions. Then the structure $\langle P, S, A, A \rangle$ completely characterizes the class of learning subjects, while $\langle C, D, E, E \rangle$ represents the class of teachers. Education process is then an information transfer from the system of teachers into the system of learning subjects.

We have completed the list of relevant components of the supervised learning situation. The next task is to interlock these components and construct a model of supervised learning which will reveal useful information about learning and instruction. In order to accomplish this, we have to add to components various functors, representing the mutual interaction of these components: A teaching functor $T$, a solution functor $J$, and a problem functor $K$.

At time instant $t$ a concept $C$ is taught and by its acceptance turned into a fragment of knowledge $T(C) \in K$. Simultaneously a problem $p$ is chosen by $R(C) = p$. Using the amount of knowledge available at time instant $t$, the learner solves the problem $p$ and contributes to his knowledge via the solution functor $J(p)$. Depending on the level of his achievement $P(p) = O(p))$, the teacher turns to the next concept, or poses a new problem, or changes his teaching strategy. A detailed formal description of this process will be the subject of the forthcoming paper (Domotor, 1970).
3. SUMMARY AND CONCLUSIONS

Modeling is a precisely defined subject. It is a scientific method of approaching problems. The basic idea is to discover or to choose a conceptual construct which will reveal useful information about the posed problem, satisfying certain requirements. It is a channel that a scientist chooses or builds in order to communicate with the unknown real life situation. The parameters of the channel (complexity, accuracy) are given, and the task is to construct a least complex channel with the maximal information capacity.

The mathematical model of supervised learning presented in this paper has been developed with the particular purpose of applying the model building technique to social science problems. In the presented model the details of the learning process have been omitted. The structure of the system of knowledge $K$ is described by a mapping $\mathcal{K}: K \times C \times S \rightarrow K$ which assigns to every accepted concept, solved problem, and the old state of knowledge a new, in some sense, bigger state of knowledge:

$$k' = \mathcal{K}(k, \mathcal{I}(c, p)),$$

where $p = \mathcal{R}(c)$, and $k', k \in K$.

In a concrete situation the growth of knowledge is represented by a growing tree of relationships among the accepted concepts which are assigned to nodes of the tree. The idea is somewhat related to Hunt's Information-processing model of learning (Hunt, 1962) and to Grzegorczyk's model of Scientific research (Grzegorczyk, 1964). Also see Scandura (1971a).
A THEORY OF STRUCTURAL LEARNING
DETERMINISTIC THEORIZING IN STRUCTURAL LEARNING: THREE LEVELS OF EMPIRICISM

JOSEPH M. SCANDURA

In spite of the diversity which presently exists in behavioral theorizing, reference to probabilistic notions is all-pervasive. Even support at the .05 level of significance is often enough to elicit whoops of glee from most cognitive theorists. Given this milieu, it is not too surprising that (aside perhaps from computer simulation types and a few competence theorists (e.g., Miller and Chomsky, 1963)), no one seems to have seriously pursued the possibility that deterministic theorizing about complex human learning may actually be easier than stochastic theorizing. And yet, this is precisely what in my own work I have found to be the case.

The purpose of this article is to describe the "rudiments" of a potentially powerful and internally consistent deterministic partial theory of structural learning, which could make it possible to explain, and hopefully also to predict, certain critical aspects of the behavior of individual subjects in specific situations. The term "rudiments" is used because at the present time relatively few implications of the theory have been drawn out. The emphasis so far has been on establishing a fit between behavioral reality and the basic constructs and hypotheses of the theory.

As suggested by the title, there are really three different partial theories, each of which must be tested in a different way. First, there is a theory of structured knowledge—or, more accurately as we shall see below, theories of structured knowledge. These theories deal with the problem of how to characterize knowledge. (The knowledge had by any given individual constitutes a theory in its own right.) Second, there is a theory of idealized behavior which tells how knowledge is selected for use, and how it is learned. This theory applies only where the subject is unencumbered by memory or by his finite capacity to process information. The third theory is still more general and tells what happens when memory and information processing capacity are taken into account. These three theories are not independent of one another, although, as we shall see, research on any one can progress independently of the others and this includes empirical testing.

1. PRELIMINARY OBSERVATIONS

Before describing these partial theories, some general background may be helpful.
There are three main ideas which my title conveys. The label "structural learning" sets the whole tone for the title, so we consider that first. Structural learning refers to the knowledge a person may have and the behavior (and learning) which this knowledge makes possible. More specifically, structural learning is concerned with complex human learning and behavior which cannot naturally be studied without giving explicit attention to what the subject knows before he enters the learning or behaving situation. Any attempt to study mathematics learning, for example, with reference only to the stimulus situation would be folly to the nth degree. Individual differences in prior knowledge and other intellectual skills in mathematics may be very great indeed, and these differences must be taken explicitly into account in any theory that is to provide a viable account of complex mathematics learning. It should be noted parenthetically that one of the primary requisites for selecting tasks in most traditional studies has been that prior learning be of minimal importance. The reference here, of course, is to experiments on serial and paired-associate learning, classical conditioning, and the like.

Dependence on prior knowledge, then, is important to my conception of structural learning. But this alone is not sufficient. The knowledge involved must also have a reasonably clear structure. In this sense, mathematics, for example, tends to have a clearer structure than, say, the social studies or the humanities. The fact that grammarians, like Harris and Chomsky, have been able to make as much progress as they have in linguistics attests to a good deal of structure in language as well.

The second dominating phrase in my title is "deterministic theorizing". In view of the tradition in psychology against this type of theorizing, it is instructive to consider the paradigm most typically used in testing behavioral theories. First, assumptions are made about how individuals learn or behave. When stated in their clearest form, as in the stochastic theories of mathematical psychology, the basic assumptions are stated in terms of probabilities. Second, inferences are drawn from these assumptions yielding predictions about group statistics—that is, about characteristics of the distributions of responses made by the experimental subjects. Third, on the basis of the experimental results obtained, inferences are made about the basic assumptions.

Of course, there is no harm in this as long as it is recognized that the initial assumptions deal with probabilities and not with individual processes. But this fact has not always been made as explicit by theorists as might be desirable. What needs to be made clear with such probabilistic theories is that what any given subject does on a given occasion may have little or nothing to do with the particular assumptions made. For example, in stochastic models of paired-associate learning it is usually assumed that each subject has the same probability of learning on each trial. Even the most superficial analysis of relevant data, however, indicates clearly that the probability of success for different subjects may vary greatly. And one cannot attribute this to the fact that the probability of learning is a random variable. This would still not explain the fundamental fact that the probability of success of many subjects tends to be either uniformly high, or low, over different trials.

How much better it would be to have a theory which would tell us explicitly what a given subject will do on specific occasions—a theory which leaves errors in prediction to inadequacies in observation and measurement, and does not make these errors an explicit part of the theory itself. Ideally, such a theory would satisfy the classical condi-
tions for a deterministic theory in the hard sciences—theories which say, in effect, that given such and such basic hypotheses and these initial conditions, this is what should happen. Given a theory of this sort, probability would enter only where one wanted to make predictions in relatively complex situations where the experimenter practically speaking could not, or did not wish to, find out everything he would need to know and specify in order to make deterministic predictions. In effect, a truly adequate deterministic theory would make it possible to generate any number of stochastic theories by loosening one or another of various conditions which must be satisfied in order for the deterministic theory to apply. (In this regard, see the comments below on levels of empiricism and conditional hypotheses.)

In order to be completely honest, I must mention one further reason why deterministic theorizing appeals to me. I am basically lazy. I have done a good deal of traditional behavioral research, but I dislike with a passion poring over reams of raw data or computer printouts, especially when I know that, no matter what statistics are used to summarize the data, I am losing much, if not most, of what is important. It is perhaps this distaste as much as anything else which has moved me to search for a new and better way to do empirical research on complex human learning. How much nicer to have data which is clearcut, no means or variances to compute, no analyses of variance, or canonical correlations, or factor analyses—just looking. In this regard, I can't resist the temptation to repeat a little story about an experience I had as a post-doctoral student being initiated into mathematical psychology at Indiana University. The time was the summer of 1962, and the field was bright and promising. As part of my orientation, I was routed about to visit a number of the more prominent names on campus, including one very fine physiologist. Caught up by the emphasis on mathematics given by the psychologists, I asked him what kinds of mathematics he found most useful in his work, and how he used it. His answer was, "We count." After getting over my initial shock, I began to see the logic of his answer, and have been trying to meet his ideal ever since.

Finally, let us consider what is meant by "levels of empiricism". Recall first that any theory is but a partial model of reality. It deals adequately with certain phenomena in the sense of providing an adequate explanation for them, but not others. Theories do not apply universally. To make the point in its most trivial sense, we need only note that existing theories of thermodynamics, for example, are not likely to be very useful in explaining paired-associate learning—or vice-versa. As a more realistic example, learning theories such as Hull's provide a far better account of certain simple behavioral phenomena than they do, for example, of the learning of complex mathematical structures. (Partial theories must not be confused with so-called miniature theories of mathematical psychology. Partial theories deal with only certain phenomena of given, broad-based realities. Miniature theories deal intensively with highly restrictive phenomena such as paired-associate learning.)

The general difficulty with most theory construction in psychology, today, is that very little attention has been given to specifying conditions under which theories are not presumed to hold. To date, the sole approach to this problem has been an ad hoc empirical one in which experimental evidence is gradually accumulated over relatively long periods of time.
It is my feeling that much can be done along these lines, while theories are actually being constructed. This does not obviate the need for empirical testing, of course. No one believes that we can ever do away with that. But I do think that we can do away with a good deal of it, if theorists would give more explicit attention in their work to identifying these negative conditions.

In constructing a theory, whether it be a mathematical theory or a scientific theory, the theorist has some model, or models, in mind at the time. These models arise basically from particular segments of reality—but more important here, they usually deal with only certain aspects of that reality. The rest is simply ignored.

This approach may be a viable one in mathematics, where one aims for abstraction. One never knows where mathematical theories may ultimately prove useful (i.e., be applied), and it would undoubtedly be a mistake to tie them in too closely to any particular model, by specifying aspects of these particular models with which the theory does not deal.

This is not true in science, however, where the ultimate aim may be to devise theories which deal with more of the particular reality in question. A theorist may have many more kinds of phenomena in mind in attempting to construct a theory than he can possibly handle at one time. To get around this problem, he may purposefully ignore for a time certain of these phenomena to facilitate constructing what might be called a partial theory—a theory which deals with part of the reality but not all of it.

In constructing such a partial theory, it is critically important that the theorist do so in a way which is compatible with the broader reality. Thus, for example, the ultimate aim of competence theorists such as Chomsky (1968) and Miller and Chomsky (1963) is not just to characterize the knowledge had by an idealized human subject—that in itself might be attempted in any number of different ways. What these theorists want is a theory of knowledge which is likely to be compatible with a more encompassing behavior theory once one is developed (e.g., see Miller and Chomsky, 1963, 433-488). In such cases, it will generally be in the theorist's interest to know just what aspects of reality his present theory does not consider. Stated differently, he must know what boundary conditions must be satisfied in order for his partial theory to apply. Theoretical predictions based on partial theories are necessarily dependent (on such conditions).

In order to test a partial theory, then, the empirical situation must accurately reflect these boundary conditions. Otherwise, the partial theory will simply not be applicable—by definition. Perhaps the best known example has to do with linguistics, where grammarians, such as Chomsky (1957), assume an idealized knower—a knower who can use whatever rules are attributed to him without error, and wherever they might be needed. This type of theory seems to be having increasingly important implications for psychology, but it must be remembered that a competence theory of this sort applies only in those situations where the idealized performer assumption is reasonable to make. (There is a close relationship between these ideas and the so-called ecological approach to behavioral science [Wohlwill, 1970], which is becoming increasingly popular of late. In fact, the partial theories described below provide good examples of the kind of theories for which this approach seems to call.)
2. FOUNDATIONS OF A THEORY OF KNOWLEDGE

The first level of theorizing is concerned with the problem of how to account for the behavior of idealized subjects. More particularly, given a finite class or corpus of behaviors, the problem is one of how to characterize the knowledge underlying the corpus in a way which accounts as well for the other behaviors of which an idealized knower of that corpus may be capable. Our approach to this problem involves the invention of a finite set of rules of one sort or another which can be used to generate not only the behaviors in the given corpus, although this is an absolute minimum, but also the other behaviors one might wish to attribute to the knower (Scandura, 1971b, for an earlier but closely related version of this goal see Chomsky, 1957). (A rule may be said to account for a class of behaviors if, given any stimulus input associated with the class, the corresponding response may be generated by application of the rule (Scandura, 1963, 1970b).)

As one might suspect, there are any number of different ways in which to characterize the same given corpus. The theoretical problem is one of evaluating these various characterizations to determine which best accounts for the other behaviors one might wish to attribute to the knower (Chomsky, 1957; Scandura, 1971b; Volume I). These additional behaviors constitute the predictions.

Consider some of the alternatives. Undoubtedly, the simplest way to account for a given finite corpus is just to list the behaviors involved. Thus, for example, a list of paired-associates might be characterized as a finite set of degenerate rules (Scandura, 1968) or, equivalently, as a finite set of associations. Clearly, lists of paired associates are not the sort of corpora we usually have in mind in talking about mathematical and other complex behavior, and characterizations which consist of simple lists of associations would be essentially sterile in content. If this were all a person could learn, it would be impossible even to learn how to add numbers, addition fact by addition fact. A person could learn at most a finite number of sums, since each addition fact (e.g., $3 + 5 = 8$, $25 + 47 = 72$, and so on) would have to be learned separately.

A somewhat more realistic characterization of a corpus of behaviors derives from recent attempts in educational circles to define school curricula in terms of a finite number of operational objectives (e.g., Lipson, 1967). Each of the objectives of these curricula amounts to a class of behaviors which can be generated by a rule; the abilities to add, to multiply, to find areas of triangles, and so on, provide obvious examples. It is possible to account for the behaviors represented by such a corpus, then, by simply listing a finite set of rules. In fact, this is essentially what has been done by curriculum constructors who have followed this approach. The curricula consist essentially of long lists of rules for achieving the (operational) objectives, one rule for each objective.

Clearly, exactly the same idea might be applied in characterizing the knowledge had by individual subjects. A list type of characterization of this sort would have the major advantage of requiring a very simple performance mechanism. Thus, if knowledge is characterized as a list of discrete rules, which operate independently of one another, then a more general theory of performance would need to tell only how such rules are put to use. Since the rules are discrete, no interactive mechanisms need be postulated.
This advantage, however, is also its major disadvantage. Because the characterizing rules are discrete, they cannot account for behaviors which go beyond the given corpus, except in the most trivial sense. For example, suppose the characterization only included rules for adding, subtracting, multiplying, and dividing. In this case, the subject would be unable to even generate the addition fact corresponding to a given subtraction fact, although one might reasonably expect this type of behavior from a person who was well versed in arithmetic. One might counter, of course, that it would be a small thing simply to add a new rule to the original list.

\[ c - a = b \implies a + b = c \]

We might even use the distinguishing label "relational rule" since it operates on the elements of a binary relation. Indeed, this is precisely the sort of reply one might expect from curriculum constructors of the operational objectives persuasion. When confronted with the criticism that their objectives do not constitute a mathematically (or otherwise) viable curriculum, they would simply say we can add more objectives.

The trouble with this sort of argument is that it misses the point entirely. Not only would such an approach be ad hoc—which really says nothing by itself except to convey some ill-defined dissatisfaction—but it would be completely infeasible where one is striving for completeness. To see this, it is sufficient to note that a new rule would have to be introduced for every conceivable interrelationship, and that the number of such interrelationships is indefinitely large. One could easily envision a number of rules so large that no human being could possibly learn all of them. There would not be sufficient time in a single lifetime. The sum total of all mathematical knowledge which is presently in print, for example, is so vast that no one has, or could, possibly acquire all of it. As vast as this knowledge is, however, a really good mathematician is capable of generating any amount of new mathematics which does not appear in print anywhere. That is, he can create. Much of the new mathematics might be utterly trivial, of course, but the very fact that it exists at all strongly suggests that any characterization such as that described above would almost certainly miss much that is important.

We can get a far more powerful and simple characterization by allowing rules to operate, not just on ordinary stimuli, but on other (lower order) rules as well. More specifically, allowing rules to operate

\[
\text{Higher order rules on rules are common in various branches of mathematics where they go under the label of functions on functions, but the idea seems not to have generally pervaded either computer science or formal linguistics. In formal linguistics, for example, where the goals closely parallel ours, no one seems to have seriously proposed the use of higher order rules. The closest linguists have come in this regard has been to introduce the notion of grammatical transformation between phrase markers (Chomsky, 1957). Rather than higher order rules, transformations correspond more closely to what we have here called relational rules (see Scandura, Volume I).}
\]

There are two good reasons why this has probably not been done in the past. First, even grammatical transformations have so far resisted mathematical treatment (Nelson, 1968); and, second, no existing approach to psychology known by the writer provides any real motivation for introducing them. Gagné's (1965) view on problem solving, or rules on rules, and Miller,
in this way makes it possible to generate new rules and these rules, in turn, may make it possible to generate what might appear to be completely different kinds of behavior. For example, suppose that an idealized knower has mastered the two rules:

\[(1) \quad a, b \rightarrow a + b \]
\[(2) \quad [x, y \rightarrow x \circ y] \Rightarrow [x, y \rightarrow x \circ' y],\]

where (1) represents a rule for generating sums of pairs of, say, integers and (2) represents a (higher order) rule which, given a rule of the form (1) for any binary operation, generates a rule for performing the corresponding inverse operation (denoted \(o'\)). Such a rule would connect, for example, not only addition of numbers with subtraction, but composition of all sorts with the corresponding inverse operations, whether these operations involve permutations, rotations, rigid motions, or whatever. In this case, application of rule (2) to rule (1) yields rule

\[(3) \quad a, b \rightarrow a - b,\]

where "-" is the inverse of "\(+\)". Application of rule (3), in turn, makes it possible to generate differences between any given pair of integers \(a\) and \(b\) where \(a > b\). But, then, isn't this just a simple instance of the sort of thing we have in mind when we think of creative behavior? If the extrapolation involved seems too tame to qualify for this distinguished label, consider the following example in which we add another level to the analysis. In this case, we assume in addition to rules (1) and (2) that the idealized knower has also mastered rules,

\[(4) \quad [x, y \rightarrow x \circ y] \Rightarrow [x, y \rightarrow x \circ y] \quad \text{(Note: } x, y, o \text{ are different from } x, y, o_{\text{'}}, \text{ respectively.)}\]
\[(5) \quad [(x \rightarrow y), (y \rightarrow z)] \Rightarrow [x \rightarrow z]\]

Rule (4) may be thought of as denoting knowledge of generalized homomorphic relationships between pairs of systems such as the system (A) of integers under addition and, say, the system (B) of rational numbers under addition. Rule (5) is extremely general and makes it possible to generate the composite (rule) of any pair of given rules such that the output of one of the rules serves as the input of the other.

Knowing these rules would make all kinds of behaviors possible. For example, the idealized knower would be able to subtract, not only in the first system (A) but in the second system (B) as well. To see this, we need only observe that application of rule (5) to rules (4) and (2), yields rule

\[\text{Galanter, and Pribram's (1960) TOTE hierarchies come close, however.} \]
\[\text{This is unfortunate, since there is a very simple and intuitively sound reason for allowing rules to operate on (classes of) rules. The main one is just this: There is a very simple and intuitively compelling performance mechanism by which higher and lower order rules may be combined so as to generate completely new kinds of behavior. Furthermore, as shown in the next section, some empirical support for this mechanism has already been obtained.}\]
(6) \([x, y \rightarrow x \circ y] \Rightarrow [x, y \rightarrow x \circ' y]\)

Application of rule (6) to rule (1), then, yields rule

(7) \([a, b \rightarrow a +' b]\) or \([a, b \rightarrow a - b]\) where \(+' = -\).

Rule (7) is the subtraction rule for system B. The basic relationships are represented schematically in Figure 1.

Fig. 1. A schematic representation of the basic relationships described in the text. Solid arrows refer to (pre)learned rules and dotted arrows to derivable rules. Rule (5) by which rules (4) and (2) are combined to give (6) is not represented since this would require a third dimension and would complicate the diagram without adding any clarification.

More details and further examples may be found in Chapter 6 of Volume I.

In summary, the essentials of the theory of knowledge as outlined are just these. (1) The knowledge of any given individual at any given
stage of learning can be characterized in terms of a finite set of rules. This implies among other things that there may be as many different theories of knowledge as there are individuals—or, equivalently, as many theories as there are conceivable curricula to be mastered. (2) Rules may act on classes of rules as well as on simple stimuli. Allowing rules to act in this way amounts to a simple but conceptually major revision of existent competence theories. (3) For purposes of the theory, it is assumed that the rules may be combined at will and without error as needed. Stated differently, the idealized knower is assumed to have mechanisms available for putting the rules attributed to him to use.2

3. FOUNDATIONS OF AN IDEALIZED THEORY OF STRUCTURAL LEARNING

The third point above is a critical boundary condition of the theory of knowledge. The theory applies only at the analytical level in the sense that generative grammars account for language behavior. The relevance of the theory to actual human behavior is dependent on our ability to spell out mechanisms which are both adequate to account for how rules may be combined and which are reflected in the actual behavior of human subjects.

It is to this task that we now turn—the task of introducing mechanisms of idealized performance, learning, and motivation into our formulation. The purpose of adding such mechanisms to the theory of knowledge is to obtain an extended theory which deals explicitly with the way in which available knowledge is put to use. This more encompassing theory is still a partial theory, however, one which applies only where subjects are unencumbered by either memory or their intrinsically limited capacity to process information. It should be emphasized, however, that it is a theory which is assumed to apply no matter what knowledge an idealized subject has available. Thus, even though the knowledge had by different individuals may vary greatly, the same theory of idealized behavior is assumed to hold over all individuals.

The basic assumption on which this theory rests is that people are goal-seeking information processors. In this case, much of what a subject knows becomes irrelevant once a goal situation is specified. Thus, at any given point in time, only a small fraction of the rules available to a knower may be applicable—namely, those rules which may be used directly or indirectly in satisfying the given goal.

There are three basic kinds of situation with which any viable theory must deal. One type of situation is where the subject knows one or more rules which apply in the given goal situation. The second is where the subject does not explicitly know a rule which applies in the goal situation. The third is actually a refinement of the first, and deals with the question of why, when a subject has more than one rule available, he selects the rule that he does. Why not one of the others? As we shall see, these problems are closely allied with what have traditionally

2There will always be behavior, of course, which cannot be generated by any given finite set of rules. Roughly speaking, when translated into behavioral terms, Gödel's (1931) Incompleteness Theorem suggests that no matter how bright an individual, there will always be certain behaviors he will not be capable of performing.
been called performance, learning, and motivation, respectively.

The first case is simplest to deal with. We need only assume that:

(A) Given a goal situation for which a subject has at least one rule available, the subject will apply one of the rules.

Thus, for example, if a subject's goal is to find the sum of two numbers, and he knows how to add, then he will actually use an addition rule.

As trivial an assumption as this may appear, it is an assumption. It does not follow logically that just because a subject wants to achieve a certain goal and has one or more rules available for achieving it, that he will necessarily use one of them.

Furthermore, the assumption has a number of important implications. One of these is that it provides an adequate basis for determining what might be called a subject's behavior potential, relative to a given class of rule-governed (RG) behaviors. It may be noted in this regard that it is one thing to devise a procedure (rule) which accounts for a given class of RG behaviors and quite another to identify that subclass of behaviors of which a given subject is capable. The first problem is an analytical one and involves inventing a procedure which accounts for the given class of RG behaviors. No psychological assumptions are involved.

Determining a subject's behavior potential, however, necessarily depends on what can be assumed about the mechanisms which govern human behavior. The basic idea goes like this: Given any familiar class of RG behaviors, like the class of addition tasks, we can usually identify those rules (algorithms) which the subjects in question are likely to use in solving the problems. We do not automatically know which aspects of these algorithms any given subject is capable of, however. To find out, we must test the subject. But on which instances is he to be tested --how are they determined? The standard approach, of course, is just to select a random sample of test instances and then make probabilistic predictions about future performance on other instances in the class.

This approach is rejected in favor of systematic selection of test instances and deterministic prediction on individual items. To see how this can be accomplished, we first note that every algorithm for solving a given class of (RG) tasks can be represented by a directed graph (see Volume I, Chapter 2). For example, the task of generating the next numeral in Base Three Arithmetic can be represented as follows.

In Figure 2, the arcs correspond to rules which are assumed to act in atomic fashion. That is, success on any one instance of such a rule is tantamount to success on any other, and similarly for failure. We have obtained sufficient empirical evidence over the past seven years to demonstrate the existence of such rules in a wide variety of situations (e.g., Scandura, 1966d,1969a). The points correspond to branching rules, that is, decisions which must be made in carrying out the algorithm on particular test instances.
Sample

Stimuli → Responses
0 → 1
1 → 2
2 → 10
10 → 11

Total Graph

Paths (Subgraphs)

Fig. 2. Sample stimuli and responses for the task of generating the next numeral in Base Three Arithmetic, together with the (total) graph of a procedure for generating the behavior, and four graphs representing the four behaviorally distinguishable paths through this procedure.

The subgraphs at the bottom of Figure 2 correspond to the four possible paths through this procedure which may be used in solving particular problems. Since the constituent rules are all atomic, it follows that each of these paths also acts in atomic fashion. Hence, to determine the behavior potential of a given subject with respect to this algorithm, we need only select one test instance for each path. In this case, the base-three stimulus (response) numerals 101 (102), 2 (10), 112 (120), and 222 (1000), correspond respectively to the four possible paths. Accordingly, the behavior potential of a given subject on this class of tasks can be uniquely specified by his performance on just these four test instances—as long as the atomic assumption is valid. (Hence, the assessment is conditional.) Any other set of four stimulus representatives of these paths, of course, would do equally well. Although its role was hidden in describing this method of assessing behavior potential, the methods' validity depends directly on the simple performance mechanism. According to this mechanism, if a subject has a particular path available for solving a given task, then he will use it and use it consistently on all instances to which it applies. That is, of course, assuming that the subject's goal remains the same.
None of this is idle theoretical speculation. Over the past several months one of my students, John Durnin, has collected a good deal of evidence which provides support which goes far beyond the bounds of what is normally considered sufficient evidence. In a total of 204 predictions, utilizing a variety of tasks and subjects of greatly differing abilities and grade levels (from the preschool through graduate school), we have had a grand total of seven errors in prediction. A sample of this data is given in Table 1 for a procedure involving eight paths.

Table 1

<table>
<thead>
<tr>
<th>Paths</th>
<th>College Student Test</th>
<th>College Student Test</th>
<th>High School Student Test</th>
<th>High School Student Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
</tr>
<tr>
<td>2</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
</tr>
<tr>
<td>3</td>
<td>+  +</td>
<td>+  +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
<td>+  +</td>
</tr>
<tr>
<td>5</td>
<td>+  +</td>
<td>+  +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>+  +</td>
<td>+  +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>+  +</td>
<td>+  +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>+  +</td>
<td></td>
<td>-  +</td>
<td></td>
</tr>
</tbody>
</table>

Note—"+" indicates correct response.
"-" indicates incorrect response.

Let us next consider what happens when a subject has not explicitly learned a rule for achieving a given goal. In this case, the subject has a problem in the classical sense—a problem situation, a goal, and a barrier between them.

The major theoretical problem is to explain what happens when a subject is confronted with such a situation. If the problem can be formulated in a way that lends itself to prediction, so much the better. Why certain people are able to solve some problems for which they have never learned a specific rule, whereas others cannot, is a question of paramount interest. We want to know exactly what is involved, and why subjects perform as they do.

As a first approximation at least, it again appears that a very simple mechanism may suffice. This mechanism may be framed as a hypothesis as
(B) Given a goal situation for which the subject does not have a learned rule immediately available, control temporarily shifts to the higher order goal of deriving a procedure which does satisfy the original goal condition.

With the higher-order goal in force, the subject presumably selects from among the available and relevant higher order rules in the same way as he would with any other goal. Furthermore, where no such higher order rules are available, one might suppose that control would revert to still higher order goals. Theoretically, this process could continue indefinitely, but I suspect that a subject would tire of it, or run out of higher order rules, as quickly as would a theorist attempting to describe what is happening.

To complete things, we need a third hypothesis which allows control to revert back to the original goal once the higher order goal has been satisfied. We can state this as follows:

(C) If the higher order goal has been satisfied, control reverts back to the original goal.

When we say that a higher-order goal has been satisfied, of course, what we mean is that some new rule has been derived, such that that rule, when applied to the stimulus situation, satisfies the original goal criterion.

Although implicit in what has been said, it is important to note that each of the hypothesized mechanisms is assumed to work at all levels. For example, hypothesis (A) applies in higher order goal situations as well as in simple ones.

These assumptions provide an adequate basis for generating predictions in a wide variety of problem solving situations. Suppose, for example, that the problem posed to a subject is to convert a given number of yards into inches. Consider two possible ways in which a subject might solve the problem. The first is to simply know, and have available, a rule for converting yards directly into inches: "Multiply the number of yards by thirty-six." In this case, the subject need only apply the rule according to hypothesis (A). The other way is more interesting, and involves all the mechanisms described above. Here, we assume that the subject has mastered one rule for converting yards into feet, and another for converting feet into inches. The subject is also assumed to have mastered a higher order rule which allows him to combine learned rules (in which the output of one matches the input of the other, as is the case, for example, with rules for converting yards into feet and feet into inches) into single composite rules.

In a situation of this sort, the subject does not have an applicable rule which is immediately available, and, hence, according to hypothesis (B), he automatically adopts the higher order goal of deriving such a procedure. Then, according to the simple performance hypothesis (A), the subject selects the higher order composition rule and applies it to the rules for converting yards into feet and feet into inches. This yields a new composite rule for converting yards into inches. Next, control reverts to the original goal by hypothesis (C) and, finally, the subject
applies the newly derived composite rule by hypothesis (A) to generate the desired response. This sequence of events is depicted in Figure 3.

![Diagram of hypothesis mechanism](image)

**HYPOTHESES**

1. \( H_R \) (applied to \( R_1 \) & \( R_2 \)).
2. \( \text{(A)} \) (applied to composite of \( R_1 \) & \( R_2 \)).

Fig. 3. A schematic representation of the hypothesized mechanism for problem solving. \( R_1 \) and \( R_2 \) represent rules for converting yards into feet and feet into inches, respectively. \( H_R \) refers to the higher order rule for generating composite rules.

Although we are still in the process of refining our procedures and collecting more data, Lou Ackler and Chris Toy have run enough subjects under one condition to suggest that we are on the right track. What we did was to teach each subject how to use two simple rules, comparable to those described above (e.g., for converting yards into feet). These rules are denoted \( r_{11} \) and \( r_{12} \) in Table 2. As shown in the table we were successful in teaching these rules to all of the children in the sense that they could apply them uniformly well on all instances (of the respective rules). Then, each subject was tested to see if he could solve a problem requiring for its solution the composite rule, denoted \( r_{11} \circ r_{12} \). As shown, only one of the subjects was initially successful on this type of problem. Next, we taught the subjects with neutral materials how to combine pairs of simple rules such as the ones they had been taught. This time we were successful with all but one subject. (To accomplish this we also had to teach many of the subjects what it was they were trying to do—that is, find a rule which could be used to solve problems such as that requiring \( r_{11} \circ r_{12} \) above. In short, we taught them a decision making capability for determining whether or not they had achieved the higher order goal. More details on this are given in Scandura, Volume I.)

At this point, we taught each subject a new pair of rules (indicated by \( r_{21} \) and \( r_{22} \)) and then tested him to see if he could solve the corresponding composite problem, which required \( r_{21} \circ r_{22} \) for its solution.
TABLE 2

Summary of Experimental Procedure and Results

<table>
<thead>
<tr>
<th>Age of Subject</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>5</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>8</th>
<th>6</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_{11} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{12} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{11} \circ r_{12} )</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HR</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{21} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{22} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{21} \circ r_{22} )</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{31} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{32} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{31} \circ r_{32} )</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{41} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{42} )</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>( r_{41} \circ r_{42} )</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: "+" indicates S reached criterion.
"-" indicates S did not reach criterion.
"•" indicates that S was tested only (without training).

As can be seen in Table 2, all but one of the subjects succeeded on the test problem whereas only one of them had before. Furthermore, the one subject who failed was the same subject who had previously failed to learn the higher order rule when it was taught. This same pattern of teaching and testing was repeated two more times as shown, with precisely the same results.

While no empirical data are available, it has been possible to analyze a number of other, more complicated problem situations in very much the same way (see Volume I), including problems taken from Polya's (1962) pioneering yet atheoretical discussion of mathematical problem-solving. This includes taking into account the role of heuristics. A very similar type of analysis can also be applied to discovery learning, and, indeed, even to simple association learning (see Volume I). The
situation is very much like problem solving in which there are a number of simple problems presented in sequence, rather than just one. It would be misleading to imply, however, that this is a routine undertaking. To the contrary, it seems to require a good deal of experience, familiarity with the subject matter, and good intuition about how Ss actually do things. Most important, it usually takes time to come up with a viable analysis. Nonetheless, I am satisfied that this can be done in principle; what remains is to test these analyses empirically to see if the three hypotheses introduced above are sufficient to account for the performance of actual Ss (under the idealized conditions required by the theory).

The important point of all this is that learning can be viewed as a problem-solving process. Subjects learn as a result of being exposed to problem situations which require that they combine available rules in new ways. Once a problem has been solved, however, no further learning is assumed to take place upon repeated presentations of similar problems. In that case, the subject simply applies the newly learned rule.

By systematic application of our simple principles (of performance), then, it would be possible to derive all kinds of implications about learning and performance. In particular, highly specific predictions might be made about individuals who enter the learning situation with given sets of rules and who are then subjected to particular sequences of problem situations. Such analyses would have obvious implications for instructional theory. (It must be remembered, of course, that all such predictions would necessarily be limited to empirical situations which satisfy the conditions of level two theorizing.)

Suppose now that a S has more than one way of achieving a given goal and that we want to know which way he will choose. As suggested above, this problem of rule selection is basically one of motivation. To see this, we ask what theorizing about motivation involves, and how this relates to our earlier discussion. We might be tempted to define the task of motivation theory as one of explaining and/or predicting which goals subjects will adopt in given situations and let it go at that. This would not be sufficient, however, for that would not tell us where such goals come from in the first place, nor how they relate to the situation at hand.

In any given situation, the observer almost always has some idea of what a given S is trying to accomplish. Thus, for example, he may not know what sort of building an architect will design, but he can be quite sure that it will be a building, under certain circumstances at least. Similarly, he can usually be fairly certain that the next move made by a chess master will be a good one, although he may not know what the specific move will be. He can also be reasonably confident that, faced with a simple theorem, a competent mathematician will come up with a valid proof, but generally speaking, he will not know what kind of proof it will be. An analogous statement may be made about a competent fifth-grader on simple addition problems. The observer may not know, say, how quickly the sums will be given, but he will generally know that they will be correct (cf. Suppes and Groen, 1967).

Looked at in this way, the motivation theorist's task is to say something additional about what a S will actually do in any given situation, whether this involves explaining why the architect designed the building he did, why the chess master made his particular move, or why the mathematician used an indirect proof, or the child, a certain shortcut
in addition. More generally, the key question for motivation theory is to explain (and/or predict) why the $S$ took (will take) the path he did (does).

The problem comes in where the $S$ has more than one rule available for achieving the initial goal. It was assumed in this case that the $S$ would use one of the available rules (Hypothesis (A)), but nothing was said about which one. It is my contention that the answer to this question of "which one" lies at the base of what we normally think of as motivation, especially as it is realized in structural learning and performance.

Unfortunately, space does not permit anything approaching the discussion which this problem warrants. (The problem is discussed at length in Scandura (Volume I).) For present purposes, it is sufficient to assume that $S$s are systematic in their selections. I do not believe that people are intrinsically unpredictable, even in so complex a field as motivation.

If this is true, it would seem that perhaps one could gain insight into what a person might do in the future on the basis of what he has done in the past. But, then, do not we do just this almost every day? With experience, we begin to sense the way in which particular people are likely to behave in given situations, and may therefore tend to act accordingly. We frequently know ahead of time, for example, how the boss will react to a request for a raise, or what kind of activity Janie will engage in during free play, or what kind of homework will be left undone until last.

The task of the motivation theorist is to translate such intuitions into empirically testable hypotheses. A doctoral student, Francine Endicott, and I have been working on this problem for several months now, and at first we were not particularly pleased with our results. To be sure, the data almost always supported our hypotheses in a gross probabilistic sense, but they could hardly be called deterministic. By using past selections as a guide, we have been able to do much better and have recently obtained an accuracy rate of about 35% correct predictions. What we did in these experiments was to provide each $S$ with an opportunity to learn two distinct procedures (Rules A and B) in the same manner as was done in the assessment (of behavior potential) study. The stimuli were identical but the responses generated by the two procedures, could easily be discriminated. After learning both procedures each $S$ was presented with a general goal, which could be satisfied by using a path of either procedure. For testing purposes, stimuli on which $S$ had precisely the same choice to make between paths were viewed as equivalent. As in the assessment study, $S$ was tested twice on each equivalence class. According to our assumption, it was hypothesized that $S$ would select the same paths on corresponding Test 1 and Test 2 stimuli.

The results are summarized in Table 3. This table shows that whenever a $S$ selected a path of Rule A on Test 1, he almost invariably (52 times out of 54) selected a path of Rule A on Test 2. Rule B selections were consistent with the hypothesis 64 times out of 77.\(^3\)

\(^3\)A generalization of the memory-free theory is reported in Volume I in which motivation and learning (problem solving) are governed by essentially the same basic mechanisms. Data are reported which show that
TABLE 3

Results of Rule Selection Study

<table>
<thead>
<tr>
<th>Test 1</th>
<th>Rule A</th>
<th>Rule B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule A</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td>Rule B</td>
<td>13</td>
<td>64</td>
</tr>
</tbody>
</table>

To recapitulate, it should be re-emphasized that everything which has been said so far about learning, performance, and motivation only applies in situations where memory and the limited capacity of human subjects to process information do not enter. The proposed mechanisms have all assumed an information processor with an essentially unlimited ability to process information, and with perfect memory for previously acquired knowledge.

This definitely does not imply that the theorizing so far is of little value. That conclusion would be wrong on at least two counts. First, there are many practical situations in structural learning where memory is of minimal concern. In problem solving, for example, the S is almost always given all of the paper, pencils, and other memory aids that he needs. Typically, we also do our best to insure that the necessary lower-order rules are readily available, even to the extent of making textbooks available. The concern is generally with whether or not the individual can integrate available knowledge to solve problems. Considerations such as whether he can do it in his head or not, time to solution, and so on, are of secondary concern (cf. Scandura, 1967f). Second, questions of memory can usually be eliminated in experimentation by insuring that relevant rules and memory aids are available to the subject. This can normally be accomplished by training.

4. TOWARD A THEORY OF MEMORY AND INFORMATION PROCESSING

Any fully adequate theory of structural learning, of course, must deal with more than just idealized behavior. In particular, such a theory must as a minimum take both (long-term) memory and information processing into account. Insofar as memory is concerned, there must be mechanisms for storing and retrieving information in long-term memory. In addition, hypotheses are needed to deal with the processing of information, and particularly the limited amount of information which human beings can process at any given time. Thus, for example, an adequate theory should make it possible to account for the differential ability of human subjects to perform mental arithmetic, even where the Ss know perfectly well how to compute.

In most theorizing about memory, there has been an unfortunate tendency to confound these two kinds of problem. Much of the more recent selections made with one kind of task may be used to predict selections with completely different tasks.
work, for example, has been heavily influenced by a new technique for measuring retention, which was introduced by Peterson and Peterson (1959). The basic idea of their experiment was (1) to present CCC nonsense syllables (2) have the S count backward by threes or fours, and (3) test him to see if he could remember the given nonsense syllable after some intervening period ranging from about 0 to 30 seconds. Contrary to the then prevailing expectation of most psychologists, they found that retention decreased rapidly over this short period, and psychologists had a new game to play. The basic paradigm is still in wide use today.

The difficulty with this type of study is that it does not distinguish operationally between mechanisms associated with the storage and retrieval of information from long-term store, on the one hand, and the limited ability of human Ss to process information, on the other. Thus, in a Peterson and Peterson type experiment, a S may attempt to retain a nonsense syllable, say, either by continuing to process the information, by a process typically referred to as rehearsal, or by storing it in long-term memory. Under these circumstances, it is difficult to say anything definitive about either type of mechanism as a result of the experimental data obtained.

For present purposes, it would obviously be desirable to have a theory of structural learning which deals with the two kinds of problems raised above, and which at the same time is compatible with our earlier theorizing. Specifically, we need to ask how the memory-free theory may be supplemented so as to take both (long-term) memory and information processing into account. No hard answers, unfortunately, are available at the present time, particularly insofar as memory is concerned. All that can be done here is to sketch one approach to the problem which may hold some promise. Another (related) approach is sketched in the next article.

Insofar as long-term memory is concerned, nothing basically new seems to be required; the basic mechanisms of the idealized theory appear to be adequate as they are. What does need to be done is to increase the domain of applicability of these mechanisms. Specifically, rules are needed for storing and for retrieving information. Storing rules act on observables, as do other rules, but the outputs of such rules are strictly internal. Retrieving rules, on the other hand, act on stored (internal) units of knowledge (which serve as stimuli) and generate observables.

What these rules do is to relate new knowledge with knowledge which has been acquired previously. For example, in order to store (i.e., give the correct meaning to) the statement, any function continuous on a closed interval is uniformly continuous, S must clearly know ahead of time what continuous functions, closed intervals, and uniformly continuous functions are. The storing rule combines these meanings into a new meaning which corresponds to the statement, taken as a whole (Scandura, 1970b). This has the effect of tying in (i.e., locating) the desired meaning with previously acquired knowledge.

Retrieval rules, on the other hand, provide the subject with a basis for regenerating knowledge from the recall cues—for example, from a statement like "What can be said about functions which are continuous on closed intervals?"

Difficulties in recall are explained either in terms of what is (or is not) stored or the availability (or lack) of appropriate retrieval rules. For example, if a S memorizes a statement like that given above
without understanding it, and is asked at recall to explain the idea in his own words, then no one would reasonably expect the $S$ to succeed. Similarly, if the $S$ stored the meaning and was asked to repeat the statement verbatim, he would not likely be able to do more than come up with a rough paraphrase. Without adequate storing rules in the first place, of course, recall would be completely lacking according to this view. Even where a $S$ has definitely learned (stored) something, he may still not be able to "recall" it because he lacks the necessary retrieval rules. Young children, for example, are frequently able to do things, like solve arithmetic problems, indicating that they have learned how, but be quite incapable of describing what they did. Although we cannot go into the problem here, this sort of analysis appears to provide relatively simple explanations for a number of well-known phenomena, such as retroactive inhibition and reminiscence. (Details are given in Volume I).

It should be emphasized, however, that the theory is essentially deterministic, and applies only where one is dealing with highly structured materials, where one can make reasonable assumptions about the kinds of rules used in storage and retrieval. The theory is not designed to handle data from typical short-term memory experiments. (Even here, however, it can be suggestive (Scandura, Volume I).) Rather, the theory calls for quite a different kind of memory experimentation—experimentation with relatively complex and more highly structured materials, where explicit attention is given to the goal conditions imposed on the $S$ and the kinds of storage and retrieval capabilities with which he enters the situation.

The only fundamentally new hypothesis involves information processing. The basic assumption is that each individual subject has a fixed finite capacity for processing information. While this capacity may vary somewhat over individuals, it is assumed to be of the order, $7 \pm 2$ units of information. (The term bits of information is avoided since it implies a connection with information theory which is not intended.) The classic work of Miller (1956) is obviously related, but his results were based largely on averages and relatively simple tasks. (It is not clear just how (or whether) Miller's work on card sorting is related to information processing in the sense described.) It is important that these results be extended to individuals and generalized to more complex tasks. We assume that this capacity remains constant for individuals, whether one is adding numbers, storing information, or solving problems—as well as in repeating strings of digits, as Miller had his subjects do.

Demonstrating this to be the case, however, is not a trivial problem. Another student, Donald Voorhies, and I have been working on the problem trying to refine our experimental procedures to the point where we can get a fair test of the hypothesis. We still have some way to go but the results of our pilot data were reasonably good almost from the beginning and this, in retrospect, is probably what kept us going. In each case, after a certain degree of complexity was reached there was a sharp "drop off" in performance. Even this, however, required meticulous attention to detail. First, the procedure in question had to be broken down into its basic states and operators. Details are postponed until Volume I but the basic idea is closely related to Suppes' (1969a) S-R characterization of finite automata and my (1970b) reinterpretation in terms of rules (cf. Chapter One). Second, we had to get each $S$ to use this procedure exactly as prescribed. The smallest of deviations could materially affect the results.
Another major roadblock was that we could not tell ahead of time with a new task where the "dropoff" would occur. What was needed was a general scheme for calculating memory load for any given rule—but developing one did not prove to be a simple task. We have recently come up with something which seems promising, however. Some of the data available at the time of this writing are summarized in Figure 4.

Fig. 4. The performance of four subjects on the indicated tasks with percentage of perfect responses plotted against memory load. For comparative purposes, repeating n digits had a calculated memory load of n; repeating n digits and then saying "1" had a memory load of n + 1; repeating n digits after saying "1" had a memory load of n + 2; addition of two 2, 3, and 4 digit numbers without carrying had memory loads of 7, 8, and 9, respectively; with carrying, the memory loads were 8, 9, and 10.

4 Basically, the technique involves calculating for each step of the given algorithm (1) the number of states needed to determine future states, (2) the number of operators needed to determine future operators, and (3) the number of subsets of the needed states and operators which must be distinguished in completing the procedure. (See Volume I for a more parsimonious version.)
5. CONCLUDING COMMENTS AND IMPLICATIONS

The foundations of three partial theories of structural learning have been described and some relevant data have been reported. First, a partial theory of structured knowledge was proposed, in which it was argued that the knowledge had by any given S may be characterized in terms of a finite set of rules. By allowing rules to operate on other rules (in the set), it was shown how new rules could be generated. Examples were also given to show how these new rules, in turn, could account for creative behavior. With the addition of several performance assumptions, this theory was extended so as to account for learning, performance, and motivation under idealized conditions where behavior is unencumbered by memory. Finally, we outlined how memory and information processing might be dealt with, and reported some preliminary data in favor of our main hypothesis. Even the most encompassing theory, however, does not deal with a number of behavioral phenomena, specifically the ultra short-term after images reported by Sperling (1960), Averbach and Coriell (1961), and others. Whether the theory might be extended further to account for these phenomena is difficult to say. But, in any case, this might well be left until later given the large number of questions raised by the theory as it presently exists.

The theory itself represents a sharp departure from existing theories of cognitive behavior, although it does have some things in common with existent competence and information-processing theories. The differences even here, however, are not minor, but have a fundamental effect, both on theoretical adequacy and on the very kinds of empirical questions one asks. Probably the most basic departure is the idea of introducing different levels of empiricism, and the possibility of deterministic theorizing at each of these levels. According to this view, it is possible to do behaviorally relevant empirical research at at least three quite distinct levels. Although all competence models, such as those proposed by Chomsky in linguistics, purport to deal with knowledge, concern traditionally has been limited primarily to the so-called mature speaker or hearer who effectively knows all there is to know about the language. In the present formulation, it is just as reasonable to talk about the knowledge had by different individuals, naive ones as well as mature. This is an extremely important characteristic in dealing with subject matters like mathematics, science, or even language, where knowledge is not a static thing, but grows with experience.

An even more basic departure is allowing rules to act on other rules. This seems to me to be the only real hope we have at present with which to account for creative behavior within an algorithmic framework. There is a good deal more detailed work to be done, but so far the main roadblocks appear to be ones of detail and not of principle.

The distinction between idealized theorizing and related empiricism, on the one hand, and the more complete theory, including memory, on the other, is equally basic. By ignoring the effects of memory and information processing capacity, for example, it has been possible to deal with quite complex behavior, such as problem solving and motivation, in a very precise way—and even more important, in near deterministic fashion. In addition, the proposed mechanisms of memory and information processing are simpler and potentially more precise than those of existing information processing theories. Furthermore, the theory is designed primarily to apply to memory and information processing with complex structured materials, and not just with the short-term memory of lists.
of nonsense syllables, simple words, or sentences, as has been the case with most modern memory research.

Let me finally just mention some of the most promising areas of application of this work in education. Insofar as curriculum construction is concerned, it is sufficient to simply reemphasize that it is a small conceptual step from characterizing knowledge of individual Ss in terms of rules to characterizing curricula in terms of operational objectives. Unlike the current list type curricula (Lipson, 1967), however, explicit attention might be given to the identification of higher order relationships. As simple as this change may seem, its importance cannot be overemphasized. It makes it possible not only to build a good deal of transfer potential directly into a curriculum, but also to capture, I think, what subject matter specialists almost uniformly feel has been missing in current curricula of the operational objectives variety—the creative element. We have a pilot project underway at Penn at this time, in which we are attempting to apply these ideas to teaching mathematics to elementary school teachers. It is too soon to say how things will actually turn out, but so far things have been going extremely well and we hope that we will be able to teach more sophisticated mathematics in this way, and to teach it more effectively.

A second major implication has to do with testing, particularly that sort of testing used to determine mastery on the objectives which go to make up curricula of the sort indicated. Here, the groundwork has been all but completed, and application would seem to be a rather straightforward operation. In fact, several of my students have actually utilized these ideas in a developmental project aimed at diagnosing difficulties urban youngsters are having with the basic arithmetical skills. Another phase of this project has to do with remediation of these difficulties. In this regard, we are using our own home-grown version of hierarchy construction. What we do, in effect, is simply to identify the particular algorithm (rule) we want to teach the child, and break it down into atomic sub-rules. Each sub-rule, in turn, is broken down in the same way, until we reach a level where we can be sure that all of our subjects have all the necessary competencies. This breakdown corresponds directly to the hierarchies obtained in the usual manner by asking Gagne's (1962) often quoted question, "What must the learner be able to do in order to do such-and-such?" Unlike the traditional approach, however, ours provides a natural basis for constructing alternative hierarchies (since any number of procedures may be used to generate the same class of behaviors). Possibilities also exist in such areas as teaching problem solving, but our work to date has been limited to testing basic hypotheses.
American psychology is presently in the midst of an intensive re-examination of its fundamental assumptions, and this is particularly true in the study of complex human learning. Until recently, the predominant school of thought was based on the oft stated (or implied) but never fully documented belief that the study of simple learning (conditioning) would in the end provide the key to complex human behavior. In particular, it was widely assumed that although new mechanisms might have to be added no basic changes would be necessary in the mechanisms of simple learning per se (cf. Bourne, Ekstrand, & Dominowski, 1971; Suppes, Chapter 1).

Many behavioral scientists are no longer willing to make this assumption (e.g., Ausubel, 1968; Chomsky, 1959; Gagné, 1962; Miller, Galanter, & Pribram, 1960). In this view, it is not so much a question of how to integrate traditional behavioristic notions with emerging new ideas as it is of adopting new approaches. The work of Miller et al. (1960), for example, gave major impetus to the now current information processing approach to behavior, an approach which is still being refined and extended. Newell, Shaw, and Simon (1958) went further and proposed that traditional theories be replaced by computer programs. (Testing such theories empirically, however, has always posed a problem.) The Anglo-Saxon awakening to Piagetian type research represents a third main new current. Although these new approaches differ in significant ways, there has definitely been a gradual shift from an empiricist philosophy to one based on structuralism (cf., Piaget, 1970; Scandura, 1967a).

*I am grateful to Paul Rosenbloom and Dana Scott for a number of helpful suggestions concerning the manuscript. One of the key examples developed in the paper was suggested by Professor Rosenbloom. Professor Scott went through the manuscript definition by definition and helped me to clarify the exposition on a number of important points. His comments concerning definition 12 were particularly helpful. Any errors that remain, of course, are my responsibility alone.

5 Although Gagné seems to belong to the new school in fact and activity, he is perhaps more nostalgic concerning mediation theory than the others.

6 In an attempt to retain a strict behavior orientation, Bourne, Ekstrand, and Dominowski (1971) have recently offered a rule-following view
Serious questions have also been raised concerning the basic adequacy of the kinds of empirical evidence with which psychologists have been most concerned. Chomsky (1957, 1968), for example, has proposed that cognitive theories begin with the problem of characterizing competence quite apart from how such competence is put to use or acquired. Certain features of what has been called Hypothesis theory (e.g., Levine, 1966; Restle, 1962) also differ rather sharply with traditional behavioristic notions. In particular, the effects of memory are partialled out of both the theory and much of the data.

Scandura (see part one of this chapter) has recently proposed a largely informal theory of structural learning which among other things synthesizes and extends these views. As is generally true of the information processing approach, this theory is deterministic, as opposed to probabilistic, in form. Unlike most existing theories, however, it consists of three interrelated partial theories: a theory of knowledge, an idealized, memory-free theory of structural learning, and an enriched theory which takes into account memory as well.

The theory of knowledge is essentially a theory of the observer. It is a theory whose aim is to invent a finite set of rules, and laws governing their interaction, which make it possible to account for a given class of behaviors of potential interest to the observer.

Essentially all existing competence theories, of course, limit the observer in his account to a finite set of rules. The mind is finite (although large in capacity) by almost any reckoning and this seems to be a reasonable requirement for any competence theory which purports to have relevance for psychology. There are fundamental differences, however, in the ways in which the rules in a given account may interact. In the study of formal systems (Nelson, 1968), for example, where the goal is to account for (generate) theorems from the axioms, the inference rules in the characterizing set may be arbitrarily composed (i.e., applied one after another). Transformational grammars (Chomsky, 1968) allow transformations (rules) on phrase structures as well. Phrase structures are essentially chains of phrase structure (rewriting) rules. The present theory of knowledge extends these ideas by allowing rules to act on rules at arbitrary levels. In spite of its generality, however, this theory is still a partial theory (of competence), and says nothing whatever concerning behavior.

The idealized theory of structural learning begins where the theory of knowledge leaves off and brings the behaving subject into the picture. In particular, the theory does two things: (1) It provides an explicit and operational way for determining which (parts) of the rules (introduced by the observer to account for the behavior of interest) the subject has mastered. (2) It spells out the mechanisms which govern learning and performance under conditions where the subject is unencumbered by memory (i.e., all relevant information is available) and by his limited capacity for processing information. As is shown below, these mechanisms are assumed to apply to the rules available to the subject. The rules attributed to the subject, however, are relative to those introduced by the

of complex human learning. According to this view rules are introduced to account for observed behavior but subject's do not actually use rules (to generate behavior). This view contrasts sharply with those of such theorists as Newell et al. (1958) and Miller et al. (1960).
observer so that explanation and prediction in the idealized theory depends (in part) on the adequacy of a given theory of knowledge.\footnote{In effect, different observers may introduce different rule sets and, thus, make different predictions in essentially identical physical situations. Many psychologists might balk on this point. Since observation is a common requirement of all science, why single it out in the study of complex human learning? Unfortunately, I cannot give a completely definitive reason. It may be relevant to point out, however, that generally speaking the role of the observer, and in particular the importance of spelling out his role in the science, becomes increasingly important the more complex the phenomena. In relativity theory, for example, the frame of reference of the observer is an essential ingredient. In dealing with complex human behavior, is it not reasonable to suspect that the measuring units (i.e., rules) used by observers to account for given (not observed) classes of behavior might be different? And, if this is true, can we really expect to devise adequate theories, particularly deterministic ones, which ignore these differences?}

The theory of memory is obtained by adding more structure to the idealized theory. The goal of this theory is to deal with (permanent) memory and the limited capacity of humans to process information in a way which does not destroy, but rather generalizes, the essentials of the idealized theory.

The purpose of this paper is to formalize and in certain cases to extend the foundations for the first two levels of Scandura's theory. We conclude with some brief comments concerning the form that a theory of memory might take. More detail, examples, relationships to other research, and data which support some of the basic assumptions of the theory are given in Volume I.

1. FOUNDATIONS OF THE THEORY OF KNOWLEDGE

Following the traditional distinction in axiomatics between theories (descriptions of systems) and systems themselves, there are two basic kinds of behaviorally relevant theories of knowledge. The former type is descriptive and is well illustrated by Piaget's description of developmental stages.\footnote{Piaget, of course, does not stop at description. His theory also deals with the ontogeny of the developmental stages. Even here, however, his concern is with the epistemic subject and not with the individual.} Both he and some of his followers have attempted to describe certain essential features of behavior. The goal, then, is to taxonomize and characterize such descriptions (of behavior). The present approach falls in the latter category. Here, the goal is not to describe behavior but to show how it may be generated.\footnote{There are certain advantages of each approach for a theory of behavior. Generative grammars provide a more natural basis for psychological theorizing of the information processing variety, and although I cannot fully justify my claim, I feel that this approach provides a more natural basis for dealing with individual behavior. Conversely, descriptive theories seem to provide a more natural basis for dealing with group behavior. In fact, an extension of Nercut' (1967, 2) reasoning suggests that descriptive theories may be a necessity in dealing with certain pathological patterns of behavior.}
stimulus-response pairs and our job is to provide specifications for how this class of pairs may be generated.

The purpose of this section is: (1) to specify the nature of the stimuli (inputs) and responses (outputs) to which the theory relates, (2) to define precisely what is meant by rules and their extensions, (3) to identify some of the kinds of higher order rules allowed in the theory, (4) to specify the form of the theory itself (including the laws governing the interaction among rules), (5) to illustrate the theory, and (6) to suggest a general axiomatization and to propose some conjectures.

The theory so defined provides a schema to which specific realizations of the theory must conform. Any particular realization of the theory involves a finite set of rules which, given the laws governing their interaction, accounts for the behavior of interest. A specific rule set, together with the laws of interaction, constitutes a theory of the given behavior. In this sense, a theory of knowledge plays the same role as does a grammar of a language (i.e., the grammar is a theory of the language, Chomsky, 1957).

1.1 Nature of the Stimuli and Responses

Among the undefined terms of the theory of knowledge are elements (E); and the relation of being an input element for an output element (cf. Hocutt, 1967). Since the theory is designed to have behavioral relevance, we single out those input-output pairs that are potentially observable and refer to them as stimulus-response (S-R) instances (or equivalently (S×R) pairs). The class of S-R instances is a subset of the Cartesian product set E×R. We denote this subset by E×R. The terms "input" and "output" are used more broadly to refer to arbitrary inputs and outputs of rules (whether they are stimuli and responses or not).

Although these S-R instances correspond to potentially observable events in the real world, we are not concerned in the theory of knowledge with how the S-R instances in E×R are determined. They are simply given. (Equivalently, the contents of E×R depend on the interests of the theorist.)¹⁰

¹⁰Presumably, S-R instances are determined on the basis of some set of criteria known to the theorist and of interest but which for one reason or another (e.g., there may be too many of them to list) he has never made public. The reason that the theorist may be afforded this degree of freedom is the common culture he shares with other scientists. The linguist, for example, may take phonemes, morphemes, words, sentences, or the paragraph as his basic unit of analysis depending on which aspects of language concern him most. The structural learning theorist has even more latitude.

Put differently, the theorist always keys on certain properties of the stimulation he receives. The rest are ignored. These properties have the effect of defining classes of functionally equivalent stimulation, stimulation which for purposes of the theorist would be rendered indistinguishable. This stimulation may range from highly specific discrete behavior to rather continuous behavior (or stimulation) which takes place over a period of time. When this stimulation results from behavior, the term response (R) is used. Stimulus (S) refers to stimulation felt to effect behavior.
The given set $\mathcal{S}:\mathcal{R}$ of S-R instances, then, could form the starting point of our development, and for many purposes this is the most convenient thing to do. For other purposes, however, especially where we are concerned with perceptual and decoding phenomena, this is not sufficient. In this case it is not sufficient to simply identify the S's and R's of interest. It is also important to specify the respective structures of these stimuli and responses.

To accomplish this, we begin with a set of unanalyzable (atomic) stimuli and responses and let the inputs and outputs of interest be complexes (or arrays) of the atomic stimuli and responses. Examples of atomic stimuli and responses are provided by the black and white dots from which images are formed in newsprint and on the television screen, the numerals 0-9 in arithmetic, and the letters of the alphabet in English.  

We use the terms stimulus situation and behavior, respectively, to distinguish such complexes from unanalyzed stimuli and responses.

**Definition 1:** An atomic stimulus is a symbol (indivisible set of properties) from some alphabet $A_s$.

**Definition 2:** An atomic response is a symbol from some Alphabet $A_r$. The alphabet may contain a finite or infinite number of elements. We let $A = A_s \cup A_r$.

If we restrict ourselves to visual stimulation, then, certain inputs (outputs) consist of linear sequences of atomic stimuli (responses). Other inputs (outputs) consist of two-dimensional arrays of atomic stimuli (responses)—and similarly for three-dimensional stimuli (responses). Normally, of course, not all arrays will be meaningful and of interest. Those inputs and outputs that are, are distinguished with the labels stimulus situation and behavior, respectively. (Some readers will note the analogy with formal languages.)

We can summarize this as

**Definition 3:** An input (output) array is a one, two, or three-dimensional finite arrangement (sequence) of atomic stimuli (responses).

An array $A$ is a subarray of array $B$ if and only if (iff) it is a proper part of array $B$.

**Definition 4:** A stimulus situation is an input array which is of interest to the theorist.

---

11 Atomic stimuli and responses may in principle be thought of as determined by taking the intersection of the partitions imposed on the real world by each criterion imposed by the theorist. Intuitively, stimuli (responses) may be thought of as atomic if and only if their components cannot be distinguished by any criterion of even potential interest to the theorist. The stimulus, $56^2$, is not likely to be atomic in this sense in arithmetic because the observer will almost certainly want to distinguish among the various digits of which the stimulus is composed.

12 e.g., where all stimuli and responses in $\mathcal{S}:\mathcal{R}$ are atomic.
Definition 5: A behavior is an output array which is of interest to the theorist.

The phrase "of interest to the theorist" is admittedly vague. In the next section we shall see that identifying a stimulus situation (behavior) as being of interest implies the availability to any observer of a perceptual (decoding) rule for detecting (constructing) it. Ordinarily, the scientific observer will concern himself primarily with those stimulus situations (behaviors) that can be detected (constructed) via rules commonly available in a given culture or subject population.

Definition 6: Two stimulus situations (behaviors) are functionally distinct if and only if (iff) they are not identical arrays.

Notice that just because one observable is functionally distinct from another does not necessarily mean that they have nothing in common. The stimulus situations "54" and "56," for example, have a "5" in common.

Because any given environmental event may give rise to more than one functionally distinct stimulus situation (behavior), stimulus situations (behaviors) may be thought of as specifying both that which is relevant and that which is irrelevant in the situation. The functionally distinct stimulus situation 24, for example, specifies that "2," "4," "+" "5," "6," and "," together with the structural relationships among them, are relevant; whereas, other aspects of the environment—for example, the context in which the stimulus situation appears—is not relevant. This property of functionally distinct stimulus situations is important to keep in mind later on where one or more functionally distinct stimulus situations (e.g., 2, 4) are embedded in what for other purposes is itself a stimulus situation (e.g., in 56). We summarize this point by

Definition 7: A stimulus situation (behavior) $S'$ is a substimulus (subbehavior) of $S$ iff it is a subarray of $S$.

To complete our specification of observables, we need

Definition 8: If $S$ is the stimulus situation for behavior $B$, then the $S$ and $B$ are said to form an S-B pair or instance.\(^\text{14}\)

For convenience, we sometimes also use the term "behaviors" to denote instances—for example, in the phrase "the set of rules introduced to account for a class of behaviors (i.e., S-B instances)."

The basic data of a theory of knowledge is a given set $\mathcal{S}$ of S-B pairs. Finally, we observe

\(^{13}\) In this more technical use of the term, we distinguish between "behavior" (singular) and "behaviors" (plural).

\(^{14}\) Note that $S$ is used to denote both stimulus and stimulus situation. Since we shall almost always refer to (S-R) or (S-B) instances, the intended meaning will always be clear from context.
Definition 9: If each stimulus situation in \( \mathcal{A} \) is atomic and each behavior in \( \mathcal{B} \) is atomic, then we denote the given set \( \mathcal{A} : \mathcal{B} \) and refer to its elements as S-R instances.

Examples:

1. Let \( \mathcal{A} = \{a, B, 0, 1\} \)

\[ \mathcal{A} = \{\text{the set of strings of the form } "zB" \text{ where } z \text{ is a string of } a \text{'s (e.g., aaaaaB).} \]

\[ \mathcal{B} = \{\text{the set of strings of the form } "Bw" \text{ where } w \text{ is a binary numeral (e.g., B101).} \]

In \( \mathcal{A} : \mathcal{B} \) the output for any input \( zB \) is \( Bw \) where \( w \) is the binary numeral representing the number of \( a \)’s in \( z \).

In this case, \( B \) is the only atomic stimulus (response), but given any pair of distinct stimuli, one is a substimulus of the other (e.g., \( aaB \) is a substimulus of \( aaaaaB \)).

2. Let \( \mathcal{A} = \{0, 1, \ldots, 9, +, -, x, \} \cup \text{Aux} \) where \( \text{Aux} \) is a set of auxiliary signs.

\[ \mathcal{A} : \mathcal{B} = \{\{S\rightarrow B | S \text{ is a computational problem (e.g., } 95 \times 4) \text{ and } B \text{ is its solution (e.g., } 380)\} \]

In this case there are no meaningful atomic stimuli although \( 0, 1, \ldots, 9 \) may serve as atomic responses.

3. Let \( \mathcal{A} = \{w | w \text{ is a word in English}\} \cup \{S\} \) where \( S \) is a nonterminal symbol.

\[ \mathcal{A} : \mathcal{B} = \{\{S\rightarrow B | B \text{ is an English sentence}\} \]

1.2 Rules

Certain input-output pairs naturally belong together in the sense that the theorist "knows" at least one rule which applies to each input and generates each of the corresponding outputs. A set of such instances is referred to as the extension of a rule; and a description of the rule, as the intension (cf., Rogers, 1967; Scott, 1967). As we shall see below, any (denumerable) number of intensions may have the same extension. For present purposes, we distinguish further between the semantics of rules and their descriptions (cf., Engeler, 1963, for a formal treatment and Scandura, 1967b, 1969a, for behavioral implications).

The main purpose of this section is to define semantics in terms of extensions. In the process, a number of related ideas are also formalized. The basic idea parallels that of the previous section. We take

\[ ^{15} \text{In recursive function theory (e.g., Rogers, 1967; Scott, 1967) intensions are formulated as descriptions. Reference to syntax is important in behavioral science for certain purposes (e.g., see the next section) but alone it does not appear to be sufficient. The descriptive approach provides no natural way to determine, for example, to what degree a subject knows and can use a particular rule. In Part 2 we show how reference to semantics makes this possible.} \]
certain rules as basic, equate their extensions and semantics, and construct all of the remaining rules in terms of them. This amounts to specification of some maximal level of detail beyond which further refinement of a rule is unwarranted. If the theorist's concern is with the mechanics of arithmetic, for example, the appropriate level of analysis may involve rules whose basic steps include rules for generating number facts. We shall see below that there is a close relationship between atomic stimuli (responses) and basic (what we shall call atomic) rules. In particular, once the atomic stimuli (responses) are identified, there is a minimal level beyond which a rule cannot be refined (i.e., broken down). Thus, for example, if the numerals 0, 1, 2, ..., 9 are taken as atomic, then the basic number facts correspond to minimal atomic rules. Any further "breakdown" of such atomic rules (e.g., to specify how the number facts are actually generated) would necessarily lead to consideration of more environmental detail--e.g., representation of the numerals as successors (of zero).

In any case, by starting at an atomic level, a strictly extensional approach to semantics is possible. We start by defining certain minimal (atomic) extensions, and then construct the semantics of more complicated rules from these minimal extensions. Our development shares certain characteristics with Engeler's (1968) treatment of algorithmic bases but is more general (although less formal).

Rather than begin with our more general version, we first develop the idea for $d:R$ and then show how it can be generalized to $d:B$

**Definition 10:** An atomic operating rule (atomic or) is a set of input-output (I-O) pairs in which each input has a unique output.

$$
\text{atomic or } = \{(I_i, O_i) | O_i = O_j \text{ whenever } I_i = I_j\}
$$

The set of inputs of an atomic operating rule is called its domain (Dom) and the set of outputs, its range (Ran).

Atomic or's correspond to classes of S-R instances in which the behavioral scientist (as distinct from the competence theorist) is not interested either in distinguishing among the instances or in the details of a (rule based) account of these instances. Depending on the interests of the theorist, any of the following might be considered atomic:

1. knowledge of the addition facts through $9 + 9 = 18$,
2. the ability to carry in addition,
3. the ability to read (where reading ability is involved in the behavior of concern but is not itself the object of study and can safely be assumed of the subjects--e.g., college students),
4. the ability to compute.

---

16 The existence of some such minimal level is guaranteed by the fact that each rule is essentially equivalent to a Turing machine.

17 The theorist who chooses to think of a stimulus (e.g., numeral) as an indivisible whole is not likely to be concerned with how its response is derived from it. Conversely, if he were interested in the nature of the derivation, he would have to deal with particular properties of the stimulus, and effectively treat it as a composite.
Definition 11: An atomic decision making capability (atomic dmc) is an ordered finite partition \( (P_1, P_2, \ldots, P_n) \) of a class \( C \) of elements \( E \) where \( P_i \) is a set of the form \( \{ E | E \text{ satisfies predicate } P'_i \} \) for \( i = 1, 2, \ldots, n-1 \) and \( P_n \) is the complement of \( \bigcup_{i=1}^{n-1} P_i \) relative to \( C \).

In an automaton, each atomic decision making capability (partition) contains exactly two equivalence classes (i.e., the branches in the state diagram of every automaton are binary with input elements 0 and 1). Other examples of dmc's are:

1. \( \{ \{x| x > 3\}, \{x| x \leq 3\} \} \) where \( C = \{x| x \text{ is a real number}\} \)
2. \( \{ \{n| n = 2n - 3\}, \{n| n > 2n - 3\}, \{n| n < 2n - 3\} \} \) where \( C = \{n| n \text{ is an integer}\} \)

Some observations:

1. Each element in the class \( C \) associated with a given atomic dmc is in exactly one element (i.e., equivalence class) of the atomic dmc.

2. Atomic dmc's correspond to decisions in computer programming and play an essential role in the theory. They make it possible to move from (operating) rules with finite domains to rules with infinite domains. Physically speaking, this effects a change from stimuli which can be represented by bounded stimulus arrays to stimuli of arbitrary size. While not critical in the idealized (memory-free) theory proposed later on, stimulus size will almost certainly be a factor where memory and other physiologically based characteristics of the human are taken into account.

3. Notice the close relationship between an atomic dmc and the behavior expected of subjects in the typical concept attainment study where subjects are required to sort elements (i.e., stimuli) in a given universe into two piles, one containing exemplars of the concept (i.e., stimuli which satisfy a predicate) and the other containing non-exemplars (i.e., stimuli which do not satisfy the predicate). (Note: The sorting operation itself is due to a pair of constant valued rules which act on the stimuli in respective elements of the partition formed.)

4. Because of the close relationship between atomic dmc's and predicates, we frequently suppress the difference and speak of the predicates (relations) themselves. This is particularly common in logic and computer programming where binary partitions (involving truth or falsity and yes or no, respectively) are common.

Rules are defined in terms of atomic or's and dmc's.

Definition 12: A rule \( r \) is a labeled finite directed graph whose "arrows" are atomic or's labeled with predicates and whose "nodes" are atomic dmc's. One node is labelled "Start." Each node, which contains predicates not used to label arrows emanating from it, is labelled "Halt."

A stimulus \( S \) and a rule \( r \) together determine a response \( R \) as follows: The dmc labelled Start is applied first. The atomic or applied next is the one labelled with that predicate in the starting dmc which contains \( S \).
Control then flows to the dmc determined by the preceding arrow (i.e., at the head of the arrow). In a similar manner, control proceeds from the nth atomic dmc, to the designated nth atomic or, to the n + 1st atomic dmc. This continues until the computation terminates. This occurs where an atomic or is undefined on a stimulus or no next atomic or is specified.

Definition 13: An atomic rule (atomic r) is a rule consisting of one arrow, with one Start and one Halt node.

Definition 14: A subgraph A is called a subrule of rule B if the subgraph has a unique entrance node from the complement of A in B, and becomes a rule upon proper labelling. Specifically, the entrance is labelled "Start" and each node, which contains predicates not used to label arrows emanating from it, is labelled "Halt."

Definition 15: Rule A is a closed subrule of rule B if it is a subrule of B in which there is exactly one node designated "Halt."

We denote rules schematically in terms of modified flow diagrams. For example

![Diagram](attachment:flow_diagram.png)

Each rule, then, corresponds to a computer program, where the arrows and nodes correspond to (labeled) operation instructions and test instructions, respectively. In our illustration, dmc_1 (Start) corresponds to "If m = 0, do or_1; else do or_2;" or_1 corresponds to "(n,0) \rightarrow (n);" or_2 to "((n,m),(n+l, m-1));" and dmc_2 to Halt.

Furthermore, closed subrules correspond to subroutines in computer programming. Indeed, any rule A can be redefined in terms of closed subrules as long as each atomic rule in A is in exactly one closed subrule of A. In this case, the "arrows" of the redefined rule correspond to closed subrules. A set of closed subrules with this property is called a complete disjoint set.

Definition 12 is incomplete in an important sense; it ignores both "perceptual" and "decoding" processes. It is simply assumed that stimuli are somehow automatically encoded (via "Start"), and responses...
automatically decoded (via "Halt"). Nothing is said about the details of either process. This is no great loss where a rule is designed to account for a given class of S-R instances. In this case, the stimuli and responses are unitary so that nothing essential can be gained by including perception and decoding in an account. If the stimulus is the display 24, for example, then a rule which generates a response +36 from this stimulus must do so by acting on the stimulus as a whole. The first step in an addition rule, for example, might be represented

\[
\begin{array}{cc}
24 & 24 \\
\downarrow & \downarrow \\
+36 & +36 \\
\hline \\
0 & 6
\end{array}
\quad \text{and the last}
\begin{array}{cc}
24 & 24 \\
\downarrow & \downarrow \\
+36 & +36 \\
\hline \\
0 & 6
\end{array}
\]

Most would agree, of course, that this is the stimulus and that this stimulus generates the indicated response. But in reality, it is typically not the whole stimulus which elicits the response associated with the stimulus (e.g., the two units digits elicit 1). Furthermore, the response (behavior) is really just 60 and not the entire complex. Any complete account must reflect these details.

Fortunately, perception and decoding can be dealt with by extending our concept of rule. The essential idea involved is that of incorporating (atomic) capabilities into each rule which extract atomic stimuli (or, more generally, substimuli) from stimulus situations and which construct behaviors from atomic (sub)behaviors.

In the illustration above, for example, a verbal account might go as follows: (1) Locate and encode the two digits on the right—notice that the "response" here is a movement and (automatic) encoding of a substimulus (consisting of two atomic stimuli). (2) The encoded digits, then, serve as a (sub) stimulus for the (sub) response 10. (3) (Automatically) decode 10 by putting "0" beneath the units column and retaining the 1 (or placing it slightly above and to the left of the "0"). (4) This subresponse (is automatically perceived and) serves to elicit a move to the tens column. (5) Here, the tens digit from the first summation, together with the new tens digits, constitute the stimulus for the response 6. (6) Finally, "6" is placed in the tens place of the answer.

From this example it might appear that encoding (decoding) simply involves internalizing (externalizing) some substimulus (subbehavior). This is not entirely true, however, as can be seen by considering \( \triangle \) and \( \vartriangle \). Both displays are typically encoded as triangle. Triangle, however, is not a substimulus but rather a property which defines a class of stimulation in the real world. What we have called a substimulus (stimulus) is in actuality a canonical representative of a class of real world stimulation. Thus, the substimulus "5" in "45," for example, represents a class of equivalent displays (e.g., [5, 5, five, ...]). We shall see at the end of Section 1.5 that perceptual knowledge grows as rules interact, and that this growth has the effect of imposing finer and finer partitions on the environment.

Notice finally that when encoding and decoding rules are added to our formulation, stimuli and responses (behaviors) rarely serve as inputs and outputs, respectively, for the atomic or's in a rule. What we
call stimuli and behaviors can, however, be recovered from arbitrary rules and that in part is what the formalization below is designed to make possible.

We strengthen our formulation by adding two kinds of atomic capability.

**Definition 16:** An atomic encoding (perceptual) rule (atomic p rule) is a set of input-output pairs with a finite Range in which each input is a stimulus and each output is a class of stimuli which includes the stimulus.

Since atomic p rules effectively partition the environment, they may alternatively be defined in that way (i.e., as partitions). Intuitively, atomic p rules involve the direct identification of stimulus properties. Where a property is a substimulus, the p rule may be thought of as locating the substimulus in the environment, and then, inserting the substimulus in a class of stimuli. (Appropriately restricted atomic p rules correspond to movements of reading heads of Turing machines together with their basic capability of "reading" 0's and 1's on tapes.)

Each of the following are atomic p rules: (1) a rule which partitions a class of objects according to color (size, shape, etc.), (2) a rule which partitions a class of numerals according to the number of digits, (3) a rule which "reads" the digits 0, 1, 2, ..., 9.

**Definition 17:** An atomic decoding rule (atomic d rule) is a finite set of input-output pairs in which each input is a class of (sub) behaviors and each output is a (sub) behavior in this class.

Intuitively, atomic d rules involve altering the environment by constructing observables so that the environment has specified properties. Where the property is a subbehavior, the d rule involves selecting the subbehavior (which is a canonical representative) and constructing it in the environment in a particular location. (Atomic d rules correspond in Automata theory to movements followed by "writing" on tapes.)

Each of the following are atomic d rules: (1) a rule which constructs objects having specified shapes (colors, sizes, etc.), (2) a rule which constructs a numeral with a specified number of digits, (3) a rule which "writes" the digits 0, 1, 2, ..., 9 in particular locations.

By way of summary, then, rules as previously defined operate exclusively on encoded stimuli and generate undecoded responses. Adding atomic p rules and atomic d rules to our formulation of rules has the effect of adding "arrows" to our directed graph characterization which involve insertion into classes (encoding) and extraction from classes (decoding), respectively. Thus, what was an atomic rule before involves three operations, an atomic encoding (perceptual) rule, an atomic operating rule, and an atomic decoding (operational) rule.

---

13As we shall see in discussing the idealized (memory-free) theory, operational tests of encoding rules may be obtained by requiring subjects to classify stimuli. In decoding, the subject is required to construct responses of various types.
and an atomic decoding rule. (Strictly speaking, atomic operating rules must be restricted to finite domains unless looping of atomic p rules is allowed because the ranges of the latter are finite.) In the study of perception, encoding rules may be expected to play the primary role. Decoding rules would appear critical in studies involving skills of various kinds (e.g., speaking, writing).

Encoding rules make it possible to spell out the manner and order in which substimuli are "extracted" from the environment. It is no longer necessary to assume that all stimuli are extracted automatically as wholes. In a like manner, decoding rules make it possible to detail the process by which responses are generated. Responses typically are not generated instantaneously, but are constructed from subresponses over a period of time. (As an exercise, devise a precise directed graph characterization of the verbal account of addition. That is, formulate a general rule for adding numbers which includes both p and d rules.)

In order to place the role of (atomic) p and d rules in perspective, it is instructive to consider the rule \( \frac{(A + L)}{2} \) for summing arithmetic series. (A is the first term of the series, L the last, and N the number of terms.) Observe that identifying A and L is essentially a matter of locating them. (This is not entirely true, unless A and L are bounded from above.) A complete account of how N is determined, however, requires more than just encoding (insertion into classes). A and L are substimuli, whereas N is a property derivable from substimuli. (One possible rule might involve encoding the terms one by one and counting.) It should be emphasized, however, that in many applications this distinction is unimportant and may be suppressed.

Although the atomic p (and d) rules in a rule generally speaking do not operate directly on stimulus situations (nor directly generate behaviors), it is always possible to recover the stimuli (and behaviors) associated with a rule given only the substimuli—inputs (and subresponses—outputs) associated with each of its constituent atomic p (and d) rules. That is, stimulus situations and behaviors are implicit in any rule which accounts for them. Given a particular rule, a stimulus situation associated with the rule may be determined by tracing through the rule as follows: Go to the starting dmc. Pick one element in one equivalence class that belongs to the domain of some atomic or. Construct (in the environment) a canonical representative of this element. Each time an encoding rule is encountered, move to the appropriate location in the environment and construct a canonical element in the output class (of the encoding rule). (Notice that elements in the extensions of p rules, or's, and dmc's are classes of observables.) The process is repeated until the computation terminates. The configuration of substimuli which results is defined to be a stimulus situation associated with the rule. The domain of a rule r is the set of stimulus situations associated with the rule.

Behaviors are generated by simply applying the rule to stimulus situations. That is, the behavior associated with a rule r and stimulus situation S is determined by applying r to S. The behavior is the configuration formed from the subresponses of the atomic d rules in r. The range of a rule, then, is the set of behaviors generated by application of the rule to its domain.

In effect, it is possible not only to build up rules out of atomic rules but also to detail how stimuli (behaviors) may be recovered from
rules. Recall that in Section 1.1 we promised to specify more precisely the sense in which certain arrays might be of interest. Such interest is a direct function of the availability of appropriate perceptual (and decoding) rules in a given culture or subject population.

We summarize this in

Definition 13: The extension $\text{Ext} (r)$ of a rule $r$ is the set consisting of all S-B pairs such that $S$ is in the domain of the rule and $B$ is the output generated by application of $r$ to $S$.

Notice that any number of rules may have the same extension. For example, consider the borrowing and equal additions methods for subtraction. The extension of an atomic rule, however, is essentially (up to one-to-one correspondence) the same as the set of input-output pairs of the constituent atomic or. (The extension of the atomic rule by definition consists of S-B instances, whereas the set of input-output pairs of the atomic or includes the encoded counterparts of these S-B pairs.)

A rule is said to be finitary if and only if its extension contains a finite number of S-B instances. Rules with infinite extensions may be built up from (finite numbers of) finitary rules by looping back on the constituent rules in the usual way.

Our final point is that rules can often be refined by considering processes underlying one or more of their atomic rules. The associated $\mathcal{J}:\mathcal{B}$ classes of S-B instances, however, place limits on the degree of refinement possible. The substituents involved in any rule account involving perception, for example, may not involve discriminations finer than the atomic stimuli. In order to add more detail it would be necessary to redefine the given set of S-B instances.

More generally, we state

Definition 19: (a) Rule $r'$ is a refinement of rule $r$ if there is a one-to-one correspondence between the atomic operation (p, or, and d) rules of $r$ and the closed subrules of $r'$ such that the extension of each atomic rule in $r$ is the extension of the corresponding closed subrule of $r'$. (b) Rule set $K_1$ is a refinement of rule set $K_2$ if there is a one-to-one correspondence between the rules in $K_1$ and $K_2$ such that each rule in $K_1$ is a refinement of the corresponding rule in $K_2$.

Observation: If $r'$ is a refinement of $r$, then the extension of $r'$ equals the extension of $r$ (i.e., $\text{Ext} (r') = \text{Ext} (r)$).

To see this we observe that the atomic p rules in $r$ correspond in one-to-one fashion with closed p subrules of $r'$ with the same extensions. But, the domains of rules are determined uniquely by the extensions of their p rules. Also, rules $r$ and $r'$ generate exactly the same outputs for any given input since the closed subrules of $r'$ have exactly the same effect as the atomic operation rules of $r$. Hence, $\text{Ext} (r') = \text{Ext} (r)$.

In thinking about perception (decoding) it is important to recognize that different perceptual (decoding) rules may operate at different levels. Specifically, in Section 1.5 we shall see that perceptual rules generated (by application of higher order rules) may involve a finer
basis (i.e., sub-substimuli) than the perceptual rules from which they are generated. Later on, it will become apparent that learning to perceive in general involves moving from gross to finer distinctions concerning the environment. According to Piaget, for example, the young child is able to distinguish between figures which are and are not (topologically) closed before he is able to distinguish, say, between squares and circles, both of which are closed.

1.3 Programs: Descriptions of Rules

Although rules may serve as inputs and outputs of other rules (cf. Section 1.4), rules themselves may not serve as stimulus situations or behaviors. Observable counterparts of rules do exist, however, and these we shall call programs. Programs essentially are descriptions of rules in a suitable language.

For present purposes we shall simply assume that some suitable language exists, say ordinary mathematical English or some programming language. Modifying Scott (1967) slightly, this may be accomplished by introducing n-tuples of predicate terms \( \langle P_1, \ldots, P_n \rangle \) for the atomic decision making capabilities and function terms \((F, E, \text{ and } D)\) for atomic operating, encoding, and decoding rules, respectively. In addition, we need labels \( (L) \), and the following special terms: Start, Go, to, If, Then, Do, else, :, ;, and Halt.

**Definition 20:** An instruction is a string of one of the following six forms

- Start: Go to \( L \) (start instruction)
- \( L \): Do \( E \); go to \( L' \) (encoding instruction)
- \( L \): Do \( F \); go to \( L' \) (operation instruction)
- \( L \): If \( P_1 \) then go to \( L_{1}' \); if \( P_2 \) then go to \( L_{2}' \), \ldots; if \( P_{n-1} \) then go to \( L_{n-1}' \); else go to \( L_n \) (test instruction)
- \( L \): Do \( D \); go to \( L' \) (decoding instruction)
- \( L \): Halt (halt instruction)

where \( L, L' \) and \( L_{i}' \) are members of the set of labels \( \mathcal{L} \); \( F, D, \text{ and } E \) are function terms in \( \mathcal{F}, \mathcal{D}, \text{ and } \mathcal{E} \), respectively, and \( P_i \) is a predicate term in \( \mathcal{P} \).

**Definition 21:** A program is a finite set of instructions containing exactly one start instruction and containing for each label that occurs anywhere in any instruction in the program, exactly one instruction that begins with that label.

In our system it will turn out that programs may denote rules which act on classes of programs (as inputs). The instruction symbols of which such programs are composed, therefore, will be of an essentially
higher order; the symbols denote operations and predicates which act on
classes of predicate and function symbols. The total number of predi-
cate and function symbols needed in any particular (finite) account, of
course, will always be finite.

The distinction between rules and programs can be suppressed in
many applications where the theorist is uninterested in the rules used
in interpretation (e.g., in encoding symbols, and then assigning them
meanings) and description (e.g., in representing meanings, and then
decoding the representations). In simple arithmetic computation, for
example, no distinction is usually made between numbers and numerals
(number names). In this case it is particularly convenient to assume for
each higher order rule that it includes just one atomic rule which com-
bines encoding and interpretation, and one atomic rule which combines
description and decoding.19

1.4 Kinds of Rules

Definition 22: Simple rules are rules whose extensions do not
include stimuli or responses that are programs for other rules.

Definition 23: Higher order rules (ho rules) are rules that are
not simple; their extensions contain stimuli and/or responses
that are programs for other rules.

Although it is possible to conceive of ho rules which operate ex-
clusively at the syntactic level (i.e., which make no reference to
meaning), such rules have played a central role in linguistic grammars
(cf., Chomsky, 1965) and are not of central concern here. For present
purposes, it is more natural to think of ho rules as actually inter-
preting the programs on which they operate. We shall not, however, have
much to say in this section about the details of this process (See
Scandura, 1970b, Chapter 7, Volume I, and Section 1.5 for further dis-
cussion). It is assumed that encoding and interpretation (and descrip-
tion and decoding) are combined as indicated above. In this case, it is
possible to discard the initial and terminal atomic rules with no loss
of generality and just talk about rules which operate on and generate
rules (i.e., meanings). This practice is adopted throughout this section,
except where we discuss ho rules involved in perception. In particular,
when we refer to the extension of a ho rule we mean the set of input
rule-output rule pairs.

The main purpose of this section is to identify some of the basic
kinds of ho rules. No claim is made regarding the exhaustiveness of
these basic types, however.20

---

19 Where a distinction between rules and programs is desired, there
are two kinds of relevant data (observables); one involves paraphrasing
and the other application of rules corresponding to programs (cf. Section
2).

20 Nonetheless, these types do form a basis for a wide variety of
other types. Further, it appears (although we shall not attempt to
prove it) that a relatively small subset of the identified types may be
sufficient for generating the others in much the same way that "not" and
"and" form a basis for generating all of the basic connectives used in
the statement logic (e.g., "or," "implies").
Definition 24: A composition rule $c$ is a rule which maps pairs of rules represented by $\langle P_1, P_2 \rangle$ and $\langle P_2, P_1 \rangle$ into composite rules represented by $\langle P_1, P_2 \rangle$.

For simplicity, we define $c$ as a rule with an extension of the form

$$c = \{(r_1, r_2; r_3) \mid \text{Ran } r_1 \subseteq \text{Dom } r_2$$

for all $(r_1, r_2) \in \text{Dom } c$ and $\text{Dom } r_3 = \text{Dom } r_1$

and $\text{Ran } r_3 \subseteq \text{Ran } r_2\}$$

Composition rules act on pairs of rules in which the outputs of one serve as inputs for the other. Because the individual rules $r_1$ and $r_2$ shared common labels some relabeling was necessary in forming the composite. (This can always be accomplished in an algorithmic fashion.) In view of what follows, it may also be worth noting that the dmc's of the nodes in the composite rule are identical to (certain of) the dmc's of the nodes in $r_1$ and $r_2$. This is not always the case (cf. Definition 26).

Definition 25: The inverse $c'$ of a composition rule $c$ undoes the composition rule. That is, the extension of $c'$ is of the form $c' = \{(r_1 \circ r_2, r_2; r_1)\}$
Definition 26: A simple generalization rule \( sg \) is a rule whose extension is of the form

\[
sg = \{(r_1, r_2; r_3) \mid r_1(S) = r_2(S) \text{ for all } S \in \text{Dom } r_1 \cap \text{Dom } r_2 \}
\]

\( \text{Dom } r_2 \) and \( (r_1, r_2) \in \text{Dom } sg \);

also \( \text{Dom } r_3 = \text{Dom } r_1 \cup \text{Dom } r_2 \) and \( r_3(S) = r_1(S) \) if \( S \in \text{Dom } r_1 \); \( r_3(S) \) otherwise.

Simple generalization rules act on pairs of rules and generate equivalent rules with domains that are unions of the given ones. It is important to notice in this connection that the \( \text{dmc} \) in the initial node of \( r_3 \) is different from any of the \( \text{dmc} \)'s associated with \( r_1 \) or \( r_2 \). In particular, this (initial) \( \text{dmc} \) is formed by forming a union and accepts any stimulus that is accepted by either the initial \( \text{dmc} \) of \( r_1 \) or the initial \( \text{dmc} \) of \( r_2 \). That is, the initial \( \text{dmc} \) is \( \{ \text{Dom } r_1 \cup \text{Dom } r_2, \text{Dom } r_1 \cap \text{Dom } r_2 \} \).

Definition 27: A conjunction rule \( cj \) is a rule whose extension is of the form

\[
cj = \{(r_1, r_2; r_3) \mid \text{Dom } r_1 = \text{Dom } r_2 \text{ for all } (r_1, r_2) \in \text{Dom } cj; \}
\]

also \( \text{Dom } r_3 = \text{Dom } r_1 = \text{Dom } r_2 \) and \( r_3(S) = r_1 \times r_2(S) \).

For example, let \( r_1 \) add pairs of whole numbers and \( r_2 \) multiply pairs of whole numbers, then \( r_3 \) adds and multiplies pairs of whole numbers (i.e., the outputs are pairs consisting of sums and products). It would be a simple matter to generalize \( cj \) by allowing the initial \( \text{dmc} \) of \( r_3 \) to differ from the \( \text{dmc} \)'s of \( r_1 \) and \( r_2 \). In particular, by letting this initial \( \text{dmc} \) be \( \{ \text{Dom } r_1 \cap \text{Dom } r_2, \text{Dom } r_1 \cap \text{Dom } r_2 \} \), \( cj \) can be defined on arbitrary pairs of rules with non-empty intersections. (Intersection \( \text{dmc} \)'s play a central role in learning how to perceive, and as we shall see in Section 1.5 they are crucial in accounting for \( \aleph \)-\( \mathfrak{B} \) classes.)

Definition 28: An elimination rule \( el \) is a rule which eliminates unnecessary subrules of a rule. More specifically, the extension of an elimination rule is of the form

\[
el = \{(r, r-r') \mid r \text{ contains closed subrule } r' \text{ whose inputs and outputs are identical and } r-r' \text{ is the rule obtained by } "\text{detaching}" \ r' \text{ from } r \}
\]

For example, represent \( r = \frac{2n+1}{3} \) by

\[
x^2 \xrightarrow{+1} (2n+1) \xrightarrow{-1} (2n+1-1) \xrightarrow{\text{assoc}} 2n+(1-1) \xrightarrow{\text{identity}} 2n \xrightarrow{+3} 2n/3
\]

where the circles are labeled nodes.
Then, for some suitable el, el (r) is represented by
\[ x^2 + 3 \]

In the following, r is a rule and p, a subrule in r for determining stimulus properties (i.e., for perception). Hence, rule r may be denoted \( r = p + (r - p) \).

**Definition 29:** A restriction rule res is a rule whose extension is of the form

\[ res = \{(p + (r-p), p'; p' + (r-p)) | p \text{ and } p' \text{ are perceptual rules such that } \text{Dom } p' \subseteq \text{Dom } p \} \]

For example, let \( r = p_N + (r - p_N) \) be a rule for summing arbitrary arithmetic series (e.g., \( \sum (A + L)/2 \)) where \( p_N \) is a perceptual rule for finding the number of terms N for arbitrary arithmetic series (e.g., \( p_N = (D + L - A)/D \) where A is the first term of the series, L is the last, and D is the common difference between terms). Also, let \( p_N' \) be a perceptual rule for finding N for arithmetic series of the form \( 1 + 3 + 5 + \ldots + (2N-1) \) (e.g., \( p_N' = \frac{A + L}{2} \)). (Presumably, \( p_N' \) is more "efficient" than \( p_N \) on its restricted domain.) In this example, r may be represented by \( \left[ \frac{(D + L - A)}{D} \right] \frac{(A + L)}{2} \) and \( p_N' + (r - p_N) \) by \( \left[ \frac{A + L}{2} \right] \frac{(A + L)}{2} \). For an experimental study involving similar rules see Scandura and Durnin (1968). (This experiment is summarized in Volume II, Chapter 2.)

**Definition 30:** A generalization rule g is a rule whose extension is of the form

\[ g = \{(p' + (r - p), p; p + (r - p)) | p \text{ and } p' \text{ are as above}\} \]

Restriction rules appear to be equivalent to the inverse of a composition rule followed by composition. A similar comment applies to generalization rules (Definition 30).

Elimination and restriction rules may play important roles in increasing efficiency of behavior. In particular, practice on a given rule, especially under timed conditions, may lead to the elimination of unnecessary subrules and, hence, result in the acquisition of a more efficient rule. Presumably, the most efficient rules consist of an atomic encoding rule followed by an atomic decoding rule. Efficiency can sometimes be further increased by restriction because restricted rules (see above) frequently involve fewer operations.

Although only treated herein in passing, the generation of new perceptual rules (cf. Section 1.5) apparently serves a similar function, particularly with young children and/or novel stimulation.

Where both efficient and relatively inefficient rules are available for solving a given (kind of) task, selection rules (cf. Definition 31) presumably serve to insure that the more efficient rules are used where needed (e.g., under timed conditions).
For example, again consider the general rule \( N[(A + L)/2] \) for summing arithmetic series and its restriction \( N'(A+L)/2 \) (where \( N' \) is determined by \( \frac{A + L}{2} \) and \( N \) by \( \frac{(D + L - A)/D}{2} \)).

**Definition 31:** A selection rule \( s \) is a rule whose extension is of the form
\[
s = \{ (r_1, r_2; r_1) \mid \text{Dom } r_1 = \text{Dom } r_2 \text{ and } r_1 = r_1 \text{ or } r_2 \}
\]

For example, one selection rule might be described "If \( r_1 \) was used on the previous trial, then select \( r_2 \); else \( r_2 \)"

The selection rule used in a study reported in Volume I, Chapter 8, was "If the stimuli are edible objects, then select \( r_1 \); else \( r_2 \)."

### 1.5 Nature of the Theory of Knowledge

With this background, a theory of knowledge \( \mathcal{K} \) may be defined as an \( n + 3 \) tuple.
\[
\mathcal{K} = \langle A, \mathcal{A}:\mathcal{B}, K, r_1, r_2, \ldots, r_n \rangle
\]
where \( A \) is a finite alphabet, \( \mathcal{A}:\mathcal{B} \) is a class of pairs consisting of stimulus situations and their corresponding behaviors, and \( K \) is a finite set of rules (which may include percepts, i.e., encoded stimuli). The \( r_i, i = 1, 2, \ldots, n \) are the rules in \( K \). These rules may act on rules in \( K \) as well as on the S-B in (the pairs of) \( \mathcal{A}:\mathcal{B} \). Notice also that programs (descriptions of rules) are potentially observable and may serve as elements of \( \mathcal{A}:\mathcal{B} \) whereas rules themselves may not. Conversely, encoded percepts may belong to \( K \) but stimuli may not.

Nonetheless, for many purposes involving complex knowledge, it is convenient to ignore the details of encoding and decoding processes. In this case, we have
\[
\mathcal{K} = \langle \mathcal{A}:\mathcal{K}, K, r_1, r_2, \ldots, r_n \rangle
\]

Most of this section is devoted to making precise and illustrating what it means for \( K \) to provide an account of \( \mathcal{A}:\mathcal{K} \). The general idea (cf. Scandura, 1971a) is that the rules in \( K \) may operate on other rules in \( K \) to produce new rules. A rule set is said to account for a given S-R pair in \( \mathcal{A}:\mathcal{K} \), if a rule is eventually generated in this manner that yields \( R \) when applied to \( S \).

With this in mind, we introduce

**Definition 32:** \( K^1 = K \) and \( K^n = K^{n-1}(K) \cup K^{n-1} = \bigcup_{r \in K^{n-1}} r(K) \cup K^{n-1} \)

where \( K^{n-1} (K) \) means the rule set generated by applying all of the rules in \( K^{n-1} \) to every element in \( K \) in its domain. The potential knowledge \( PK \) associated with a rule set \( K \) is
For example, consider the theory $\mathcal{K}$ with alphabet $\{a, B, 0, 1\}$, $\mathcal{A} = \{(xB, By) | x \text{ is a string of } a\text{'s}, y \text{ is the binary numeral which represents the number of } a\text{'s}\}$, and $K = \{r_1, r_2, \circ, i\}$ where $r_1 = xxBy \xrightarrow{a} xBOy$, $r_2 = xxabY \xrightarrow{a} xBy$, $\circ = r, r' \xrightarrow{a} r \circ r'$ for all $r, r'$ and $i = r \Rightarrow r$ for all $r$. (Note: $\circ$ is called the generalized composition rule and $i$ the identity. The $n$-fold application $\circ$ is denoted $\circ^n$.) Then

$$
K^1 = \{r_1, r_2, \circ, i\}
$$

$$
K^2 = \{r_1, r_2, \circ, i, r_1 \circ r_2, r_2 \circ r_1, \circ^2\} = K^1 \cup \{r_1 \circ r_2, r_2 \circ r_1, \circ^2\}
$$

$$
K^3 = K^2 \cup \{r_1 \circ r_2 \circ r_1, r_1 \circ r_2 \circ r_2, r_2 \circ r_1 \circ r_1, \ldots, \circ^3\}
$$

$$
\vdots
$$

$$
K^n = K^{n-1} \cup \{r_1 \circ r_2 \circ \ldots \circ r_n \mid i_j = 1 \text{ or } 2 \text{ for } j = 1, \ldots, n, \circ^n\}
$$

In general, if $K$ contains the identity rule $i$, then $K^n = K^{n-1}(K)$ for $n \geq 2$ and $K^\infty = \bigcup_{n \geq 1} K^n(K)$. We make this simplifying assumption in what follows.

Two comments are worth noting. First, $K^n$ may be thought of as an upper bound on the knowledge that might possibly be acquired via the $n$-fold application of rules in $K$. $PK$ is the asymptotic state of knowledge—that is, the knowledge which might be acquired by a knower characterized by $K$ given an indefinite amount of time (for rules to operate on and generate new rules in all possible ways). Second, in the idealized learning theory introduced in Section 2, the growth of knowledge takes place in a more restricted, although closely related manner. A specified amount of knowledge (i.e., some number of rules—possibly zero) is added each time the knowing subject addresses a new task.

**Definition 33:** (a) A rule $r \in K^n$ is said to account for an S-R pair in $\mathcal{A} : \mathcal{E}$ iff $r(S) = R$.

(b) A rule $r' \in K^n$ is said to account for a rule $r'' \in K^{n+1}$ iff there are rules $r_1, r_2, \ldots, r_i$ in $K^n$ such that $r'(r_1, r_2, \ldots, r_i) = r''$.

In the above example, (a) $r_1 \circ r_1 \circ r_2 \in K^3$ accounts for aaaaB - B100 because aaaaB $\xrightarrow{r_1}$ aB0 $\xrightarrow{r_1}$ aB0 $\xrightarrow{r_2}$ B100. (b) $\circ \in K^3$ accounts for $r_2 \circ r_2 \circ r_1 \circ r_4 \in K^4$ because $\circ (r_2, r_2, r_1, r_1) = r_2 \circ r_2 \circ r_1 \circ r_1$ where $r_1$ and $r_2 \in K^3$.

---

24 This rule corresponds to the "alternating" selection rule studied in Volume I, Chapter 8.

25 I am indebted to Paul Rosenbloom for suggesting this example.
Definition 34: (a) An S-R pair in $\mathcal{A} : \mathcal{R}$ is said to be \textbf{nth order generable} from a rule set $K$ iff there exists a rule $r \in K^n$ which accounts for the S-R pair.

(b) A rule $r''$ is said to be \textbf{nth order generable} (n ≥ 2) from a rule set $K$ iff there are rules $r', r_1, r_2, \ldots, r_i \in K$ such that $r''(r_1, r_2, \ldots, r_i) = r''$.

(c) A rule is \textbf{first order generable} from a rule set $K$ iff it is in $K$. (Note: Rules may operate on given S-R pairs in determining $K^n$).

For example, (a) $aaaa\emptyset - \emptyset 100$ is third order generable because there is a rule $r_3 \circ r_1 \circ r \in K$ which accounts for the pair. Similarly $r_2 \circ (r_1 \circ r_1 \circ r_1)$ is fourth order generable because $r_1$, $r_2$, and $r_2 \in K$ and $o^3(r_2, r_2, r_1, r_1) = r_2 \circ r_2 \circ r_1 \circ r_1$.

Definition 35: A rule set $K$ in a theory of knowledge $\mathcal{K}$ is said to account for $\mathcal{A} : \mathcal{R}$ iff for each S-R pair in $\mathcal{A} : \mathcal{R}$, there is a finite number $n$ such that S-R is nth order generable from $K$.

Evaluating a theory of knowledge empirically involves determining whether or not arbitrary S-R pairs in $\mathcal{A} : \mathcal{R}$ are derivable from $K$. Certain S-R pairs, of course, will be trivially easy to account for so that to make a convincing case for the theory one usually concentrates on aberrant cases or cases which otherwise make it possible to distinguish between alternative theories. (This is standard practice in linguistics.)

Although $n$ is generally restricted to a finite number, presumably Definition 34 could be extended to allow for generability at asymptote (i.e., generability via rules derivable only at asymptote). It is not clear, however, where this might be useful, if at all. Another and possibly more useful alternative might be to restrict the size of $n$ (e.g., $n \leq 7$). This alternative could be of some value in applications where memory is a factor. For certain purposes it might also be useful to define the notion of a "uniform account" of $\mathcal{A} : \mathcal{R}$ in terms of a maximum $n$. In this case, we might look for conditions on $\mathcal{A} : \mathcal{R}$ and $K$ such that providing an account implies providing a uniform account. (For example, any account in a behavioral objectives type theory (see Section 1.6) is a uniform account.)

Some further definitions are

Definition 36: A theory of knowledge $\mathcal{K}$ is \textbf{simple} iff $K$ contains the identity.

Definition 37: A theory of knowledge $\mathcal{K}$ is \textbf{atomic} iff each rule in $K$ is atomic.

Definition 38: A theory of knowledge $\mathcal{K}$ is \textbf{finitary} iff each rule in $K$ is finitary (i.e., has a finite domain).

The example given above is simple and atomic, but not finitary.

Definition 39: Let $K$ and $K'$ be rule sets in Theories $\mathcal{K} = \langle \mathcal{A} : \mathcal{B} , K, r_1, \ldots, r_i \rangle$ and $\mathcal{K}' = \langle \mathcal{A} : \mathcal{B} , K', r_1', \ldots, r_m' \rangle$ respectively. Then, $K$ and $K'$ are said to be \textbf{equivalent in computing power} iff every S-R pair in $\mathcal{A} : \mathcal{R}$ either can be accounted for by both $K$ and $K'$ or by neither one.
Structural Learning: Definitional Foundations

For example, define $\mathcal{X}$ as above and let $K' = \{r'\}$ where $r' = \begin{array}{c}
\text{Start} \\
\circ \\
\text{Halt} \\
r_1 \\
r_2
\end{array}$

(The node can be described as "If $x \not= \emptyset$ and the string is of the form $xxBy$, do $r_1$; else do $r_2$. If $x = \emptyset$, Halt".)

It is instructive to consider a second example in which $K$ and $K'$ are identical except for the generalized composition rule. Although the extensions are the same, $K$ contains one composition rule ($\circ$) and $K'$ another ($\circ'$) where

$$
\begin{array}{c}
\text{o} = \text{Start} \\
\circ \rightarrow \\
\text{Halt} \\
\text{or}_1
\end{array}
$$

and

$$
\begin{array}{c}
\text{o'} = \text{Start} \\
\circ' \rightarrow \\
\text{Halt} \\
\text{or}_1
\end{array}
$$

with

\[
\begin{align*}
\text{or}_1 &= \text{Select two rules $r_1$ and $r_2$ (in $K$ or $K'$)} \\
\text{or}_2 &= \text{Form the composition $r_1 \circ r_2$} \\
\text{dmc} &= \langle A = \{(r_1, r_2) \mid \text{Ran } r_1 \cap \text{Dom } r_2 \neq \emptyset\}, \Lambda \rangle \\
\text{dmc}' &= \langle A = \{(r_1, r_2) \mid \text{Ran } r_1 \cap \text{Dom } r_2 \neq \emptyset, S \in \text{Dom } r_1, \text{and} \text{R } \in \text{Ran } r_2\}, \Lambda \rangle
\end{align*}
\]

The important thing to notice here is that $\text{dmc}'$ (as opposed to $\text{dmc}$) makes direct reference to the S-R pair to be accounted for—i.e., $S$ and $R$ are among the arguments of $\circ'$. (Definitions 22-23 anticipated this possibility.) Allowing $S$'s and $R$'s to enter in this way makes for more efficient search. Obviously inappropriate rules can be eliminated from consideration relatively quickly via decisions rather than allowing a derivation to proceed only to find that the derived rule does not account for the given S-R pair. In short, such decisions may sharply reduce the number of false starts. With small rule sets such as $K$ and $K'$, of course, efficiency is

\[2^{26}\text{Strictly speaking, or}_1 \text{ is a nondeterministic operating rule.}\]
not critical but it becomes increasingly so as rule sets become larger. Furthermore, as we shall see in discussing the idealized (memory-free) theory of learning in Section 2, decisions involving S-R pairs (or equivalently, goal situations) seem to more adequately reflect what human subjects are likely to know and do.

Finally, we consider briefly what is involved in the more general form of the theory of knowledge which involves S-B (rather than S-R) pairs. In this case, we need to consider perception and decoding as well as the simple generation of responses from stimuli. Undoubtedly the easiest way to accomplish this would be to simply add on an atomic encoding and an atomic decoding rule to each rule in K, where K is associated with a corresponding $\mathcal{S} : \mathcal{R}$ class. More generally, of course, we must allow encoding and decoding rules to act on arbitrary substimuli.

The key question in this case is whether all of the respective encoding and decoding of stimuli and behaviors in $\mathcal{S} : \mathcal{B}$ must be done directly (i.e., by one of the encoding or decoding rules attached to the rules in K)---or, whether new encoding and decoding rules may be derived indirectly. In particular, are new perceptual and decoding rules derivable via higher order rules as in the simplified form of the theory above. And, if so, what are the specifics of the process? Although tentative because of the almost complete absence of relevant data, I believe that the above mechanisms account also for the growth of perceptual and decoding skills.

To see how this can be accomplished, we first note that atomic encoding and decoding rules involve stimuli and behaviors, respectively, and insert into or extract from classes (cf. Scandura, 1970b). Such rules do not distinguish among stimuli (behaviors) in any given class. For example, a rule may encode each of a class of stimuli as triangles without distinguishing between large and small ones.

Insofar as generating new encoding (decoding) rules is concerned, the important point is that different encoding (decoding) rules will generally divide up the "environment" (i.e., $\mathcal{S} : \mathcal{B}$) in different ways. Hence, combining such rules via application of higher order rules may generate new encoding (decoding) rules. The basic process involves forming intersections of given dmc's (cf. the discussion following Definition 7). In effect, the rules in K at any particular stage determine which stimuli may be distinguished. As knowledge grows (in $K^2$), finer and finer distinctions may be made. In devising a characterizing rule set, then, the theorist must not only make judgments concerning (internal) operations and decisions but also how the environment is to be initially partitioned.

The exact nature of the process is best seen in terms of an example. We consider a number of stimuli, each of which may be classified according to two independent dimensions.

<table>
<thead>
<tr>
<th>Letter First</th>
<th>Letter Second</th>
<th>Triangle</th>
<th>Circle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small a $#$</td>
<td>$#$ a</td>
<td>Small $\Delta$</td>
<td>$\bigcirc$</td>
</tr>
<tr>
<td>Large $\bigtriangleup$ a</td>
<td>$\bigtriangleup$ A</td>
<td>Large $\bigtriangleup$</td>
<td>$\bigcirc$</td>
</tr>
</tbody>
</table>

Some of Piaget's work, however, is suggestive. His finding that young children are frequently unable to deal with two or more dimensions simultaneously seems to be particularly relevant.
The behaviors are simply descriptions of the stimuli. For example, the behavior associated with "A" is "Small Triangle." Thus

$$\mathcal{A} : \mathcal{B} = \{(a^\#, \text{Small-Letter First}), \ldots, (\#A, \text{Large-Letter Second}), (\triangle, \text{Small Triangle}), \ldots, (\bigcirc, \text{Large Circle})\}$$

In this case, of course, it would be a simple matter to devise a rule set $K$ which accounts for $\mathcal{A} : \mathcal{B}$ directly. A more interesting possibility is provided by

$$K' = \{x, y\} \rightarrow \text{Small}, [x, y] \rightarrow \text{Letter First}, \ldots,$$

where $\{x, y\}$ is the class containing stimulus $x$ and stimulus $y$, and $\Phi$ maps each pair of rules (e.g., $[\triangle, \bigcirc] \rightarrow \text{Large}, [\bigotimes, \bigcirc] \rightarrow \text{Circle}$) into a rule whose domain is the intersection set and whose response is a conjunction (e.g., $[\bigcirc] \rightarrow \text{Large Circle}, \text{or just } \bigcirc \rightarrow \text{Large Circle}$). (It is worth noting that $p$ rules can partition the environment only to the level of atomic stimuli.)

It is easy to see that $K'$ provides an adequate account of $\mathcal{A} : \mathcal{B}$. But, it is even more important to notice how naturally $K'$ can be generalized to deal with larger $\mathcal{A} : \mathcal{B}$ classes by introducing larger input sets (e.g., $\{A, O, H, \ldots\}$) and/or new lower order rules (e.g., $\ldots \rightarrow \text{Hard}$).

### 1.6 Applications

As defined in Section 1.5 a theory of knowledge is extremely general and allows for all of the possible types of competence theory mentioned above. For example, if each rule in $K$ is required to act independently of the others, a behavioral objectives type theory is obtained. In this case $\mathcal{A} : \mathcal{C}$ includes the S-R pairs associated with a given, usually large, class of tasks to be performed. (Each task defines an equivalence class of $\mathcal{A} : \mathcal{C}$.) $K$ is a set of rules which accounts for these S-R pairs, one (or more) rule(s) for each task. Hence, such a theory is of the form $
abla = \langle \mathcal{A} : \mathcal{C}, K \rangle$ where $\mathcal{A} : \mathcal{C}$ equals the union of the extensions of the rules in $K$. Notice that no operations (rules) are allowed on the rules in $K$. By way of illustration, consider a theory in which $\mathcal{A} : \mathcal{C}$ consists of the tasks of addition, subtraction, multiplication, and division (i.e., the set of number pairs paired with their respective sums, differences, products, or quotients) and $K$ contains the addition, subtraction, multiplication, and division algorithms (rules).

Allowing the rules in $K$ to be composed in accounting for $\mathcal{A} : \mathcal{C}$, we get a type of theory which includes generative (phrase structure) grammars. A generative grammar can be characterized by the form $\langle \mathcal{A} : \mathcal{P} \rangle$.

---

28 $[x,y] \rightarrow e$ denotes the rule described by the program: Start: Go to 1; 1: Do $e$, go to 2; 2: If $[x,y]$ go to 3, else go to 5; 3: Do $r$, go to 4; 4: Do $d$, go to 5; 5: Halt. $e$ is an atomic encoding rule which inserts stimuli into $[x,y]$ or its complement, $r$ maps the class $[x,y]$ into $e$ and $d$ decodes $e$ (i.e., chooses a particular representative of class $z$).
\( K, \ast \) where \( \ast \) is a composition rule which acts on the rules in \( K \) but is not itself in \( K \). For example, we might let \( J : \mathcal{K} \to \mathcal{K} \) be any set of pairs whose first element is \( S \) (for sentence) and whose second elements are terminal strings (sentences) of the form \( a^n b^n \) where \( x^n (x = a \text{ or } b) \) means \( x \) repeated \( n \) times. In this case, an adequate account is provided by letting \( K \) contain the two rules \( r_1 = S \to ab \) and \( r_2 = S \to aSb \). The rule \( \ast \) may operate on \( r_1 \) and \( r_2 \) as many times as necessary to generate a rule adequate for generating any particular terminal string. The pair \( S \to aabb \), for example, can be generated by the composite rule \( r_2 \circ r_2 \circ r_1 \) (i.e., apply \( r_2 \) twice, then apply \( r_1 \)).

Adding transformation rules and allowing them to act on phrase structures, yields a type which includes transformational grammars. In this case, a transformational grammar can be characterized by the form \( \langle J : \mathcal{K}, T, t_1, \ldots, t_n, \ast_K, \ast_T \rangle \) where \( T \) is the set \( \{t_1, \ldots, t_n\} \) of transformations which act on (sets of) chains of rules in \( K \) and \( \ast_K \) is composition restricted to the rules in and derivable from \( K \). Similarly, \( \ast_T \) acts on \( T \) together with \( \ast_K \). Consider, for example, the theory \( \mathcal{L} \) with the set of strings \( \{x \to y \} \) as a transformational grammar. In providing an account of a given pair in \( \mathcal{L} \), \( \ast_K \) may operate on rules in \( K \) and \( \ast_T \) to produce new (transformational) rules. \( \ast_K \) is a transformational rule of sorts since it too operates on rules in \( K \).) The new rules produced in turn operate on the rules in \( K \) to generate new rule chains which when applied to the given input, yield the given output. For example, the pair \( S \to abaaba \) may be accounted for as follows: \( \ast \) produces \( \ast \) followed by \( t, \ast_T \). \( \ast_K \circ t \) applied to the rules in \( K \) yields \( r_1, r_2 \to (r_2 \circ r_2 \circ r_1) \).where \( r_1 \circ r_2 \circ r_1(S) = ababa \) (i.e., \( S \to aS \to baS \to ababa \)).

In grammars for natural languages, the chains of phrase structure rules in the domains of transformational rules are identical save for the terminal rewriting (phrase structure) rules. Consider, for example, a transformation which acts on phrase structures of the form \( N_1 - V - N_2 \) (e.g., John hit Jim) and generates new phrase structures of the form \( N_3 - V - be - V - en - by - N_1 \) (e.g., Jim was hit by John). Each such phrase structure can be thought of as a chain of rewrite rules, identical except for the terminal rewrite rules which output specific words. (Chomsky and Miller (1963, 300-301) disallow transformations on undeveloped phrase markers which correspond here to chains of rewrite rules sans the terminal
Structural Learning: Definitional Foundations

ones. They argue that this requires duplication of phrase structure (rewrite) rules after transformation.

We briefly consider an application which involves meaning. In this case, we have

\[ \chi = \langle (S, R) | S \text{ is a base } b \text{ numeral and } R \text{ is the equivalent base } a \text{ numeral}, K = \{ r_1, r_2, g, *, i \}, r_1, r_2, g, *, i \rangle \]

where (1) \( r_1 \) is a rule for interpreting base \( b \) numerals. Its exact nature is unimportant here (For details see Scandura, 1970b; Chapter 7, Volume I) but its extension has base \( b \) numerals as inputs and can be thought of as having meanings as outputs (i.e., numbers—sets containing appropriate numbers of elements). (2) \( r_2 \) is a rule for generating base \( a \) numerals given (as input) arbitrary numbers. (3) \( g \) is a rule which operates on rules for changing numerals in one base (e.g., \( b \)) into numerals in another base (e.g., \( a \)) together with a new pair of base numerals \((b', a')\). One element in its domain is \( (r_1 \circ r_2, b', a') \) where \( b' \) and \( a' \) are given by the S-R pair to be accounted for. \( g \) generates new rules for changing (new) numerals in base \( b \) into numerals in base \( a' \). Notice that \( g \) generates an indefinitely large class of rules.

(4) \(* \) is a composition rule and \( i \) is the identity.

Given the rule set \( K \) and the to-be-accounted-for pair S-R (i.e., a pair of numerals in base \( b \) and \( a' \), respectively, we get

\[ K^2 = K \cup \{ r_1 \circ r_2, *, \circ g \} \]

since \( *(r_1, r_2) = r_1 \circ r_2 \) and \( *(r, g) = * \circ g \)

\[ K^3 = K^2 \cup \{ (r_1 \circ r_2)' \} \cup A \]

where \( A \) contains the necessary compositions and \( g(r_1 \circ r_2, b', a') = (r_1 \circ r_2)' \) where \( (r_1 \circ r_2)' \) is a rule for changing a numeral in base \( b \) into a numeral in base \( a' \).

This example shows explicitly how \( K^n \) may depend not only on \( K \) but also on the S-R pair to be accounted for. Also notice that although our theory provides a semantic account, this is not necessary in working with numerals. It would be a simple matter to construct a purely syntactic account which makes no reference to meaning. Presumably, the role of semantics only becomes crucial where \( \mathcal{J} : \mathcal{K} \) is sufficiently complex (e.g., as in natural language translation).

1.7 Axioms, Comments, and Conjectures

In general, the nature of PK will depend on the nature of \( K \) and the nature of \( \mathcal{J} : \mathcal{K} \) will determine characteristics of any \( K \) which accounts for it. For instance, notice that \( \mathcal{J} : \mathcal{K} \) may or may not contain pairs including programs that correspond to rules in \( K \). This suggests the following kinds of questions (among others) for future research: (1) Given properties

30A number of examples of rules like \( g \) can be found in Scandura, Durnin, Ehrenpreis & Luger (1971).
concerning K, determine properties of PK which follow. (2) Given properties of $\mathcal{S}:\mathcal{K}$, determine properties of any K which accounts for it. The more conditions that can be placed on $\mathcal{S}:\mathcal{K}$, of course, the easier in practice it will be to find a suitable K. Finding necessary and sufficient conditions is the ultimate goal. (3) Given two theories $\mathcal{K} = \langle \mathcal{S}:\mathcal{K}, \mathcal{K}, ... \rangle$ and $\mathcal{K}' = \langle \mathcal{S}:\mathcal{K}, \mathcal{K}', ... \rangle$ such that both K and K' account for $\mathcal{S}:\mathcal{K}$, determine and develop conditions for comparing K and K'.

As described so far, the theory is essentially a schema which may take on any number of forms depending on the particular use to which it is to be put. The axioms for a behavioral objectives type theory, for example, are likely to be far more specific than those required to characterize the more general form of the theory.

Even in the latter case, of course, there are many different forms any particular theory might take. If concerned with the ontogeny of knowledge (e.g., the characterization of knowledge in a way which might reflect its growth from birth), for example, it might seem reasonable to assume that each rule in K has a finite domain. Indeed, in order to insure only the simplest of competencies at birth, one might further require that each rule be atomic and consist entirely of one-instance operating rules (e.g., generalized activity, instincts, reflexes) or decision making capabilities that are two element partitions, one set containing one (encoded) stimulus and the other containing the absence of this stimulus. (The latter corresponds to making decisions on the basis of 0's and 1's.) Two such operating rules might be described, "Suck when something soft is in mouth," and "Spit when something soft is in mouth." A related dmc might involve distinguishing between situations in which one's stomach is full, and situations where it is not. Composition, probably restricted initially to a small class of rules, might provide an example of a higher order rule. These examples suggest at once the possible importance of generalized tendencies and "instincts" to the growth of knowledge and the difficulties likely to be involved in identifying all of the crucial ones.31

If, on the other hand, a theory is concerned with "ongoing" knowledge, then this level of detail makes no sense. People really do know rules that have infinite domains (although they cannot possibly apply the rule in all cases) and a viable theory of knowledge should reflect this. Allowing such rules in K seems to leave the general scheme unchanged and, furthermore, it is a simple matter to show how rules with infinite domains can be constructed from rules with finite domains by application of higher order rules to lower order ones. The basic idea is simply that the introduction of dmc's (nodes) into (new) rules may lead to "loops" that can be repeated an arbitrary number of times.

Whatever the specific nature of a theory of knowledge, however, it is almost certain to share some common properties. Two that are likely

---

31It is worth noting in this regard that although Piaget has been centrally concerned with the ontogeny of knowledge he has for the most part ignored individual differences and concentrated on that which is common to human beings generally. Indeed, in the case of Piaget, $\mathcal{S}:\mathcal{K}$ would correspond roughly to the class of behaviors generated via adult logic. K, then, would correspond to the initial base upon which knowledge grows through the various developmental stages.
Structural Learning: Definitional Foundations

to prove of general value in applications are the following

**Axiom 1:** For each S-R pair in $\mathcal{J}: \mathcal{K}$ which is first order generable from PK, there exists a finite number $n$ such that the S-R pair is $n$th order generable from $K$.

Axiom 1 disallows "limit" behaviors which cannot be derived in a finite number of steps.

**Axiom 2:** For each class of S-R pairs in $\mathcal{J}: \mathcal{K}$, there is some finite set of rules which accounts for it.

Axiom 2 insures that $\mathcal{J}: \mathcal{K}$ is not pathological and is designed to avoid problems involving diagonalization arguments (e.g., Rogers, 1967), criticisms by certain linguists (cf., Marcus, 1967) interested in natural languages, and the like.

The ontogeny of knowledge is partially described by the following

**Definition 40:** $K_1$ is an innate basis for $K_2$ iff for some finite number $n$, each rule in $K_2$ corresponds to some rule in $K_1^n$ with the same extension.

**Definition 41:** $K_1$ is a finitary innate basis for $K_2$ iff $K_1$ is an innate basis in which each rule has a finite domain.

**Definition 42:** $K_1$ has greater simple computing power than $K_2$ iff the union of the extensions of the simple rules in $K_1$ contains the union of the extensions of the simple rules in $K_2$. (Recall that the extensions of simple rules do not involve rules.)

As will become apparent in Section 3, the notion of simple computing power has relevance in determining performance where memory is a factor.

**Finitary Conjecture:** Each $K_2$ has a finitary innate basis $K_1$.

Further, there is a finitary innate basis in which each rule is either an atomic rule with one input or a simple decision rule consisting of a dmc that partitions a two element class in half adjoined with two discrete atomic rules with one-element domains.

**Knowledge Potential Conjecture:** If $K_1$ is an innate basis for $K_2$, then $\exists m \geq n$ such that $K_1^m \geq K_2^m$.

If true, this latter conjecture could have a number of interesting implications. For example, suppose $K_1$ and $K_2$ represent the competencies had by two newborn babies, where $K_1$ is an innate basis for rule set $K_2$, but $K_2$ has greater simple computing power than $K_1$. Then, it would follow that the baby characterized by $K_1$ would have greater knowledge potential but lower initial ability. Thus, behavior that might be accounted for directly (simply) by $K_2$ would require $K_1^n$ for some $n > 1$. Coupled with the well known fact that boys tend to develop more slowly than girls but generally catch and surpass them in certain areas (while presumably the reverse is true in other areas), this observation might provide interesting food for thought by women libertarians. Perhaps the Frenchmen have been right all along. More seriously, this implication
could provide a basis for explaining a number of phylogenetic anomalies (e.g., that baby chimps are initially "smarter" than human babies) and, indeed, could place the whole nature-nurture question in new perspective.

2. FOUNDATIONS OF THE IDEALIZED (MEMORY-FREE) THEORY

In the theory of knowledge $\mathcal{K}$ the theorist's task is to devise a finite rule set $K$ which accounts for a given class $\mathcal{A} : \mathcal{B}$. In order to have behavioral relevance, each rule known to a given subject(s) and relevant to the given class $\mathcal{A} : \mathcal{B}$ must be embedded in some rule in $K$ in a sense made explicit below. Put differently, $K$ must include all of the rules the subject or subjects are likely to know.

The idealized (memory-free) theory\textsuperscript{32} of structural learning, then, begins where $\mathcal{K}$ leaves off. There are two major additional problems with which the behavior theorist must deal. First, he must have some way of identifying the knowledge had by the subject in terms of that identified in his role as observer (qua competence theorist). Second, the theorist must have some way of explaining and predicting the subject's learning and performance.

The basic approach rests on the following assumptions:

1. The behavior theorist either knows or can manipulate the subject's goal (i.e., what he is trying to do) with some known degree of specificity; where this is not the case the theory can only be used for explanatory purposes, not prediction.\textsuperscript{33}

2. Performance is goal directed. This assumption provides a basis for determining which (parts) of the rules in $K$ individual subjects know.

3. Learning and motivation depend on shifting control between given and higher order goals in a specified manner.

The formalism which follows involves these assumptions. For convenience, we limit our examples to $\mathcal{A} : \mathcal{K}$.

\textsuperscript{32}Because this theory applies only where all of the relevant rules are immediately available to the subject, several people have suggested the label "Memory-Available."

\textsuperscript{33}Such an approach has been criticized (e.g., see Witz, Chapter 4; Knifong, 1971) as not providing a "dynamic" model which directly reflects observed behavior. In reaction, I would simply point out that the present approach provides a basis not only for explanation (cf. Piaget in Flavell, 1963, Furth, 1969; Witz, Chapter 4) and description (e.g., Knifong, 1971) but also for prediction, something which the so-called dynamic models have been unable to provide. Indeed, although I shall not attempt to justify my belief, I suspect that the dynamic models described by Witz, for example, will turn out to correspond to situations where the observer knows relatively little about what the subject is trying to do ahead of time, and can only determine this after the fact.
2.1 Goal Situations

Goals provide the theorist-observer with criteria for judging behavior. They are also assumed to control the subject's behavior. Definitions 1-4 make these notions precise.

Let Dom be a class of inputs (I) or stimulus situations (S) and Ran be a class of outputs (O) or behaviors (B). P is a binary predicate which defines a nonempty binary relation p in Dom x Ran.

**Definition 1:** A goal G is a class \[ \{ O \mid (I,0) \in p \text{ for some } I \in \text{Dom} \} \]

All goals can be described by an imperative statement of the form: "Find a behavior B such that B satisfies goal criterion (predicate) \( P(S,B) \) for some \( S \in \text{Dom} \)." Equivalently, "\( \exists B, P(S,B) \)."

Notice that, whereas goals may be defined in terms of nonobservables (e.g., \( I, O \)), goal descriptions are strictly observable. Interpreting goal descriptions yields goals. For example, the goal \( \{ 0 \mid 0 = a + b \text{ where } a, b \text{ are natural numbers} \} \) may be described "Find a response R, such that \( R = a + b \) for the pair of natural numerals \( a, b \)."

By themselves, goals are insufficient for determining performance. They tell where one is going but not where one is. Stimulus situations or, more generally, inputs, provide the occasion for responding and together with goals give

**Definition 2:** A goal situation \( < I_0, G > \) is a class \[ \{ O \mid (I_0,0) \in p \text{ where } I_0 \in \text{Dom} \} \]

Notice that goal situations can be characterized as pairs consisting of an input \( I_0 \) and a goal \( G \). All goal situations can be described by a statement of the form: "Find a behavior B such that B satisfies \( P(S_0, B) \) where \( S_0 \in \text{Dom} \)." In practice, of course, goal situations may be described in a variety of equivalent ways. The statement "Find \( 3 + 4 \)," for example, is usually interpreted to mean "Find \( n \) such that \( n = 3 + 4 \)."

In addition to the stimulus \( S_0 \), we may on occasion also want to refer to other stimulation in the environment, possibly including programs for rules. In this case we use the term environmental complex \( E \).

**Definition 3:** A goal environment is a pair consisting of a goal situation \( < I_0, G > \) and an environmental complex \( E \).

Clearly, there is a close relationship between dmc's and goals. The former are partitions whose elements are categories (sets) of outputs. The latter are (single) categories containing just those outputs which "satisfy" a goal. More precisely, the relationship is specified by a map \( t \in \text{Definition 4: If dmc is a decision making capability, then } t_n(\text{dmc}) \text{ is that goal given by the nth category in dmc.} \]

For example, \( t_1(\{ \langle R|R = 5 + 4 \rangle, \{ R|R \neq 5 + 4 \} \} ) = \{ \langle R|R = 5 + 4 \rangle \} \) and \( t_2(\{ A = \{ r \mid (5,4) \in \text{Dom} r, \text{ and for all } R \in \text{Ran} \, r, R = a + b \text{ for some } (a,b) \}, \text{A (i.e., complement of A)} \} ) = \text{A}. \)
There are two senses in which a response may satisfy a given goal situation. The goal itself requires only that \( B \) be a member of a particular class (e.g., the class of sums \( a + b \)). We call this the \textit{a priori} criterion of the goal situation. This criterion defines only the range (type) of behaviors allowed and is independent of the particular stimulus. The stronger\(^{34}\) sense of satisfying a goal situation depends also on the stimulus (e.g., \( B = 3 + 4 \)). In this case, the criterion involved is referred to as \textit{stimulus dependent}. This distinction between \textit{a priori} and stimulus dependent criteria plays an important role in the learning and performance mechanism described below.

2.2 \textbf{Nature of the Idealized (Memory-Free) Theory}

A memory-free theory of learning \( \mathcal{L} \) may be defined as an \( n + m + 3 \) tuple \( \mathcal{L} = \langle \mathcal{S}; \mathcal{B}, K, r_1, \ldots, r_m, S, < S_1, G_1 >, \ldots, < S_m, G_m > \rangle \)

where \( \mathcal{S}; \mathcal{B}, K, r_1, \ldots, r_m \) are as in \( \mathcal{L} \), \( S \) is a behaving subject, and \( < S_1, G_1 >, \ldots, < S_m, G_m > \) are goal situations.

If the theory is to have predictive value, then the goal situations must be capable of manipulation and equally meaningful to both \( S \) and the observer. Our main task in this section is to show how \( \mathcal{L} \) provides a basis for determining what \( S \) knows and how \( S \) performs and learns.

First, we introduce the notion of path of a rule and show how the paths of a rule impose a partition on the rule's extension.

\textbf{Definition 5:} A (completed) \textit{computation} of stimulus \( S \) by rule \( r \) is the finite sequence \( S = S_1, o_1, o_2, \ldots, o_n = B \) (where \( S \) is in the domain of \( r \) and \( o_1 \) is the output generated by the preceding atomic operating rule \( o_1 \)) obtained by applying in turn the atomic operating rules of \( r \) to \( S \). Notice that \( \text{DMC}'s \) are involved in determining which \( o \) is applied next but do not appear in the computation itself.

\textbf{Definition 6:} A \textit{simple path} of a rule \( r \) on stimulus \( S \) is the finite sequence of atomic operating rules \( o_1, o_2, \ldots, o_n \) of \( r \) obtained by deleting the stimulus and outputs from the computation of \( S \) by \( r \).

Intuitively, the simple path associated with any given stimulus \( S \) and rule \( r \) consists precisely of those atomic operating rules of \( r \) which are applied to produce the corresponding output, and in the order in which they are applied (including repeats).

\textbf{Definition 7:} The set of \( S-B \) pairs associated with a simple path of rule \( r \) is the class of \( S-B \) pairs that the simple path computes.

\textbf{Definition 8:} A \textit{path} (form) is determined from a simple path by eliminating consecutive repetitions of subsequences. (Each repetition may be replaced by \( * \).)

For example, the simple path \( o_1, o_2, o_3, o_4, o_5, o_6 \) is replaced with the path \( o_1 \).

\(^{34}\)Obviously, if \( B \) satisfies a stimulus dependent criterion, then it necessarily satisfies the corresponding \textit{a priori} criterion.
Definition 9: Two simple paths are equivalent if they generate the same path form (up to \(*\)).

Theorem A: "Is the same form as" is an equivalence relation on the set of simple paths associated with a rule. Further, this equivalence relation partitions the extension of the rule.

Proof: "Is the same form as" is an equivalence relation because each simple path is mapped individually and unambiguously into a path form. Hence, each simple path is in some path. These paths are either identical or distinct so the relation satisfies the reflexive, symmetric, and transitive properties.

To show that this equivalence relation imposes a partition on the extension we can assume without loss of generality that the rule is defined on its entire domain. Then, each S-B pair is associated with some simple path of the rule and, in turn, with some path. To show that the equivalence relation is mutually exclusive (i.e., that the classes of S-B pairs associated with the paths of the rule are disjoint), we need only note that there is one and only one simple path, and hence path, associated with each S-B pair.

Theorem B: Each rule has a finite number of different paths.

Proof: The observation is obviously true of rules with a finite number of simple paths. Suppose the rule has an infinite number of simple paths. Then, all but a finite number of them must include at least one repetition (i.e., one repeated subsequence) because each rule contains at most a finite number of atomic or's. We need to show that there can be at most a finite number of different repetitions. But, this follows directly since there are only a finite number of subsequences which can be repeated.

Corollary A: The paths associated with a rule partition its extension into a finite number of equivalence classes.

Corollary B: The paths of each subrule of a rule impose a finite partition on the extension of the subrule.

As an exercise, consider the rule 

\[ r_1 = \text{Start} \xrightarrow{a} \text{Halt} \]

\[ r_2 = \text{Start} \xrightarrow{\text{non-atomic symbols}} \text{Halt} \]

where \( r_1 = xxBy \rightarrow xBOy \) and \( r_2 = xx\text{aBy} \rightarrow \text{xBly} \) (\( x \) is a string of a's and \( y \) is a string of 0's and 1's). Identify a computation, a simple path, and a path of this rule. Also identify the partition imposed by the paths of this rule on its extension.
In discussing the problem of assessing behavior potential we use the following terminology.

**Definition 10**: A goal situation \(<S_0, G>\) is said to be **resolved** iff some behavior B is observed which either satisfies \(<S_0, G>\) or does not satisfy it. We say that a subject achieves the goal of \(<S_0, G>\) if he generates a behavior which satisfies \(<S_0, G>\). In this case, he is also said to succeed. He fails if he does not succeed.

In \(\mathcal{A}\) the subject is theoretically given all of the time he needs to respond. Practically speaking, however, to allow for situations where the computation may not Halt, we must impose some time criterion.

**Definition 11**: A **trial** consists of a goal situation \(<S_0, G>\) and its resolution. An **assessment episode** is a trial in which \(<S_0, G>\) is satisfied by an \(r\) in \(K\). An **assessment program** (criterion referenced test) is a finite sequence of assessment episodes, one for each path of each rule in \(K\). A **learning episode** is a trial in which \(<S_0, G>\) is **not** satisfied by an \(r\) in \(K\). A **training program** consists of a finite sequence of learning episodes.

**Definition 12**: A goal \(<S_0, G>\) is satisfied by a rule \(r\) if \((S_0)\) satisfies \(<S_0, G>\). We also say that \(<S_0, G>\) is satisfied by the particular path of \(r\) associated with \(<S_0, G>\) where the output \(B_0\) of the computation associated with \(S_0\) and \(r\) satisfies \(<S_0, G>\).

A subject \(S\) is said to **know** a path iff he can perform perfectly in all goal situations satisfied by that path. The rule corresponding to the set of known paths of a rule \(r\) is denoted \(r_S\) (for subject's rule). The set of all known rules in \(K\) is denoted \(K_S\).

Given any computable function (i.e., set of input-output pairs generable by some rule), it is well-known that there is a countably infinite number of rules which might account for it. In practice, of course, the number of such rules that might reasonably be employed by a human subject is typically quite small (and often just one). Subtraction, for example, is typically performed via one of two methods, borrowing or equal additions.

In general, it is possible to distinguish two or more rules extensionally (i.e., in terms of observables) just to the extent that they impose different partitions on their common extension. Hence, for purposes of prediction, we identify all rules imposing identical partitions. That is, we do not discriminate among such rules since for purposes of explanation and prediction we cannot distinguish among them. Indeed, because we shall eventually want to allow rules to act on rules, we identify all subrules with identical partitions. (We do not attempt a formal justification for this statement but simply note that in operating on rules a rule may act on subrules.)

**Definition 13**: Two or more rules (subrules) are said to be **partitionally equivalent** (equivalent) iff they impose identical partitions on a common extension.

Where alternative rules (e.g., borrowing and equal additions—see Durnin & Scandura, 1971) impose different partitions on an extension, the situation is somewhat different. Assuming (as we do below) that paths of a rule are either totally available or completely unavailable,
then different patterns of behavior (i.e., successes and failures) on a
given extension may more directly reflect one of the rules rather than
the others. The pattern of successes and failures in subtraction, for ex-
ample, depends to a great extent on whether borrowing or equal additions
is used. Furthermore, it is reasonable to expect that some subjects may
use a combination of two or more such rules.

For this reason, whenever it seems likely that two or more nonequi-

damental rules (with a common extension) may be used in generating behavior,
we replace the nonequivalent rules with a single rule that is equivalent
to them. To see how such a rule may be devised it is sufficient to note
two things: (1) A set of n nonequivalent rules with a common extension
imposes a refined partition on the extension which consists of all n-fold
intersections of equivalence classes involving one equivalence class from
each of the n partitions associated with the n nonequivalent rules. More
exactly, if \([A_{11}, A_{12}, \ldots, A_{1m}], [A_{21}, A_{22}, \ldots, A_{2m}], \ldots, [A_{n1}, A_{n2},
\ldots, A_{nm}]\) are partitions imposed by nonequivalent rules \(r_1, r_2, \ldots,
\ldots, r_n\), then the refined partition is \([A_{11} \cap A_{21} \cap \cdots \cap A_{1m}]\) where
\(1 \leq i_1 \leq m_1, 1 \leq i_2 \leq m_2, \ldots, 1 \leq i_n \leq m_n\). (2) It is always possible
to devise a new rule which imposes the refined partition on the common
extension. Furthermore, any other rule which imposes this refined parti-
tion is equivalent.

From now on we shall assume that the extensions associated with the
rules in \(K\) are disjoint and further that no two rules have the same exten-
sion. In addition, to insure that \(K\) provides an adequate basis for \(\mathcal{E}\), it
is implicitly assumed that all rules known to \(S\) that are relevant to \(\mathcal{A}\)
\(\mathcal{B}\) are embedded in the rules in \(K\). It is also assumed that the rules in
\(K\) are refined to the point where each atomic or and dmc acts in an all-
or-none fashion. (This is always possible in principle since if I inter-
pret the notion properly every Turing machine has this property.)

The following two axioms provide a basis for assessing behavior
potential (i.e., finding out which paths of which rules in \(K\) the \(S\) knows).

**Axiom 1:** There is exactly one nonempty criterion referenced test
and one training program.

**Axiom 2:** If a subject achieves an \(\langle S_0, G \rangle\) that can be satisfied
by a path of a rule \(r \in K\) (there is exactly one such rule), then
he knows the path (i.e., can achieve all \(\langle S_0, G \rangle\)'s for that path);
if he fails on one such \(\langle S_0, G \rangle\), then he will fail on all.

**Theorem (Assessing Behavior Potential):** Given a subject \(S\) and a
class of \(S-B\) pairs which can be accounted for by an \(r \in K\), then there
is a finite number of \(\langle S_0, G \rangle\)'s which can be used to determine
which paths of \(r\) the \(S\) knows (i.e., to determine the corresponding
\(r^S\)).

**Proof:** We prove the theorem by showing how to select the \(\langle S_0, G \rangle\)'s.
Each \(r \in K\) has a finite number of paths and thus partitions the
given class (of \(S-B\) pairs). Select one \(S_0-B_0\) pair from each equi-
valence class in this partition. Taking Axiom 2 into account, the
\(\langle S_0, G \rangle\)'s which correspond to these \(S_0-B_0\) pairs can be used as
follows to determine which paths of \(r\) are known: If the subject
achieves an $\langle S_0, G \rangle$ satisfied by a path of $r$, then he knows that path; otherwise he does not.

This theorem provides a basis for assessing a subject's behavior potential (relative to a given $K$).

**Corollary A.** Given any $K$, a finite number of $\langle S_0, G \rangle$'s is sufficient for determining $K$.

**Proof:** This follows because the number of rules in $K$ is finite.

**Corollary B:** If all rules in $K$ are atomic, then just one $\langle S_0, G \rangle$ is needed for each rule in $K$.

**Proof:** This follows because atomic rules have only one path.

A. As an exercise, consider the rule $r$ above together with the related rule $r'$

\[
\begin{align*}
\text{Start} & \quad r_1' \quad r_2' \quad \text{Halt} \\
& \quad r_3' \quad r_4'
\end{align*}
\]

where $r_1' = xxxxBy \rightarrow xB00y$

$ r_2' = xxxxabY \rightarrow xB01y$

$ r_3' = xxxxaaBy \rightarrow xB10y$

$ r_4' = xxxxaaaBy \rightarrow xB11y$

\[dmc' = \langle \{xxxxBy\}, \{xxxxabY\}, \{xxxxaaBy\}, \{xxxxaaaBy\}, \{By\} \rangle.\]

Also let the common extension be $\{(wB, Bz) | w$ is a string of a's and $z$ is the binary numeral representing the number of a's}. Identify a single rule which is partitionally equivalent to $r$ and $r'$ collectively.

B. Suppose $r$ is in $K$ and that $\xi$ is tested and responds as indicated below.

\[
\begin{align*}
aaaB & \rightarrow 11 \\
aab & \rightarrow 1
\end{align*}
\]

Use the axioms to identify which expressions (of the form $wB$, $w$ a string of a's) $\xi$ may be expected to respond to successfully and which not.

C. The behavioral reality of this approach is demonstrated in Volume I, Chapter 7 (Section 3). See Durnin and Scandura (1971) for a more complete discussion of the experimental results and comparison of the algorithmic approach formulated here with item forms and hierarchical technologies.
The assessment procedure described above clearly provides an explicit basis for determining $K_g$. Our second major step is making precise the mechanisms by which the rules in $K_g$ are put to use and new rules are acquired (i.e., $S$ learns).

We begin by defining the notion of higher order goal relative to a given goal situation. Consider goal situation $<S_0, G> = [R | P(S_0, R)]$ where $P$ is the predicate of $G$. Then

**Definition 14:** The second order goal relative to $<S_0, G>$ is

$$G^2 = \{ r | S_0 \in \text{Dom} \ r, \text{Ran} \ r \subseteq G \}.$$  

Similarly,

$$G^3 = \{ r | S_0 \text{ together with the } r' \text{'s in } K_g \text{ specify an element } e \text{ of } \text{Dom} \ r, \text{Ran} \ r \subseteq G^2 \}.$$ 

$$\vdots$$

$$G^n = \{ r | S_0 \text{ together with the } r' \text{'s in } K_g \text{ specify an element } e \text{ of } \text{Dom} \ r, \text{Ran} \ r \subseteq G^{n-1} \}.$$ 

Notice that the elements in the domain of any $r$ in a goal of third or higher order consist (partly) of rules in $K_g$. Also notice that rules which satisfy higher order goals need not necessarily satisfy $<S_0, G>$. In particular, it may be that $S_0 \in \text{Dom} \ r$ and $\text{Ran} \ r \subseteq G$ but $R_0 = r(S_0)$ may not be in $<S_0, G> = [R | P(S_0, R)]$. As we shall see below, our definition allows for "false starts" in problem solving (cf. Axioms).

**Definition 15:** The $m$th stage of a trial is a triple $<S_0, G, G^m, K_g^m>$ where $<S_0, G>$ is the goal situation of the trial, $G^m$ is the $m$th order goal relative to $<S_0, G>$ where $m' < m$, and $K_g^m$ is the set of rules known to the subject (at the $m$th stage). The $m$th order goal is said to be in control during the $m$th stage of a trial.

The following axioms spell out the conditions governing learning and performance under memory-free conditions.

**Axiom 3:** At stage $<S_0', G', G^m', K_g^m>$ if there is exactly one $r \in K_g^m$ that satisfies $G^m'$, then control goes to $G^{m'-1}$ (at the $m + 1$st stage) where $m' \geq 2$ and $r$ is applied. Where $m' = 1$, the computation Halts. Otherwise (i.e., if $G^m'$ contains no rule in $K_g^m$ or more than one), control goes to $G^{m+1}$ (i.e., the next stage is $<S_0, G, G^{m+1}, K_g^{m+1}>$) if $G^{m+1}$ exists. Where $G^{m+1}$ does not exist, the computation halts (unless there is more than one rule in $K_g^m$ in which case the selection is nondeterministic).

Recall that the existence of $G^m'$ for a given $m$ depends on the availability of a suitable dmc. Dmc's unfortunately do not reside in $K_g$ as such where $K_g$ is the set of rules available to $S$ at the beginning of the trial. They are nonoperational in the sense that their presence...
cannot be determined directly. The best our assessment procedure can do is to specify a "decision" rule consisting of the dmc in question together with or's which attach common responses to stimuli in each equivalence class of the partition (dmc).

**Axiom 4:** An output rule r generated at any stage automatically becomes part of the available knowledge at the next stage (i.e.,

\[ K_{sn}^{m+1} = K_{sn}^m \cup \{r\} \]

If we denote the knowledge had by S at the beginning of trial n by \( K_{sn} \), then \( K_{sn}^m \) is the knowledge had by the subject during the mth stage of the nth trial. Notice that any finite number of rules may be added to \( K_{sn} \) on any given trial.

**Axiom 5:** If the output r generated at any stage satisfies \( G^m \), (irrespective of how many rules in \( K_{sn}^m \) satisfy \( G^m \)), then control reverts to \( G^{m-1} \) where \( m' \geq 2 \) or STOP where \( m' = 1 \) and r is applied. Otherwise, the last rule (in \( K_{sn}^m \)) selected in the course of the derivation is (temporarily) eliminated (from \( K_{sn}^m \)) and control at the next stage shifts to \( G^n \) where \( n' \) is the level in the preceding derivation at which the eliminated rule was selected.

The reason for eliminating rules (temporarily) from K is to allow for "false starts" and subsequent attempts at problem solving. Some such mechanism seems essential if we are to reflect human behavior.

Although the above definitions and axioms (in one form or another) may be expected to play a central role in the idealized theory, they alone will probably not be sufficient. In any complete theory, for example, it will be necessary to make explicit the sense in which the (idealized) theory of learning depends on the underlying competence theory. Definition 16 together with Axiom 6 provide one way of accomplishing this.

**Definition 16:** (a) Given a goal situation \( \langle S_0, G \rangle \), G is reachable from stimulus \( S_0 \) via a subset \( \mathcal{A} \subseteq K \) if there is an \( R_0 \subseteq \mathcal{R} \) which satisfies G such that \( S_0 \rightarrow R_0 \) is nth order generable from \( \mathcal{A} \) (i.e., such that there is an \( r \in \mathcal{A}^n \) (for some n) which accounts for \( S_0 \rightarrow R_0 \)). (In this case, we also say that \( \langle S_0, G \rangle \) can be satisfied by a subset \( \mathcal{A} \subseteq K \).)

(b) A rule \( r \in K \) is relevant to \( \langle S_0, G \rangle \), if there is a subset \( \mathcal{A} \subseteq K \) with \( r \in \mathcal{A} \) such that \( \langle S_0, G \rangle \) can be satisfied by \( \mathcal{A} \) but not by \( \mathcal{A} - \{r\} \).

Because no operations are performed, dmc's themselves may be assumed to take place instantaneously—or at least in some short fixed time. Of course, the operations which must follow if a dmc is to be detected do result in a measurable response latency. Latency measures may provide one method for determining the internal structure of rules. (For one approach to this problem, as well as a concise summary of its long history, see Sternberg, 1969. Also see Chapter 8, Volume I for a related, detailed discussion of information processing.)
(c) The set of rules $K_{\text{rel}} (S_0, G) = \{ r \in K | r \text{ is relevant to } (S_0, G) \}$ relevant to $(S_0, G)$ is called the relevance class of $(S_0, G)$.

(d) The relevance class of a training program $K_{\text{rel}}$ is the union of the relevance classes of the $(S_0, G)$'s in the training program.

**Axiom 6:** Let $K_{\text{rel}}$ be the relevance class of the training program of the theory $\mathcal{L}$ relative to the set $K$ of rules actually known to a $S$. Then, $K_{\text{rel}} \subseteq K$ where $K$ is the rule set (in $K$) introduced by the competence theorist.

Axiom 6 says, in effect, that the theory $\mathcal{L}$ can be no better than the theory $K$ on which it is based.

Another assumption that needs to be made explicit concerns the extent to which an experimenter may control behavior by manipulating goals. In particular, it is necessary to assume that presenting a goal situation invariably results in $S$ adopting the goal. Otherwise, the theory would lose predictive value and be useful only for explanation. More precisely, we assume not only that $S$ is able to interpret any goal situation description presented by the experimenter but also that he will do so. Furthermore, once understood this goal is assumed to be in control at the first stage of the trial.

We shall not attempt to develop these fine points or to draw out further implications of the theory. This will be the subject of future papers.

### 2.3 Applications

(1) Let $K_S = \{ r_1, r_1', r_2, r_2', \ldots, r_n, r_n', o, i, d \}$ where $r_i$ and $r_i'$ are rules (with disjoint extensions) such that $\text{Ran } r_i \subseteq \text{Dom } r_i'$ for $i = 1, 2, \ldots, n$, $o$ is composition, $i$ is identity, and $d$ is a decision rule for determining for any given $r$ and $(S, G)$ whether or not $S \in \text{Dom } r$ and $\text{Ran } r \subseteq G$. Also let $(S_0, G) = \{ r | P(S_0, r) \}$ where $S_0 \in \text{Dom } r_3$ and $\text{Ran } r_3' \subseteq G$. Then, $G^2 = \{ r | S_0 \in \text{Dom } r, \text{Ran } r \subseteq G \}$. The composite $r_o r_3'$ satisfies $G^2$ and can be generated by applying $o$ to $(r_3, r_3')$. Hence, $(S_0, G)$ can be satisfied and $S$ will learn the rule $r_3 o r_3'$ in the process.

Now let $(S_o', G)$ be such that $S_o' \in \text{Dom } r_{n+1}$ and $\text{Ran } r_{n+1}' \subseteq G$ where $r_{n+1}$ and $r_{n+1}' \in K_S$. Can $S$ satisfy $(S_o', G)$? How could $K_S$ be modified so that this would be possible?

---

As this book goes to press, I am in the middle of a project on mathematical problem solving (NSF Grant #GW6796) that among other things is aimed at drawing out implications of this theory for artificial intelligence and mathematics education. In the future I also hope to draw out implications of the theory for optimizing learning.
314 Scandura

(2) **Semantics** can be brought into the picture by modifying the example above so that $p_1$ is a program for $r_1$ and $m$ is a rule for assigning meanings to arbitrary programs of the given types. Let $K_s' = \{ m, \circ, \text{conj}, \triangleright \}$ where conj is conjunction. Furthermore, we allow programs $p_1$ in the goal environment $\langle \langle S_0, G \rangle, E \rangle$ where $p_1 \in E - S_0$. In this case, even a goal situation that can be satisfied by one of the rules $r_1 \in K_s$ (above), strictly speaking, requires a derivation. Here, the meaning rule $m$ is applied to $p_1$ in the environment to generate $r_1$ which satisfies $G^2$. (For $G^2$ to be meaningful, of course, we would need to include a suitable decision rule in $K_s'$.)

A third level derivation is required where a given goal situation can be satisfied only by a composition $r_1 \circ r_1'$. To derive $r_1 \circ r_1'$ we need a conjunction (conj) of $m$ followed by composition ($\circ$), applied to $p_1$, $p_1'$. But to get this we need to apply "conj $\circ$ composition" to $m$ and $\circ$. This latter rule results on applying "$\circ$" to conj and $\circ$.

The distinction between rules and programs may be suppressed in experiments where meaning is not at issue. Indeed, this was the case in all of the experiments reported in Chapters 6 and 7 of Volume I and the first article in this chapter. The study on verbal and symbolic statements of rules reported in Chapter 2 involves meaning more or less directly.

The reader with a computer science orientation may wonder why we have made a distinction between rules and shifting goals since the role of the latter may be incorporated into single unified procedures (rules). Thus, in example (2) it would be a simple matter to devise a unified procedure that would encode and operate on programs in the environment to produce new procedures (rules)--and, then, would "turn around" and apply these newly formed procedures to that portion of the environment that we have called simple stimuli. In effect, the interpretation and application of programs $p_1$, which we have explained above in terms of shifting goals, could also be viewed in terms of applying a single procedure to the environment.

It would appear, therefore, that these two views are mathematically equivalent. But, they are not equivalent psychologically. The present view not only allows for a greater variety of behavior (because the higher and lower rules may act separately, for example, as well as in combination), but, more importantly, also provides theoretical constraints on rule based accounts which are directly reflected in human behavior (cf. Volume I, Chapters 7 and 8). To the extent that such constraints are ignored, each behavior theory reduces to a rule; that is, the rule becomes the theory.

(3) Perceptual learning can be illustrated by extending the example on p. 299 as follows. Let $K' \subseteq K_s$ where $\emptyset$ "forms the intersection." In this case, the goal $G$ in $\langle S_0, G \rangle$ consists of pairs of properties of stimuli (e.g., size and shape). Then, assuming $K_s$ contains a decision rule for $G^2 = \{ r \mid S_0 \in \text{Dom} r, \text{Ran} r \subset G \}$, we see that $\emptyset$ can satisfy $\langle S_0, G \rangle$ while learning an $r \in G^2$.

(4) We can illustrate the "false start" phenomenon with the rule set $K_s = \{ r_1, r_1', \ldots, r_n, r_n', \circ, h, s \}$ where $r_1, r_1', \ldots, r_n$,
Structural Learning: Definitional Foundations 315

The outputs of $\circ$ and $h$ are behaviorally distinct because of the way we have defined rules in $K$. Nonetheless, these outputs (rules) may have equivalent effects on elements in their respective domains with respect to goals which are "sufficiently" indiscriminate.

See Vol. I, Chapter 10 where a fairly encompassing theory of memory is described. This theory extends and elaborates the one sketched here.
generate an adequate retrieval rule, control may shift levels as before. Once the needed input is retrieved, control returns to the goal from which the search was initiated and the process continues.

The above mechanisms can only serve to add more knowledge to M. More critical, they can only serve to make more rules in M active. Mechanisms are also needed to explain how information is deactivated. A reasonable assumption to make in this regard is that information is deactivated according to immediate needs in processing information to achieve set goals. The basic constraint is the fixed finite capacity of A.

It would be presumptuous to propose anything definitive at this stage, but one possibility is that: Goals that are no longer in control are deactivated first; then rules that are no longer needed, and finally simple elements (encoded stimuli and responses). This would help but it would still leave open what would happen, for example, in applying a rule which becomes overloaded. In this case, one might assume that elements are dropped in a predetermined manner (e.g., elements processed last, are dropped first), but there is no clear evidence for this. In fact, the primacy effect would suggest just the reverse. Another alternative would be to add more structure to the rule notion itself so that elements can be "erased" as well as generated. This alternative is developed in Volume I, Chapter 10.

Two final comments seem worth making.

(1) Notice that the distinction between short and long term store (memory) has been surpressed. It is not so much a structural (physiological) difference that is involved as the current state of the memory system--that is, part is active and the rest is not.

(2) If the theory is to provide a basis for prediction, as well as explanation, either assumptions must be made concerning activation and deactivation between trials and or techniques must be developed for manipulating or assessing the current state of activation (i.e., the contents of A) of given individuals.

4. EPILOGUE

The distinction between knowledge divorced from behavior, and knowledge attributable to particular subjects, is basic. The former view of knowledge corresponds directly to the notion of competence as used in generative linguistics. The latter view is more consistent with those of epistemologically oriented psychologists like Piaget where knowledge is directly attributable to subjects. In this case knowledge is assumed to grow according to postulated learning mechanisms.

What the present theory does (among other things) is to provide a missing link between the two views. Knowledge, viewed as competence, corresponds to a theory of the observer, and provides measuring units (rules) against which the knowledge had by individual subjects may be measured. In effect, the theory of knowledge (competence) provides a basis for the operational definition of human knowledge.

In retrospect, these two uses and their relationship to existing theoretical approaches might better have been emphasized by reserving the term "competence" for the former usage and the term "knowledge" for the latter.


Corcoran, J. *Three logical theories.* *Philosophy of Science,* 1969, 36, 153-177.


Corcoran, J. A mathematical model of Aristotle's Syllogistic. Submitted for publication.


Craik, R. C. Directed versus independent discovery of established relations. *Journal of Educational Psychology,* 1956, 47, 223-234.


Descartes, R. *Rules for the direction of the mind*. In: Italdane and Ross (Ed. & Tr.), *Philosophical works of Descartes*. Volume 1. New York: Dover, 1931.


Elkind, D. *Quality conceptions in junior and senior high school students*. *Child Development*, 1961, 32, 551-560. (b)


Hadamard, J. The psychology of invention in the mathematical field. Princeton: Dover, 1945.


Harris, Z. Distributional structures. Word, 1954, 10, 146-162.


Haselrud, G. M., and Meyers, Shirley. The transfer value of given and individually derived principles. Journal of Educational Psychology,
Bibliography 321

1958, 49, 293-298.


Kersh, B. Y. The adequacy of 'meaning' as an explanation for the superiority of learning by independent discovery. Journal of Educational Psychology, 1958, 49, 282-293.


Lovell, K. The developmental approach of Jean Piaget. In Cognitive models and development in mental retardation. Monograph Supplement, American Journal of Mental Deficiency, 1966, 70, 84-89. (b)


Miller, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.


Scandura, J. M. Familiarization and one-trial learning: A positive
Scandura, J. M. A formal representation of mathematical knowledge. Paper read at the American Psychological Association Convention, Washington, D. C., September 1, 1967. (a)
Scandura, J. M. Learning verbal and symbolic statements. Journal of Educational Psychology, 1967, 58, 356-364. (b)
Scandura, J. M. What is a rule? Journal of Educational Psychology, 1972, 63, 179-185. (c)


Sperling, G. The information available in brief visual presentations. Psychological Monographs, 1960, 74, No. 11.


Suppes, P. Psychological foundations of mathematics. Technical Report
No. 80, Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California, 1965.


Tolman, E. C. Is there more than one kind of learning. Psychological Review, 1949, 56, 144-155.


Bibliography


Witz, K. G. Representations of cognitive processes and cognitive structure in children I. *Archives de Psychologie*, 1971, (a)

Witz, K. G. Analysis of frameworks in young children. 1971. (b)

Witz, K. G. Structural changes in 4-5 year olds. Notes for a talk delivered at the Second Annual Interdisciplinary Meeting in Structural Learning, University of Pennsylvania, April, 1971. Notes prepared by D. Goodwin. (c)

Witz, K. G. Relational representations. Mimeographed notes, 1971, Section 3.3. (d)


Additional References


<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ackler, L.</td>
<td>264</td>
</tr>
<tr>
<td>Almy, M.</td>
<td>109</td>
</tr>
<tr>
<td>Anderson, J.</td>
<td>222</td>
</tr>
<tr>
<td>Apostel, L.</td>
<td>109</td>
</tr>
<tr>
<td>Arbib, M.</td>
<td>19, 20, 25, 26, 27, 28, 29, 31, 32, 33, 35, 164, 165, 166</td>
</tr>
<tr>
<td>Aristotle,</td>
<td>180, 196n, 207, 208, 220n</td>
</tr>
<tr>
<td>Atkinson, R.</td>
<td>143</td>
</tr>
<tr>
<td>Austin, G.</td>
<td>178</td>
</tr>
<tr>
<td>Ausubel, D.</td>
<td>3, 119, 275</td>
</tr>
<tr>
<td>Averbach, E.</td>
<td>272</td>
</tr>
<tr>
<td>Barenfeld, M.</td>
<td>135, 137n</td>
</tr>
<tr>
<td>Bartlett, F.</td>
<td>3, 4, 34, 225</td>
</tr>
<tr>
<td>Baylor, G.</td>
<td>135, 143</td>
</tr>
<tr>
<td>Behr, M.</td>
<td>39</td>
</tr>
<tr>
<td>Beilin, H.</td>
<td>107, 109</td>
</tr>
<tr>
<td>Berge, C.</td>
<td>77</td>
</tr>
<tr>
<td>Berlyne, D.</td>
<td>3, 26, 46</td>
</tr>
<tr>
<td>Beth, E.</td>
<td>180, 181n</td>
</tr>
<tr>
<td>Bever, T.</td>
<td>26</td>
</tr>
<tr>
<td>Bloomfield, L.</td>
<td>188</td>
</tr>
<tr>
<td>Blum, M.</td>
<td>158</td>
</tr>
<tr>
<td>Bohm, D.</td>
<td>156</td>
</tr>
<tr>
<td>Boole, G.</td>
<td>195, 196, 208, 222</td>
</tr>
<tr>
<td>Bourne, L.</td>
<td>57, 69, 70, 275</td>
</tr>
<tr>
<td>Bower, G.</td>
<td>142n</td>
</tr>
<tr>
<td>Brahe, T.</td>
<td>162</td>
</tr>
<tr>
<td>Brown, L.</td>
<td>46, 60</td>
</tr>
<tr>
<td>Brown, P.</td>
<td>110</td>
</tr>
<tr>
<td>Bruner, J.</td>
<td>178</td>
</tr>
<tr>
<td>Bundrick, M.</td>
<td>57n</td>
</tr>
<tr>
<td>Butterworth, I.</td>
<td>113</td>
</tr>
<tr>
<td>Cantor, G.</td>
<td>157</td>
</tr>
<tr>
<td>Carr, E.</td>
<td>98, 101</td>
</tr>
<tr>
<td>Cassirer, E.</td>
<td>167</td>
</tr>
<tr>
<td>Cauman, L.</td>
<td>199n, 213n</td>
</tr>
<tr>
<td>Chomsky, N.</td>
<td>11, 16, 17, 22, 25, 28, 185, 189, 191, 192, 193, 200, 222, 225, 226, 228n, 251, 254, 255, 256n, 272, 275, 276, 278, 290, 300</td>
</tr>
<tr>
<td>Church, A.</td>
<td>200, 226n</td>
</tr>
<tr>
<td>Coleman, J.</td>
<td>120</td>
</tr>
<tr>
<td>Copi, I.</td>
<td>198n, 199n</td>
</tr>
<tr>
<td>Corcoran, J.</td>
<td>179, 195, 207</td>
</tr>
<tr>
<td>Coriell, A.</td>
<td>272</td>
</tr>
<tr>
<td>Craig, R.</td>
<td>46</td>
</tr>
<tr>
<td>Craik, K.</td>
<td>135</td>
</tr>
<tr>
<td>Cramer, G.</td>
<td>176</td>
</tr>
<tr>
<td>Davis, J.</td>
<td>57n</td>
</tr>
<tr>
<td>Dean, R.</td>
<td>220n</td>
</tr>
<tr>
<td>De Groot, A.</td>
<td>142</td>
</tr>
<tr>
<td>Descartes, R.</td>
<td>180</td>
</tr>
<tr>
<td>Dienes, Z.</td>
<td>3, 56, 117, 118, 230n</td>
</tr>
<tr>
<td>Dodwell, P.</td>
<td>109</td>
</tr>
<tr>
<td>Dominowski, R.</td>
<td>275</td>
</tr>
<tr>
<td>Domotor, Z.</td>
<td>237</td>
</tr>
<tr>
<td>Duncker, K.</td>
<td>34, 77, 142</td>
</tr>
<tr>
<td>Durnin, J.</td>
<td>6, 53n, 262, 301n, 308</td>
</tr>
<tr>
<td>Easley, J.</td>
<td>123</td>
</tr>
<tr>
<td>Edelstein, L.</td>
<td>223</td>
</tr>
<tr>
<td>Ehrenpreis, W.</td>
<td>301</td>
</tr>
<tr>
<td>Ekstrand, B.</td>
<td>275</td>
</tr>
<tr>
<td>Eldredge, G.</td>
<td>60</td>
</tr>
<tr>
<td>Elkind, D.</td>
<td>112</td>
</tr>
<tr>
<td>Elsasser, W.</td>
<td>164</td>
</tr>
<tr>
<td>Endicott, F.</td>
<td>267</td>
</tr>
<tr>
<td>Engeler, E.</td>
<td>281, 282</td>
</tr>
<tr>
<td>Ernst, G.</td>
<td>143n</td>
</tr>
<tr>
<td>Finney, D.</td>
<td>46</td>
</tr>
<tr>
<td>Fischbein, E.</td>
<td>113</td>
</tr>
<tr>
<td>Flavell, J.</td>
<td>3, 4, 304n</td>
</tr>
<tr>
<td>Fodor, J.</td>
<td>26</td>
</tr>
<tr>
<td>Frank, W.</td>
<td>223</td>
</tr>
<tr>
<td>Furth, H.</td>
<td>304n</td>
</tr>
<tr>
<td>Gabor, D.</td>
<td>166</td>
</tr>
<tr>
<td>Gagné, R.</td>
<td>3, 4, 6, 7, 46, 60, 119, 225, 273, 275</td>
</tr>
<tr>
<td>Galanter, E.</td>
<td>3, 4, 16, 21, 31, 155, 163, 225, 257n, 275</td>
</tr>
<tr>
<td>Galperin, P.</td>
<td>117</td>
</tr>
<tr>
<td>Garrett, M.</td>
<td>26</td>
</tr>
<tr>
<td>Gödel, K.</td>
<td>158, 162, 259n</td>
</tr>
<tr>
<td>Goodnow, J.</td>
<td>178</td>
</tr>
<tr>
<td>Goss, A.</td>
<td>3</td>
</tr>
<tr>
<td>Gray, W.</td>
<td>113</td>
</tr>
<tr>
<td>Greeno, C.</td>
<td>237</td>
</tr>
<tr>
<td>Greeno, J.</td>
<td>5, 39, 40, 42n, 73, 97, 222</td>
</tr>
<tr>
<td>Grzegorczyk, A.</td>
<td>247</td>
</tr>
<tr>
<td>Groen, G.</td>
<td>266</td>
</tr>
</tbody>
</table>
Author Index

Hadamard, J., 183
Halmos, P., 220n
Hammond, A., 205, 223
Harris, Z., 179, 184, 185, 187, 188, 189, 192, 193, 200, 222
Hartley, J., 116
Hartmanis, J., 157
Hasselrud, G., 46
Hatcher, W., 192, 193
Haygood, R., 57, 70
Healey, D., 111
Hebb, D., 156
Herring, I., 223
Heyting, J., 116
Hilbert, D., 179, 181, 196, 209n
Hilgard, E., 4, 142n, 143n
Hockett, C, 186, 188
Hocutt, M., 9, 278
Hull, C, 26, 46
Hunt, E., 247
Inhelder, B., 110, 112, 113, 115
Jaskowski, S., 221
Jensen, R., 57
Jeeves, M., 118, 230
Johnstone, H., 222
Jones, P., 107
Kalicki, J., 213n
Kalish, O., 217n, 222, 233
Katona, G., 150n
Katz, L., 11
Keenan, E., 201n, 223
Keislar, E., 57
Kelley, J., 220n
Kendler, H., 3, 26
Kendler, T., 3, 4, 26
Kepler, J., 162, 204
Kersh, B., 46
King, R., 57n
Kino, A., 181n
Kittle, J., 46
Klahr, D., 150n
Kleene, S., 171
Knifong, D., 304
Koffka, K., 156
Kotovsky, K., 150n
Kreisel, G., 181, 183n, 185, 191
Lamb, S., 93
Langford, C., 200n
Lashley, K., 156
Laurendeau, M., 108, 111
Leblanc, H., 222
Lee, F., 46
Leibnitz, G., 167
Lemmon, E., 211n, 222
LeTourneau, J., 203, 223
Levine, M., 276
Lewis, C., 200n
Lipson, J., 235, 273
Lloyd, D., 98, 101
Longuett-Higgins, H., 166
Lovell, K., 107, 111, 112, 113, 118
Luger, G., 301
Lukasiewicz, J., 222
Lunzer, E., 111, 113
Luther, B., 48n
Lukkien, D., 48n
Lyons, J., 184, 186, 188, 191, 193
Maier, N., 55, 57
Maley, M., 48n
Maltzman, I., 3, 26
Mandler, G., 3, 4
Männzat, I., 113
Marcus, S., 277, 303
Mates, B., 222
Mayer, M., 116, 118
Mays, W., 109
McCulloch, W., 153, 154, 156
Meillet, A., 188
Mendeleev, D., 162
Merryman, C., 101
Meyers, S., 46
Millenson, J., 225
Miller, G., 3, 4, 16, 17, 21, 25, 28, 31, 155, 163, 164, 225, 228n, 251, 254, 256n, 270, 275, 276n, 300
Minsky, M., 155, 159, 160
Montague, R., 217n, 222, 233
Moore, E., 12
Morf, A., 109
Munz, J., 223
Myhill, J., 166
Neisser, U., 156
Nelson, R., 256n, 276
Newell, A., 3, 4, 131, 133, 134, 136, 137n, 138n, 143n, 144n, 275, 276n
Ogilvie, E., 112
Orton, A., 114
Osborn, C., 3, 26
Paige, J., 150
Pampu, I., 113
Papert, S., 160
Parr, R., 113
Parry, W., 199
Pavlick, E. F., 57n
<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pidgeon, D.</td>
<td>117</td>
</tr>
<tr>
<td>Pinard, A.</td>
<td>108, 111</td>
</tr>
<tr>
<td>Pitts, W.</td>
<td>153, 156</td>
</tr>
<tr>
<td>Plato</td>
<td>166</td>
</tr>
<tr>
<td>Poincaré, H.</td>
<td>196</td>
</tr>
<tr>
<td>Polya, G.</td>
<td>3, 4, 183, 227, 265</td>
</tr>
<tr>
<td>Post, E.</td>
<td>171</td>
</tr>
<tr>
<td>Postal, P.</td>
<td>11</td>
</tr>
<tr>
<td>Prawitz, D.</td>
<td>233</td>
</tr>
<tr>
<td>Premack, D.</td>
<td>186</td>
</tr>
<tr>
<td>Pribram, K.</td>
<td>3, 4, 16, 21, 31, 156, 163, 166, 225, 257, 275</td>
</tr>
<tr>
<td>Quine, W.</td>
<td>185, 221n</td>
</tr>
<tr>
<td>Reich, P.</td>
<td>93</td>
</tr>
<tr>
<td>Reid, C.</td>
<td>196n</td>
</tr>
<tr>
<td>Resnik, M.</td>
<td>211n</td>
</tr>
<tr>
<td>Restle, F.</td>
<td>276</td>
</tr>
<tr>
<td>Reynolds, J.</td>
<td>114</td>
</tr>
<tr>
<td>Richardson, J.</td>
<td>39n, 41</td>
</tr>
<tr>
<td>Rogers, H.</td>
<td>226n, 281, 303</td>
</tr>
<tr>
<td>Rose, R.</td>
<td>48n</td>
</tr>
<tr>
<td>Rosenbloom, P.</td>
<td>171, 222, 275n, 276n</td>
</tr>
<tr>
<td>Roughhead, W.</td>
<td>6, 60</td>
</tr>
<tr>
<td>Rowland, A.</td>
<td>111</td>
</tr>
<tr>
<td>Russell, B.</td>
<td>25, 157, 207</td>
</tr>
<tr>
<td>Scott, D.</td>
<td>32n, 213, 275n, 281, 289</td>
</tr>
<tr>
<td>Shantz, C.</td>
<td>111</td>
</tr>
<tr>
<td>Shaw, J.</td>
<td>3, 4, 131, 133, 134, 137n, 275</td>
</tr>
<tr>
<td>Shaw, R.</td>
<td>153, 165</td>
</tr>
<tr>
<td>Shepler, J.</td>
<td>113</td>
</tr>
<tr>
<td>Simon, H.</td>
<td>3, 4, 131, 133, 134, 135, 136, 137n, 138, 143, 144n, 149, 164, 166, 275</td>
</tr>
<tr>
<td>Skinner, B.</td>
<td>11</td>
</tr>
<tr>
<td>Slater, A.</td>
<td>112</td>
</tr>
<tr>
<td>Smith, L.</td>
<td>57</td>
</tr>
<tr>
<td>Smock, D.</td>
<td>111</td>
</tr>
<tr>
<td>Smulyyan, R.</td>
<td>216</td>
</tr>
<tr>
<td>Sperling, G.</td>
<td>272</td>
</tr>
<tr>
<td>Staats, A.</td>
<td>3</td>
</tr>
<tr>
<td>Staats, C.</td>
<td>3</td>
</tr>
<tr>
<td>Steffes, R.</td>
<td>157</td>
</tr>
<tr>
<td>Steffes, L.</td>
<td>110, 113</td>
</tr>
<tr>
<td>Sternberg, S.</td>
<td>312</td>
</tr>
<tr>
<td>Suppes, P.</td>
<td>9, 11, 15n, 19, 20, 21, 23, 25, 29, 31, 32, 35, 220n, 222, 225, 230n, 266, 270, 275</td>
</tr>
<tr>
<td>Szeminska, A.</td>
<td>109, 110, 112, 113, 115</td>
</tr>
<tr>
<td>Taback, S.</td>
<td>115</td>
</tr>
<tr>
<td>Tarski, A.</td>
<td>158, 159, 162</td>
</tr>
<tr>
<td>Thomas, H.</td>
<td>114</td>
</tr>
<tr>
<td>Thomason, R.</td>
<td>209n, 222</td>
</tr>
<tr>
<td>Tolman, E.</td>
<td>4, 156</td>
</tr>
<tr>
<td>Toy, C.</td>
<td>264</td>
</tr>
<tr>
<td>Turing, A.</td>
<td>157, 166, 282n</td>
</tr>
<tr>
<td>Underwood, B.</td>
<td>39n, 41</td>
</tr>
<tr>
<td>Uzgiris, I.</td>
<td>112</td>
</tr>
<tr>
<td>Van Engen, H.</td>
<td>110</td>
</tr>
<tr>
<td>Von Neuman, J.</td>
<td>153, 154, 156, 157, 158, 159, 161, 162, 165, 166, 167</td>
</tr>
<tr>
<td>Voorhies, D.</td>
<td>270</td>
</tr>
<tr>
<td>Wesley, R.</td>
<td>181n</td>
</tr>
<tr>
<td>Wallace, J.</td>
<td>150n</td>
</tr>
<tr>
<td>Weaver, G.</td>
<td>223</td>
</tr>
<tr>
<td>Wells, J.</td>
<td>230</td>
</tr>
<tr>
<td>Wertheimer, M.</td>
<td>34, 77</td>
</tr>
<tr>
<td>Westlake, P.</td>
<td>166</td>
</tr>
<tr>
<td>Wigner, E.</td>
<td>164, 165, 166, 167</td>
</tr>
<tr>
<td>Williams, D.</td>
<td>147</td>
</tr>
<tr>
<td>Williams, T.</td>
<td>147</td>
</tr>
<tr>
<td>Wellington, C.</td>
<td>110</td>
</tr>
<tr>
<td>Wittrock, M.</td>
<td>46, 55, 57, 65, 69, 70</td>
</tr>
<tr>
<td>Witz, K.</td>
<td>123, 126n, 127, 304n</td>
</tr>
<tr>
<td>Wohlwill, J.</td>
<td>254</td>
</tr>
<tr>
<td>Woods, P.</td>
<td>116</td>
</tr>
<tr>
<td>Woodward, E.</td>
<td>46</td>
</tr>
</tbody>
</table>
SUBJECT INDEX

Activity, cycle of, 123
Activity, element of, 125
Activity structure, 125
Adequacy continuity principle, 239
Alphabet, 171, 279
  English, 184
Assessment episode, 308
Array, 279
Aptitude test taker, 147
Association, 6
Automaton, 12
  connected, 21
  finite, 21

Behavior, 280
  meaningful, 4
Behavior class, 227
Behavior potential, 260

Category constraint, 124
Circuit, 81
Classical conditioning, 14
Complexity measure, 242
Complexity of a structure, 160
Complexity type, 157
Compound, 185
Computation, 306
Concept, 6, 128
  geometrical, 110
  limit, 115
  number, 109
  point, 115
  volume, 112
  spatial, 110
  conjunction of, 242
  research of, 73
  super position of, 241
Concept algebra C, 238
Criterion referenced test
  (assessment program), 308
Curriculum, 255

D, 0, R, 8, 51
Descendence relation, 87
Decision-making capability (atomic dmc), 283
Deductive reasoning, 181
Difficulty measure, 242
Direct majorant, 87
Direct maximum, 88

Direct minimum, 88
Direct minorant, 88
Discovery, 60

Education process, 246
Element (E), 278
Embodiment, 229
Empiricism, level of, 253
Existence principle, 239
Exposition, 60

Finitary conjecture, 303
Function, 7, 114

General game playing program, 147
General problem solver, 142
Global property, 159-161
Goal, 259, 305
  reachable, 312
  satisfied, 308
  higher-order, 263
  second order, 311
Goal situation, 305
  resolved, 308
Grammar, 189
  generative, 299
Graph, 76, 77 def.
  equivalence, 88
  strongly connected, 88
  arcs of, 77 def.

Infer, 182
Information processing capacity, 270
Information processing system,
  human, 136, 137
Innate basis, 303
  finitary, 303
Input, 279
Instance, stimulus-response (S-R),
  278, 281
Instruction, 289
Instruction space D, 238

Kernel, 185, 209
Knowledge, 180
  constituent, 180
  potential, 294, 303
  "propositional", 180
  theory of, 255, 276, 294
  -atomic, 296
-finitary, 296
-simple, 296

Knowledge potential conjecture, 303

Language
regular, 16
super iterative, 192
levels of, 184

Learning, 266
Learning episode, 308
Learning subject, 238
Lexicon, 184

Local property, 159-161
Logic theorist, 133
Logical consequence, 182
Logical mathematical experience
108
Loop, 81

Mathematical golf, 172
Mathematical model, 238
Mathematics, 107
Means-ends analysis, 142
Memory, 315
theory of, 268
Modeling, 239, 247
Motivation theory, 267

Operator, 141
Output, 279

Path, 78
composite, 78
equivalent, 307
simple, 78, 306
length of, 78

Principle, 51

Problem, 84 def.
solvable, 84

Problem algebra P, 238
Problem solving, 4, 7
semantics of, 243

Problem space, 136
Production, 172

Program, 289

Progressive deepening search, 142

Proof, 181, 195, 212, 216
finished, 216
linear, 208
mathematical, 115, 181
partial, 216
subsidiary, 212
suppositional, 215
theory of, 199

Relevance class, 313

Representation principle, 239

Representation Theorem for Finite
Automata, 16

Response
atomic, 279
effective, 7
Rule, 6, 51, 228, 283
assumption, 212
atomic, 260, 284
branching, 260
closing, 216
composition, 291
-inverse of, 291
conditionalization, 214
conjunction, 292
decoding (atomic d), 286
definitional, 219
derivation, 60
definition, 219
encoding (perceptual) (atomic p), 286
existential instantiation, 218
finitary, 288
generalization, 293
-simple, 292
global immediate, 217
global subsidiary, 217
higher order, 32, 231, 257, 290
immediate inference, 212
initial string, 209
inference, 196
inverse, 231
logical axiom, 209
lower order, 256
nth order generable, 296
operating (atomic or), 282
partially equivalent, 308
reductio, 213
relational, 230, 256
relevant, 312
restricted, 54
restriction, 293
retrieval, 315
selection, 294
simple, 290
storing, 315

subsidiary proof, 214
supposition, 216
suppositional inference, 233
two case, 214
description of, 281
denotation of, 7
domain of, 282, 287
description of, 281, 288
generality of, 46
intension of, 281
range of, 282, 287
reification, 288
semantics of, 281

S-B pair (instance), 280
Subject Index 337

S:R, 278
S-R instance (see Instance)
Scan-search procedure, 143
Set Function Language (SFL), 4
Solution space, 238
Stimulus
  atomic, 279
  effective, 7
Stimulus situation, 279
  functionally distinct, 280
Structural learning, 252
  theory of, 251, 276
  idealized (memory free) theory of, 304, 306
Subarray, 279
Subbehavior, 280
Subgraph, 81 def.
Subproof, 216
Subrule, 284
  closed, 284
Substimulus, 280
System, 229
  mathematical, 171
  production, 148
  knowledge of, 229
Task environment, 136
Teacher, 238
Testing principle, 240
Theory
  axiomatic, 232
  deterministic, 252
  partial, 253
Thinking, 108
Tote, 17, 22
Training program, 308
Trial, 308
  stage of, 311
Uniqueness principle, 239