

R. DUNCAN LUCE
ROBERT R. BUSH
EUGENE GALANTER
EDITORS

HANDBOOK OF
MATHEMATICAL
PSYCHOLOGY

VOLUME II
CHAPTERS 9-14

R. DUNCAN LUCE received both his B.S. and his Ph.D. from MIT. At present, he is Professor of Psychology at the University of Pennsylvania, a position he has held since 1959. His publications include *Games and Decisions* (with Howard Raiffa, Wiley, 1957), *Individual Choice Behavior* (Wiley, 1959), and the editorship of *Developments in Mathematical Psychology*.

ROBERT R. BUSH graduated from Michigan State and received his Ph.D. from Princeton. He has been Professor and Chairman of the Department of Psychology at the University of Pennsylvania since 1958. His previous publications include *Stochastic Models for Learning* (with F. Mosteller, Wiley, 1955) and the editorship (with W. K. Estes) of *Studies in Mathematical Learning Theory*.

EUGENE GALANTER is a graduate of Swarthmore and received his Ph.D. from the University of Pennsylvania. In 1962, he was appointed Professor and Chairman of the Department of Psychology at the University of Washington. His publications include the editorship of *Automatic Teaching* (Wiley, 1959) and *Plans and the Organization of Behavior* (with G. A. Miller and K. H. Pribram).



66-12967
152.83 L93h v. 2 \$11.95
Luce, Robert Duncan, ed.

66-12967
152.83 L93h v. 2 \$11.95
Luce, Robert Duncan, ed.

Handbook of mathematical psy-
chology. Wiley [1963-65]

MAIN

kansas city



public library

kansas city, missouri

Books will be issued only
on presentation of library card.

Please report lost cards and
change of residence promptly.

Card holders are responsible for
all books, records, films, pictures
or other library materials
checked out on their cards.

Handbook of Mathematical Psychology

Volume II, Chapters 9-14

WITH CONTRIBUTIONS BY

Saul Sternberg Noam Chomsky
Richard C. Atkinson George A. Miller
William K. Estes Anatol Rapoport

EDITED BY

R. Duncan Luce, *University of Pennsylvania*
Robert R. Bush, *University of Pennsylvania*
Eugene Galanter, *University of Washington*

New York and London

John Wiley and Sons, Inc.

Copyright © 1963 by John Wiley & Sons, Inc.

All Rights Reserved

This book or any part thereof
must not be reproduced in any form
without the written permission of the publisher.

Library of Congress Catalog Card Number: 63-9428
Printed in the United States of America

Preface

A general statement about the background, purposes, assumptions, and scope of the *Handbook of Mathematical Psychology* can be found in the Preface to Volume I. Those observations need not be repeated; indeed, nothing need be added except to express our appreciation to Mrs. Judith White who managed the administrative details while the chapters of Volume II were being written, to Mrs. Sally Kraska who assumed these responsibilities during the production stage, to Miss Ada Katz for assistance in typing, and to Mrs. Kay Estes who ably and conscientiously prepared the indices. As we said in the Preface to Volume I, "Although editing of this sort is mostly done in spare moments, the cumulative amount of work over three years is really quite staggering and credit is due the agencies that have directly and indirectly supported it, in our case the Universities of Pennsylvania and Washington, the National Science Foundation, and the Office of Naval Research."

Philadelphia, Pennsylvania
June 1963

R. DUNCAN LUCE
ROBERT R. BUSH
EUGENE GALANTER

KANSAS CITY (MO.) PUBLIC LIBRARY

11.95

6612967

Contents

| | | |
|-----|--|-----|
| 9. | STOCHASTIC LEARNING THEORY | 1 |
| | by Saul Sternberg, <i>University of Pennsylvania</i> | |
| 10. | STIMULUS SAMPLING THEORY | 121 |
| | by Richard C. Atkinson, <i>Stanford University</i> and William K. Estes, <i>Stanford University</i> | |
| 11. | INTRODUCTION TO THE FORMAL ANALYSIS OF NATURAL LANGUAGES | 269 |
| | by Noam Chomsky, <i>Massachusetts Institute of Technology</i> and George A. Miller, <i>Harvard University</i> | |
| 12. | FORMAL PROPERTIES OF GRAMMARS | 323 |
| | by Noam Chomsky, <i>Massachusetts Institute of Technology</i> | |
| 13. | FINITARY MODELS OF LANGUAGE USERS | 419 |
| | by George A. Miller, <i>Harvard University</i> and Noam Chomsky, <i>Massachusetts Institute of Technology</i> | |
| 14. | MATHEMATICAL MODELS OF SOCIAL INTERACTION | 493 |
| | by Anatol Rapoport, <i>University of Michigan</i> | |
| | INDEX | 581 |

9

*Stochastic Learning Theory*¹

Saul Sternberg

University of Pennsylvania

1. Preparation of this chapter was supported by Grant G-18630 from the National Science Foundation to the University of Pennsylvania. Doris Aaronson provided valuable help with computations; her work was supported in part by NSF Grant G-14839. I wish to thank Francis W. Irwin for his helpful criticism of the manuscript.

Contents

| | |
|---|----|
| 1. Analysis of Experiments and Model Identification | 6 |
| 1.1. Equivalent events, | 7 |
| 1.2. Response symmetry and complementary events, | 9 |
| 1.3. Outcome symmetry, | 11 |
| 1.4. The control of model events, | 12 |
| 1.5. Contingent experiments and contingent events, | 14 |
| 2. Axiomatics and Heuristics of Model Construction | 15 |
| 2.1. Path-independent event effects, | 16 |
| 2.2. Commutative events, | 17 |
| 2.3. Repeated occurrence of a single event, | 18 |
| 2.4. Combining-classes condition: Bush and Mosteller's linear-operator models, | 19 |
| 2.5. Independence from irrelevant alternatives: Luce's beta response-strength model, | 25 |
| 2.6. Urn schemes and explicit forms, | 30 |
| 2.7. Event effects and their invariance, | 36 |
| 2.8. Simplicity, | 38 |
| 3. Deterministic and Continuous Approximations | 39 |
| 3.1. Approximations for an urn model, | 40 |
| 3.2. More on the expected-operator approximation, | 43 |
| 3.3. Deterministic approximations for a model of operant conditioning, | 47 |
| 4. Classification and Theoretical Comparison of Models | 49 |
| 4.1. Comparison by transformation of the explicit formula, | 50 |
| 4.2. Note on the classification of operators and recursive formulas, | 56 |

| | |
|--|-----|
| 4.3. Implications of commutativity for responsiveness and asymptotic behavior, | 56 |
| 4.4. Commutativity and the asymptote in prediction experiments, | 61 |
| 4.5. Analysis of the explicit formula, | 65 |
| | |
| 5. Mathematical Methods for the Analysis of Models | 75 |
| 5.1. The Monte Carlo method, | 76 |
| 5.2. Indicator random variables, | 77 |
| 5.3. Conditional expectations, | 78 |
| 5.4. Conditional expectations and the development of functional equations, | 81 |
| 5.5. Difference equations, | 83 |
| 5.6. Solution of functional equations, | 85 |
| | |
| 6. Some Aspects of the Application and Testing of Learning Models | 89 |
| 6.1. Model properties: a model type as a subspace, | 89 |
| 6.2. The estimation problem, | 93 |
| 6.3. Individual differences, | 99 |
| 6.4. Testing a single model type, | 102 |
| 6.5. Comparative testing of models, | 104 |
| 6.6. Models as baselines and aids to inference, | 106 |
| 6.7. Testing model assumptions in isolation, | 109 |
| | |
| 7. Conclusion | 116 |
| | |
| References | 117 |

Stochastic Learning Theory

The process of learning in an animal or a human being can often be analyzed into a series of choices among several alternative responses. Even in simple repetitive experiments performed under highly controlled conditions, the choice sequences are typically erratic, suggesting that probabilities govern the selection of responses. It is thus useful to think of the systematic changes in a choice sequence as reflecting trial-to-trial changes in response probabilities. From this point of view, much of the study of learning is concerned with describing the trial-to-trial probability changes that characterize a stochastic process.

In recent mathematical studies of learning investigators have assumed that there is *some* stochastic process to which the behavior in a simple learning experiment conforms. This is not altogether a new idea (for a sketch of its history, see Bush, 1960b). But two important features appear primarily in the work since 1950 that was initiated by Bush, Estes, and Mosteller. First, the step-by-step nature of the learning process has been an explicit feature of the proposed models. Second, these models have been analyzed and applied in ways that do not camouflage their statistical aspect. Various models have been proposed and studied as possible approximations to the stochastic processes of learning. The purpose of this chapter is to review some of the methods and problems of formulating models, analyzing their properties, and applying them in the analysis of learning data.²

The focus of our attention is a simple type of learning experiment. Each of a sequence of trials consists of the selection of a response alternative by the subject followed by an outcome provided by the experimenter. The response alternative may be pressing one of a set of buttons, turning right in a maze, jumping over a barrier before a shock is delivered, or failing to recall a word.³

² The model enterprise is not and should not be separate from other efforts in the study of learning. It is partly for this reason that I have not attempted a summary of present knowledge vis-à-vis various models. A good proportion of the entire learning literature is directly relevant to many of the questions raised in work with stochastic models. For this reason an adequate survey would be gargantuan and soon outdated.

³ The reader will note from these examples that the terms "choice" and "response alternative" are used in the abstract sense discussed in Chapter 2 of Volume I. For example, I ignore the question whether a choice represents a conscious decision; classes of responses are defined both spatially (as in the maze) and temporally (as in the shuttlebox); a subject's inability to recall a word is grouped with his "choice" not to say it.

We shall be concerned almost entirely with experiments in which the subject's behavior is partitioned into two mutually exclusive and exhaustive response alternatives. The outcome may be a pellet of food, a shock, or the onset of one of several lights. The outcome may or may not change from trial to trial and its occurrence may or may not depend on the response chosen. When no differential outcome (Irwin, 1961) is provided, as, for example, when the experimenter gives food on every trial or when he does not explicitly provide any reward or punishment, we think of the experiment simply as a sequence of response choices. We do not consider experiments in which the stimulus situation is deliberately altered from trial to trial; for our purposes it can be referred to once and then ignored. Because little mathematical work has been done on bar-pressing or runway experiments, they receive little attention.

The elements of a stochastic learning model correspond to the components of the experiment. The sequence of trials is indexed by $n = 1, 2, \dots, N$. There is a set of *response alternatives*, $\{A_1, \dots, A_j, \dots, A_r\}$, and a set of *outcomes*, $\{O_1, O_2, \dots, O_s\}$. Each response-outcome pair constitutes a possible *experimental trial event*, E_n . A probability distribution, $\{p_{i,n}(1), \dots, p_{i,n}(j), \dots, p_{i,n}(r)\}$, is defined over the set of response alternatives for each subject, $i = 1, 2, \dots, I$, and each trial, $n = 1, 2, \dots, N$. The subject subscript is suppressed when we consider a generic sequence. The response probabilities form a *probability vector* with r elements. In most of the examples that have been studied $r = 2$ and $p_n(1) = 1 - p_n(2)$, thus making it possible to reduce the sequence of probability vectors to a sequence of scalars, p_1, \dots, p_n, \dots .

The crux of a model is its description of response-probability changes from trial to trial. One type of description is in terms of explicit transition rules or *operators*, usually independent of the trial number, that transform the response probabilities of trial n into those of trial $n + 1$. The operator invoked depends on the event that occurs on trial n . A second type of description is in terms of an explicit formula for the dependence of p_n on both n and the sequence of events through trial $n - 1$. The explicit formula approach, although less popular, is somewhat more general, as we shall see, because it can be used for models whose expression in terms of operators of the type mentioned would be cumbersome or impossible.

One or more parameters with unspecified numerical values usually appear in the formulation of a model. These parameters may be initial probability values or they may be quantities that reflect the magnitude of the effects of different trial events. The values of these parameters are usually estimated from the set of data being analyzed. In more stringent tests of a model parameters estimated from one phase of an experiment are used in its application to another phase. A useful, if rough, distinction can be made

between a *model type*, consisting of the class of models with all possible values of the parameters, and a *model* in which the numerical values of parameters are specified. Because there is no theory of parameters in this field, investigators have concentrated on determining which model type, if any, can describe a set of data, rather than which model within a type is the appropriate one. Model types themselves can be grouped into different families which reflect basically different conceptions of the learning process, and the choice between families is the fundamental problem. In this chapter, as in much of the published work, the word "model" is used to denote a model type when it is clear from the context what is meant.

The model builder's view of learning differs in its emphasis from that of many experimenters. The idea of trial-to-trial changes in the behavior of individual subjects has been basic in traditional approaches to learning. But, with few exceptions, the changes have been investigated indirectly, often by considering the effects of experimental variations on the gross shape of a curve of mean performance versus trials. Stochastic models have been used increasingly to supplement an interest in the learning curve with analyses of various sequential features of the data, features that reflect more directly the operation of the underlying trial-to-trial changes.

At first glance stochastic models appear to have been remarkably successful in accounting for the data from a variety of learning experiments (Bush & Mosteller, 1955; Bush & Estes, 1959). Recent work, however, suggests that we view the situation with caution. As more model types are investigated, the problem becomes not one of fitting a model to a set of data but of discrediting all but one of the competing models. Apparent agreement between model and data comes easily and can lead to a sense of over-confidence. There are a number of ways of dealing with this problem, such as refining estimation and testing procedures and performing crucial experiments. Criteria have been invoked that involve more than merely the ability of a model to describe a particular set of data. A model is designed to describe some process that occurs in an experiment. But in the present state of the art it is difficult to perform experiments in which no other processes, aside from those described by the model, intrude. We must therefore compromise between making severer demands on the models and acknowledging that at present their descriptions of actual experiments cannot hope to be more than approximations.

1. ANALYSIS OF EXPERIMENTS AND MODEL IDENTIFICATION

In considering what may happen on a trial, we must draw a careful distinction between *experimental events* and *model events*. Not all the events

in an experiment are identified with distinct events in the model that is applied to it. A good deal of intuition, with varying degrees of experimental support, leads to assumptions of equivalence and complementarity among experimental events.⁴ These assumptions have strong substantive implications. They also fix the number of distinct operators (events) in the model, impose constraints on them, and specify how their application to response probabilities is governed. In general, the more radical the assumptions, the simpler the model, the less the labor in analysis and application, and the greater the chance that the model will fail. A few examples will serve to illustrate some of the relevant considerations.

1.1 Equivalent Events

Each response-outcome pair constitutes an experimental event. Examples of several sets of experimental events are given in Table 1.⁵

At this level of analysis each experiment has four possible events per trial. In one analysis of the prediction experiment (e.g., Estes & Straughan, 1954) the four experimental events are grouped into two equivalence classes, $\{(A_1, O_1), (A_2, O_1)\}$ and $\{(A_1, O_2), (A_2, O_2)\}$, which define the two model events, E_1 and E_2 . This amounts to assuming that changes in

⁴ Throughout this chapter adjectives such as "equivalent," "complementary," "path-independent," and "commutative" are applied to the term "events." In all cases this is a shorthand way of describing properties of the effects on response probabilities of the occurrence of events. These properties may characterize model events. Whether they also characterize corresponding experimental events is a question to be answered by testing the model.

Occasionally I write as if an event were an active agent, as in "the event transforms the probability . . ." This is another shorthand form. It stands for "the operator corresponding to the event transforms . . ." in the case of model events with operator representations. It stands for "the occurrence of the event affects the organism so as to change its probability . . ." in the case of experimental events.

The use of the same terms in talking about both kinds of events is intended to emphasize the fact that insofar as a model is successful the properties of its events are also properties of the corresponding experimental events.

⁵ Choices have to be made even at the stage of tabulating response and outcome possibilities, as illustrated by the difference between the tables for T-maze and prediction experiments. An alternative analysis of the T-maze experiment, formally identical to the one for the prediction experiment, is illustrated in Table 2. One relevant consideration is the type of experiment to which the analysis may be generalized. Thus, if in the prediction experiment we defined the outcomes to be O_1 :correct, O_2 :incorrect, then the generalization to an experiment with three buttons and three lights might be inappropriate. For the T-maze experiment the analysis in Table 1 is to be preferred if we wish to include experiments in which on some trials neither or both maze arms are baited. On the other hand, Table 2 provides an analysis that is more easily extended to experiments with a correction procedure.

Table 1 Definition of Experimental Events in Four Experiments

(i) Two-Choice Prediction

| Response | Outcome |
|----------------------------|---------------------------------------|
| A_1 : Left button press | O_1 : Left light onset (correct) |
| A_2 : Right button press | O_1 : Left light onset (incorrect) |
| A_1 : Left button press | O_2 : Right light onset (incorrect) |
| A_2 : Right button press | O_2 : Right light onset (correct) |

(ii) T-Maze

| Response | Outcome |
|--------------------|-----------------|
| A_1 : Left turn | O_1 : Food |
| A_2 : Right turn | O_2 : No Food |
| A_1 : Left turn | O_2 : No Food |
| A_2 : Right turn | O_1 : Food |

(iii) Escape-Avoidance Shuttlebox

| Response | Outcome |
|--|-----------------------------------|
| A_1 : Jump before US from left to right | O_1 : Avoidance of US on left |
| A_1' : Jump before US from right to left | O_1' : Avoidance of US on right |
| A_2 : Jump after US from left to right | O_2 : Escape of US on left |
| A_2' : Jump after US from right to left | O_2' : Escape of US on right |

(iv) Continuous Reinforcement in Runway

| Response | Outcome |
|---|--------------|
| A_1 : Run with speed in first quartile | O_1 : Food |
| A_2 : Run with speed in second quartile | O_1 : Food |
| A_3 : Run with speed in third quartile | O_1 : Food |
| A_4 : Run with speed in fourth quartile | O_1 : Food |

response probability from trial to trial depend only on outcomes and not on responses. Reward of A_1 by O_1 is assumed equivalent to nonreward of A_2 by O_1 . This assumption, as we shall see, considerably simplifies models for the experiment. Although a comparable reduction in the number of events in the T-maze (or the analogous "two-armed bandit") experiment is possible, it has, in general, not been made (e.g., Galanter & Bush, 1959). Analysis of the shuttlebox experiment has ignored the alternation in the animal's starting position and used the equivalence classes $\{(A_1, O_1), (A_1', O_1')\}$ and $\{(A_2, O_2), (A_2', O_2')\}$ (Bush & Mosteller, 1955, 1959). Analyses of the runway experiment have grouped all experimental events into one equivalence class, resulting in a single model event (e.g., Bush & Mosteller, 1955). As an alternative, at least one theory about runway behavior proposes that the effect of a trial event depends critically on the running speed of that trial (Logan, 1960).

It should be emphasized that the reduction in the number of events by the definition of equivalence classes entails strong assumptions about what the experimental subjects are indifferent to. Even in forming the lists in Table 1 we have implicitly appealed to the existence of equivalence classes; we have assumed, for example, that all ways of turning left are equivalent.

1.2 Response Symmetry and Complementary Events

When we have determined the set of events $\{E_k\}$ for a model by defining whatever equivalence classes seem reasonable, we can introduce further simplifications by identifying pairs or sets of complementary events. In a two-choice experiment two events, E_1 and E_2 , form a complementary pair if, to put it roughly, the effect of E_1 on p is the same as the effect of E_2 on $q = 1 - p$. If the model involves a set of operators, each associated with an event, then E_1 and E_2 will be associated with *complementary operators*.

Let us suppose, for example, that the operators are linear and that E_k transforms p into $Q_k p = \alpha_k p + a_k$, where α_k and a_k are constants. Then the complementarity of E_1 and E_2 requires that when E_2 occurs $q = 1 - p$ will be transformed into $\alpha_1 q + a_1$. This requirement implies that p is transformed into $\alpha_1 p + (1 - \alpha_1 - a_1)$ when E_2 occurs and gives the relations $\alpha_2 = \alpha_1$ and $a_2 = 1 - \alpha_1 - a_1$. The result is that we have one operator and its complement rather than two independent operators.

As in the case of equivalence classes of events, it is the subject's behavior, not the experimenter, that determines whether two events are complementary. In the analysis of prediction experiments it has frequently been assumed that the two equivalence classes are complementary. In analysis

of the T-maze experiment the event pairs $\{(A_1, O_1), (A_2, O_1)\}$ and $\{(A_1, O_2), (A_2, O_2)\}$ have been assumed to be complementary. In their treatment of an experiment on imitation Bush and Mosteller (1955) rejected the assumption that rewarding an imitative response was complementary to rewarding a nonimitative response.

It is in dealing with pairs of events in which the same outcome (a food reward, for example) occurs in conjunction with a pair of "symmetric" responses (left turn and right turn, for example) that investigators have been most inclined to assume that the events are complementary. There appears, however, to be no available response theory that would allow us to determine, from the properties of two (or more) responses, whether they are symmetric in the desired sense. Learning model analyses of a variety of experiments would provide one source of information on which such a theory could be based.

At present, therefore, it is primarily intuition that leads us to assume that left and right are symmetric in a sense in which imitation and nonimitation are not. Perhaps more obvious examples of asymmetric responses are alternatives A_1 and A_2 in the shuttlebox experiment and alternatives A_1 and A_4 in the runway (see Table 1).

In the foregoing discussion I have considered the relation between the events determined by two responses for each of which the same outcome is provided, such as "left turn—food" and "right turn—food." A second sense in which response symmetry may be invoked in the design of learning model operators arises when we consider the effects of the same event (such as "left turn—food") on the probabilities of two different responses. In many models the operators that represent the effects of an event are of the same form for all responses; that is, the operators are members of a restricted family, such as multiplicative or linear transformations. These models are, therefore, invariant under a reassignment of labels to responses, so long as the values of one or two parameters are altered. Such invariance represents a second type of response symmetry.

This type of symmetry may be defined only in relation to a specified family of operators; when such symmetry obtains, then the family is *complete* (Luce, 1963) in the sense that it contains the operators appropriate to all the responses.

As an example, let us consider the Bush-Mosteller model for two responses, in which $p = \Pr \{A_1\}$ and the occurrence of E_k transforms p into $\alpha_k p + a_k$. This is, of course, equivalent to the transformation of $q = 1 - p$ into $\alpha'_k q + a'_k$, where $\alpha'_k = \alpha_k$ and $a'_k = 1 - a_k - \alpha_k$. As a result of the occurrence of E_k , the probabilities of the two responses change in the same (linear) manner. The model for changes in $\Pr \{A_1\}$ is of the same form as the model for changes in $\Pr \{A_2\}$.

Not all learning models are characterized by this sort of response-symmetry relative to a simple family of operators. For example, Hull's model (1943, Chapter 18) for changes in the probability of a response, A_1 , where the alternative response, A_2 , is defined as nonoccurrence of A_1 , incorporates a threshold notion in the relationship between the "strength" of A_1 (its ${}_S E_R$) and its probability. No such threshold applies to A_2 , and the responses are not symmetric in the sense outlined. (This model is discussed in Secs. 2.5 and 4.1.)

A second model that lacks the symmetry features is Bush and Mosteller's (1959) "late Thurstone model" discussed in Sec. 2.6. In this model the transformations induced by events on the probability of error can be expressed by applying an additive increment to the reciprocal of the probability:

$$\frac{1}{p_{n+1}} = \frac{1}{p_n} + b.$$

The form of the corresponding transformation on the reciprocal of $q_n = 1 - p_n$ is not a member of the family of additive (or even full linear) transformations.

For any two-response model in which the effects of events may be expressed in terms of operators, we can use the response-symmetry condition to impose a restriction on the class of allowed operators. We can do this by translating the condition into the requirement that the operators on $p = \Pr \{A_1\}$ and $q = 1 - p = \Pr \{A_2\}$ that correspond to an event are to be members of the same family of operators. If, for example, we require the family to be expressed by a particular function with two parameters,

$$p_{n+1} = f(p_n; a, b),$$

then symmetry dictates that for all p , $0 \leq p \leq 1$, and for all allowed values of a and b , f has the property that

$$f(p; a, b) + f(1 - p; c, d) = 1,$$

where $c = c(a, b)$ and $d = d(a, b)$. As indicated by the discussion of the "late Thurstone model," when we require that $f(p)$ be of the form $f(p) = [(a/p) + b]^{-1}$, then its operators do not satisfy the condition. The condition may be generalized in an obvious way to more than two responses.

1.3 Outcome Symmetry

The reader may have questioned the contrast between the treatments of the prediction and the T-maze experiments. In the first experimental

events containing different responses are grouped in the same equivalence class, whereas this is not done in the second. The critical difference that guides the definition of equivalence classes here appears to be the extent of outcome symmetry from what is thought to be the subject's viewpoint. From the experimenter's viewpoint the possible outcomes in many T-maze experiments could be symmetrically described by "left-arm baited" and "right-arm baited." With this terminology we can form a new event list that is identical, formally, to the list for the prediction experiment.

Table 2 Alternative Definition of Experimental Events in T-Maze Experiment

| Response | Outcome |
|--------------------|--------------------------------------|
| A_1 : Left turn | O_1 : Left-arm baited (correct) |
| A_2 : Right turn | O_1 : Left-arm baited (incorrect) |
| A_1 : Left turn | O_2 : Right-arm baited (incorrect) |
| A_2 : Right turn | O_2 : Right-arm baited (correct) |

Despite their formal identification with the events in the prediction experiment, many model builders would be loth to assume that the first two events are equivalent in the T-maze experiment. The distinction between the two experiments is more easily seen if they are extended to three alternatives. Food presented after one of three alternatives is unlikely to have the same effect as the absence of food after another of the three. It is perhaps conceivable that the onset of one of three lights would have the same effect independently of the response that precedes it. The important criterion seems to be not whether the outcomes are capable of symmetric description by the experimenter but whether they "appear" symmetric to the subject. The question whether outcomes are symmetric is, of course, finally decided by whether the behavior produced by subjects is described by a model in which symmetry is assumed.

1.4 The Control of Model Events

Experimental events, with assumptions about their equivalence and complementarity, determine a set of model events and thereby give rise to four important classes of models. These classes are defined in terms of how the occurrence of model events is controlled by the sequence of response-outcome pairs in the experiment.

If knowledge of both response and outcome is needed in order to know which model event has occurred on a trial, then the events are *experimenter-subject controlled*. For example, in the analysis of the T-maze experiment given in Table 1 there are four model events, and both the direction of the rat's turn and the schedule of rewards must be known in order to determine which event has occurred. Although the reward schedule may be predetermined, the response is not. Therefore the sequence of probability changes cannot be specified in advance of the experiment. This class of models is relatively intractable mathematically.

The second class of models is illustrated by the Bush and Mosteller (1955) analysis of the shuttlebox experiment. Here there are only two model events (avoid or escape) and the response determines the outcome (no-shock or shock). This is an example of *subject-controlled events* in which the response alone determines the model event. Any experiment in which responses and outcomes are perfectly correlated consists of subject-controlled events. This correlation is produced in the T-maze by baiting the left arm on every trial and never baiting the right arm, for example. Again, the sequence of probability changes cannot be specified in advance and in general will be different for each subject in the experiment. Even if the real subjects correspond to a set of *identical model subjects* (identical in their parameter values and initial response probabilities), they will have a *distribution* of response probabilities on later trials.

The third class of models is illustrated by the Estes and Straughan (1954) analysis of the prediction experiment. This is an example of *experimenter-controlled events* in which the outcome alone (left-light or right-light onset) determines the event. Because the outcome schedule can be predetermined, only the parameter values and initial probabilities are needed in order to specify the trial-to-trial sequence of response probabilities. Identical subjects with identical outcome sequences who behave in accordance with such a model will have the same sequence of response probabilities. Although a subject's successive response probabilities may be different, his successive responses are independent. These features of models with experimenter-controlled events make mathematical analysis relatively simple. Despite the wide use of these models, however, direct experimental evidence that favors the independence assumption has not been forthcoming, and for the prediction experiment there is a certain amount of strong negative evidence, for example, in Hanania (1959) and Nicks (1959). The onset of a light apparently has an effect on the response probability that depends on whether the onset was correctly predicted. It is not known whether there are other experiments for which models with experimenter-controlled events might be appropriate.

Models in the fourth class involve just a single event and are the simplest.

A *single-event model* may be obtained from any model with subject-controlled events by the simplification of grouping all events into a single equivalence class. If the events "left turn—reward" and "right turn—non-reward" in a T-maze experiment with 100:0 reward, for example, are assumed to have equal effects on the probability of the right-turn response, then a single-event model is applicable. A second source for a single-event model is in the application of a model with experimenter-controlled events to the prediction experiment (Sec. 1.1), under the special condition that the same outcome is provided on every trial. Models with a single event are the easiest to study mathematically (see, for example, Bush & Sternberg, 1959) but the assumptions they entail seem seldom to be met in practice (Galanter & Bush, 1959; Sternberg, 1959b).

It appears that the best understood models are poor approximations to the data, and the models more likely to apply are little understood. We probably cannot dispense with subject control of events in learning experiments, and therefore the only choice available to us is whether or not there is experimenter control as well. Insofar as we deal with experiments whose outcomes are perfectly correlated with responses, we simplify matters by eliminating experimenter control. In this chapter I shall consider, in particular, models with subject-controlled events. Although they apply only to a restricted set of experiments, there are points in their favor: they appear to be more realistic than experimenter-controlled models, and more is known about them in terms of both theory and data than about models with experimenter-subject control.

1.5 Contingent Experiments and Contingent Events

It has been emphasized that the analysis of an experiment depends heavily on assumptions made about the subject and that the analysis is not an automatic consequence of the experimental design alone. Some of the current terminology can mislead one into thinking otherwise. Experiments have been categorized as "contingent" or "noncontingent," according to whether the occurrence of an outcome does or does not depend on the subject's response (Bush & Mosteller, 1955). It has been implied that contingent experiments correspond to experimenter-subject (contingent) events and noncontingent experiments to experimenter-controlled events and that this correspondence is unambiguous.

One difficulty with this method of classifying experiments lies in the definition of outcomes. If outcomes in the T-maze experiment are "left-arm baited" and "right-arm baited," then what the experimenter does can be predetermined and is noncontingent, although the relevant model

may be a contingent one. This example seems absurd because we have confidence in our intuitions in regard to what constitutes a reinforcing event for a rat: it is surely food versus nonfood rather than left-arm versus right-arm baited. In the analogous two-armed bandit experiment the ambiguity is more obvious, especially if the subject imagines that exactly one of the two responses is correct on each trial.

Even if we reject what may appear to be bizarre definitions of outcomes, the contingent-noncontingent distinction leads to difficulties. Let us suppose that in a T-maze or bandit experiment the subject is rewarded on a preassigned subset of trials regardless of his response. The experiment appears to be noncontingent, but in developing a model few would be willing to assume that the effect of the outcome is independent of the response.

To begin one's analysis of any experiment—contingent or noncontingent—with the *assumption* that events are not subject-controlled would appear to be somewhat rash, unless the assumption is treated as a null hypothesis or an approximating device and is later carefully tested. On the other hand, analysis by means of a model that incorporates subject control should reveal the fact that a model with experimenter control alone (or a single-event model) can represent the behavior, if such is the case.

2. AXIOMATICS AND HEURISTICS OF MODEL CONSTRUCTION

Various considerations, formal and informal, substantive and practical, have been used as guides in constructing models. So far I have discussed the factors that help to determine the number of distinct alternative model events that may occur on a trial and the determinants of their occurrence. There remains the problem of the mathematical form in which to express the effects of events. Suppose that there are two alternative responses, A_1 and A_2 , and that $\mathbf{p}_n = \text{Pr} \{A_1 \text{ on trial } n\}$. Let \mathbf{X}_n be a row-vector random variable⁶ with t elements corresponding to the t possible events. \mathbf{X}_n can take on the values $(1, 0, \dots, 0)$, $(0, 1, \dots, 0)$, \dots , $(0, 0, \dots, 1)$, which correspond to the occurrence on trial n of E_1, E_2, \dots, E_t . In general, a learning model is a function that gives \mathbf{p}_n in terms of the trial number and the sequence of events through trial $n - 1$,

$$\mathbf{p}_n = F(n, \mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots, \mathbf{X}_1), \tag{1}$$

⁶ In this chapter vector random variables are designated by boldface capitals and scalar random variables by boldface lower-case letters. Realizations (particular values) of random variables are designated by the corresponding lightface capital and lower-case letters.

where the initial probability and other parameters are suppressed.⁷ This equation makes it clear that \mathbf{p}_n , a function of random variables, is itself a random variable. Because it gives \mathbf{p}_n explicitly in terms of the event sequence, we refer to Eq. 1 as the *explicit equation* for \mathbf{p}_n . In this section I consider some of the arguments that have been used to restrict the form of F .⁸

2.1 Path-Independent Event Effects

At the start of the n th trial of an experiment, p_n is the subject's response probability, and the sequence X_1, X_2, \dots, X_{n-1} describes the course of the experiment up to this trial. This sequence, then, specifies the "path" traversed by the subject in attaining the probability p_n . A simplifying assumption which underlies most of the learning models that have been studied is that the event on trial n has an effect that depends on p_n but not on the path. The implication is that insofar as past experience has any influence on the future behavior of the process this influence is mediated entirely by the value of p_n . Another way of saying this is that the subject's state or "memory" is completely specified by his p -value.

The assumption of independence of path leads naturally to a recursive expression for the model and to the definition of a set of operators. The recursive form is given by

$$\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{X}_n), \quad (2)$$

⁷ Equation 1, and many of the other equations in this chapter in which response probabilities appear, may be regarded in two different ways. The first alternative, expressed by the notation in Eq. 1, is to regard \mathbf{p}_n as a function of random variables and therefore to consider \mathbf{p}_n itself as a random variable. This alternative is useful in emphasizing one of the important features of modern learning models—the fact that most of them specify a distribution of p -values on every trial after the first. By restricting the event sequence in any way, we determine a new, conditional distribution for the random variable. And we may be interested, for example, in determining the corresponding conditional expectation.

The second alternative is to regard the arguments in a formula such as Eq. 1 as *realizations* of the indicated random variables and the p -values it defines as *conditional probabilities*, conditioned by the particular event sequence. The formula is then more properly written as

$$\begin{aligned} p_n &= \Pr \{A_1 \text{ on trial } n \mid \mathbf{X}_1 = X_1, \mathbf{X}_2 = X_2, \dots, \mathbf{X}_{n-1} = X_{n-1}\} \\ &= F(n, X_{n-1}, X_{n-2}, \dots, X_1). \end{aligned}$$

Aside from easing the notation problem by reducing the number of boldface letters required, this alternative is occasionally useful; for example, the likelihood of a sequence of events can be expressed as the product of a sequence of such conditional probabilities. In this chapter, however, I make use of the first alternative.

⁸ I omit stimulus sampling considerations, which are discussed in Chapter 10.

whose k th component gives the cumulative number of occurrences of event E_k on trials $1, 2, \dots, n - 1$. We then have

$$\mathbf{W}_n = \mathbf{X}_1 + \mathbf{X}_2 + \dots + \mathbf{X}_{n-1}.$$

Under conditions of commutativity, the vector \mathbf{W}_n gives sufficient information about the path to determine the value of \mathbf{p}_n , and although a recursive expression does not result naturally Eq. 1 is simplified and becomes

$$\mathbf{p}_n = F(\mathbf{W}_n). \quad (4)$$

Several models have been proposed in terms of an explicit equation for \mathbf{p}_n of this simple form, in which \mathbf{p}_n depends only on the components of \mathbf{W}_n .

A further simplification arises if we require not only that F be a function of the components of \mathbf{W}_n but also that it be expressible as a continuous function of an argument that is linear in these components. Under these circumstances we have not only commutativity but path independence as well.⁹ As we shall see, path-independent models need not be commutative.

Events that commute are, in a certain sense, not subject to being forgotten. In a commutative model the effect of an event on \mathbf{p}_n is the same, whether it occurred on trial 1 or trial $n - 1$. Because the distant past is as significant as the immediate past, models of this kind tend to be relatively unresponsive to changes in the outcome sequence. (An example is given in Sec. 4.) Commutativity leads to considerable simplifications in the analysis of models and in estimation procedures.

2.3 Repeated Occurrence of a Single Event

What is the effect on response probabilities of the repeated occurrence of a particular event? This is an important consideration in formulating a

⁹ If F can be written as a continuous function of an argument that is linear in the components of \mathbf{W}_n , then events must have path-independent effects. The linearity condition requires that there exist some column vector, A , of coefficients for which $\mathbf{p}_n = F(\mathbf{W}_n) = G(\mathbf{W}_n \cdot A)$. Then

$$\begin{aligned} \mathbf{p}_{n+1} &= G(\mathbf{W}_{n+1} \cdot A) = G[(\mathbf{W}_n + \mathbf{X}_n) \cdot A] = G[\mathbf{W}_n \cdot A + \mathbf{X}_n \cdot A] \\ &= G[G^{-1}(\mathbf{p}_n) + \mathbf{X}_n \cdot A] = f(\mathbf{p}_n \cdot \mathbf{X}_n). \end{aligned}$$

A slight modification of the argument, which uses the continuity of G , is needed if G does not possess a unique inverse. One implication of some recent work by Luce (1963) on the commutativity property is that the converse of this result is also true: if a model is both commutative and path-independent then F can be written as a function of an argument that is linear in the components of \mathbf{W}_n .

model. Does \mathbf{p} converge to any value as E_k is repeated again and again, and, if so, what value? Except for events that have no effect at all on \mathbf{p} , it has been assumed in many applications of learning models that a particular event has either an incremental or a decremental effect on \mathbf{p} and has this effect whatever the value of \mathbf{p} as long as it is in the open $(0, 1)$ interval. In most cases, then, repetition of a particular event causes \mathbf{p} to converge to zero or one. Although in principle, learning models need not have this convergence property, it seems to be called for by most of the experiments to which they have been applied. This does not imply that the limiting behavior of these models always involves extreme response probabilities; in some instances successive repetition of any one event is improbable.

A related question concerns the sense in which the effect of an event is invariant during the course of an experiment. Neither commutativity nor path independence requires that the effect of an event on \mathbf{p} be in the same direction—incremental or decremental—throughout an experiment. Commutativity alone, for example, does not require that F in Eq. 4 be monotonic in any one of the components of \mathbf{W}_n . Path independence implies that if the direction of effect is to change at all in the course of an experiment the direction can depend at most on p_n . These possibilities and restrictions are relevant to the question whether existing models can handle such phenomena as the development of secondary reinforcement or any changes that may occur in the effects of reward and nonreward.

2.4 Combining-Classes Condition: Bush and Mosteller's Linear-Operator Models

Despite their strong implications, neither path independence nor commutativity is restrictive enough to produce useful models. Further assumptions are needed. Two important families of models have used path independence as a starting point. The first, the family with which most work has been done, comprises Bush and Mosteller's linear-operator models. In this section we shall consider the general characteristics of the linear-operator family and its application to two experiments. The second family, to be discussed in Sec. 2.5, with applications to the same pair of experiments, is Luce's response-strength operator family. In both families the additional assumptions are invariance conditions concerning multiple-response alternatives.

The combining-classes condition is a precise statement of the assumption that the definition of response alternatives is arbitrary and that therefore

any set of actions by the subject can be combined and treated as one alternative in a learning model.¹⁰ It might appear that the assumption is untenable because it ignores any natural response organization that may characterize the organism; this issue has yet to be clarified by theoretical and experimental work. The assumption is not inconsistent with current learning theory, however. Concerning the problem of response definition, Logan (1960, p. 117–120) has written:

The present approach assumes that responses that differ in any way whatsoever are different in the sense that they may be put into different response classes so that separate response tendencies can be calculated for them Differences among responses that are not importantly related to response tendency can be suppressed. This conception is consistent with what appears to be the most common basis of response aggregation, namely the conditions of reinforcement. Responses are separated if the environment . . . distinguishes between them in administering rewards The rules discussed above would permit any aggregation that preserves the reinforcement contingencies in the situation. Thus, if the reward is given independently of how the rat gets into the correct goal box of a T-maze, then the various ways of doing it can be classed together.

The combining-classes condition is concerned with an experiment in which three or more response alternatives, A_1, A_2, \dots, A_r , are initially defined. A subset, $A_{h+1}, A_{h+2}, \dots, A_r$, of the alternatives is to be combined into A^* , thus producing the reduced set of alternatives $A_1, A_2, \dots, A_h, A^*$. We consider the probability vector associated with the reduced set of alternatives after a particular event occurs. The combining-classes condition requires this vector to be the same, no matter whether the combining operation is performed before or after the event occurs; this invariance with respect to the combining operation is required to hold for all events and for any subset of the alternatives.

The result of applying the condition is that in a multiple-alternative

¹⁰ As originally stated by Bush, Mosteller, and Thompson (1954) and later by Bush and Mosteller (1955) the actions to be combined must have the same outcome probabilities associated with them. For example, if A_1 and A_2 are the alternatives to be combined, then $\Pr \{O_i | A_1\} = \Pr \{O_i | A_2\}$ must hold for all outcomes O_i . This was necessary because outcome probabilities conditional on responses were thought of as a part of the model, and without the equality the probability $\Pr \{O_i | A_1 \text{ or } A_2\}$ would not be well defined. From the viewpoint of this chapter, the model for a subject's behavior depends on the particular sequence of outcomes, and any probability mechanism that the experimenter uses to generate the sequence is extraneous to the model. What must be well defined is the sequence of outcomes actually applied to any one of the alternative responses, and this sequence is well defined even if the actions to be combined into one alternative are treated differently.

For a formal treatment of the combining classes condition see the references already cited, Mosteller (1955), or Bush (1960b).

experiment the effect of an event on the probability p_n of one of the alternatives can depend on p_n but cannot depend on the relative probabilities of choice among the other alternatives. To see this, let us suppose that the change in p_n does depend on the way in which $1 - p_n$ is distributed among the other alternatives. Then, if alternatives defined initially are combined *before* the event, the model cannot reflect the distribution of $1 - p_n$ among them. Thus, even if we can arrange to have p_{n+1} well defined, its value will not, in general, be the same as if the alternatives were combined *after* the event, in contradiction to the invariance assumption.

Together with the path-independence assumption, the combining-classes condition requires not only that each component of the p_{n+1} -vector depend on X_n and on the corresponding component of the p_n -vector only, but it also requires that this dependence be linear. The effect of a particular event on a set of response probabilities is therefore to transform each of them linearly, allowing us to write the operator Q_k as

$$p_{n+1} = Q_k p_n = \alpha_k p_n + a_k, \quad (5)$$

where the values of α_k and a_k are determined by the event E_k . This result can be proved only when $r \geq 3$, where r is the number of response alternatives. However, if we regard $r = 2$ as having arisen from the combination of a larger set of alternatives for which the condition holds, then the form of the operator is that given by Eq. 5.

One aspect of using linear transformations on response probabilities is that the parameters must be constrained to keep the probability within the unit interval. The constraints have usually been determined by the requirement that all possible values of p_n , from 0 to 1, be transformed into probabilities by the operator. A consequence of this requirement is that with r alternatives $-1/(r-1) \leq \alpha \leq 1$. It is in the spirit of the combining-classes condition that, in principle, an unlimited number of response classes may be defined. If r becomes arbitrarily large, then we see that one implication of the condition is that negative α 's are inadmissible.

In comparing operators Q_k and considering their properties, it is useful to define a new parameter $\lambda_k = a_k/(1 - \alpha_k)$. The general operator can then be rewritten as

$$p_{n+1} = Q_k p_n = \alpha_k p_n + (1 - \alpha_k) \lambda_k, \quad (6)$$

where the constraints are $0 \leq \alpha_k \leq 1$ and $0 \leq \lambda_k \leq 1$. The transformed probability $Q_k p_n$ may be thought of as a weighted sum of p_n and λ_k and the operator may be thought of as moving p_n in the direction of λ_k . Because $Q_k \lambda_k = \lambda_k$, the parameter λ_k is the *fixed point* of the operator: the operator does not alter a probability whose value is λ_k . In addition, when $\alpha_k \neq 1$, λ_k is the *limit point* of the operator: repeated occurrence of

the event E_k leads to repeated application of Q_k , and this causes the probability to approach λ_k asymptotically. This may be seen by first calculating the effect of m successive applications of the operator Q to p :

$$Q^m p = \alpha^m p + (1 - \alpha^m)\lambda = \lambda - \alpha^m(\lambda - p).$$

If $\alpha < 1$, $\lim_{m \rightarrow \infty} \alpha^m = 0$ and therefore $\lim_{m \rightarrow \infty} Q^m p = \lambda$.

As noted in Sec. 2.3, values of λ other than the extreme values zero or one have seldom been used in practice. An extreme limit point automatically requires that α be nonnegative if p_{n+1} is to be confined to the unit interval for all values of p_n . In order to justify the assumption that α is nonnegative, we can therefore usually appeal to the required limit point rather than to the extension of the combining-classes condition already mentioned. Because of this and because multiple-choice studies with $r \geq 3$ are relatively rare, the combining-classes condition has never been put directly to the test.

The parameter α_k may be thought of as a learning-rate parameter. Its value is a measure of *ineffectiveness* of the event E_k in altering the response probability. When α_k takes on its maximum value of 1, then Q_k is an *identity operator* and event E_k induces no change in the response probability. The smaller the value of α_k , the more p_n is moved in the direction of λ_k by the operator Q_k .

For the two extreme limit points, the operators are either of the form $Q_1 p = \alpha_1 p$ or $Q_2 p = \alpha_2 p + (1 - \alpha_2)$, and these are the two operators that have most commonly been used. In general, pairs of operators do not commute, but under certain conditions (either they have equal limit points or one of them is the identity operator) they do. When all pairs of operators in a model commute, the explicit expression for p_n in terms of the path has a simple form. When the operators in a two-choice experiment are all of the form of Q_2 , it is convenient to deal with $q = 1 - p$, the probability of the other response, and to make use of the complementary operators whose form is $\bar{Q}_2 q = \alpha_2 q$.

To be concrete, we consider examples of the Bush-Mosteller model for two of the experiments discussed in Sec. 1.

ESCAPE-AVOIDANCE SHUTTLEBOX. We interpret this experiment to consist of two subject-controlled events: escape (shock) and avoidance. Both events reduce the probability p_n of escape in the direction of the limit point $\lambda = 0$. It is convenient to define a binary random variable that represents the event on trial n ,

$$\mathbf{x}_n = \begin{cases} 0 & \text{if avoidance } (E_1) \\ 1 & \text{if escape } (E_2), \end{cases} \quad (7)$$

with a probability distribution given by $\Pr \{x_n = 1\} = p_n$, $\Pr \{x_n = 0\} = 1 - p_n$. The operators and the rule for their application are

$$p_{n+1} = \begin{cases} Q_1 p_n = \alpha_1 p_n & \text{if } x_n = 0 \text{ (i.e., with probability } 1 - p_n) \\ Q_2 p_n = \alpha_2 p_n & \text{if } x_n = 1 \text{ (i.e., with probability } p_n). \end{cases} \quad (8)$$

Because events are subject-controlled, the sequence of operators (events) is not predetermined. The vector \mathbf{X}_n (Sec. 2) is given by $(x_n, 1 - x_n)$ and the recursive form of Eq. 2 (Sec. 2.1) is given by

$$p_{n+1} = f(p_n; \mathbf{X}_n) = \alpha_2^{x_n} \alpha_1^{(1-x_n)} p_n. \quad (9)$$

The operators have equal limit points and therefore commute. This makes possible a simple explicit formula for p_n . Let $\mathbf{W}_n = (s_n, t_n)$, where

$$s_n = \sum_{j=1}^{n-1} x_j$$

is the number of shocks before trial n and

$$t_n = \sum_{j=1}^{n-1} (1 - x_j)$$

is the number of avoidances before trial n . The explicit form of Eq. 4 (Sec. 2.2) is given by

$$p_n = F(\mathbf{W}_n) = \alpha_2^{s_n} \alpha_1^{t_n} p_1. \quad (10)$$

Later I shall make use of the fact that, by redefining the parameters of the model, p_n may be written as a function of an expression that is linear in the components of \mathbf{W}_n . To do this, let $p_1 = e^{-a}$, $\alpha_1 = e^{-b}$, and $\alpha_2 = e^{-c}$. Then Eq. 10 becomes

$$p_n = \exp [-(a + bt_n + cs_n)]. \quad (11)$$

It should be emphasized that, for this model, t_n and s_n and, therefore, p_n are random variables whose values are unknown until trial $n - 1$. The set of trials is a dependent sequence.

PREDICTION EXPERIMENT. For purposes of illustration we make the customary assumption that this experiment consists of two experimenter-controlled events. Onset of the left light (E_1) increases the probability p_n of a left button press toward the limit point $\lambda = 1$. Onset of the right light decreases p_n toward the limit point $\lambda = 0$.

The events are assumed to be complementary (Sec. 1.2), and therefore the operators have equal rate parameters ($\alpha_1 = \alpha_2 = \alpha$) and complementary limit points ($\lambda_1 = 1 - \lambda_2$). It is convenient to define a binary variable that represents the event on trial n ,

$$y_n = \begin{cases} 0 & \text{if right light } (E_2) \\ 1 & \text{if left light } (E_1). \end{cases}$$

Because events are experimenter-controlled, the sequence y_1, y_2, \dots can be predetermined. In some experiments a random device may be used to generate the actual sequence used. For example, the $\{y_n\}$ may be a realization of a sequence $\{\mathbf{y}_n\}$ of independent random variables with $\Pr \{y_n = 1\} = \pi$. However, insofar as we are interested in the behavior of the subject, the actual sequence, rather than any properties of the random device used to generate it, is of interest. It is shown later how simplifying approximations may be developed by assuming that the subject has experienced the average of all the sequences that the random device generates. For the purpose of such approximations, which, of course, involve loss of information, y_n may be considered a random variable with a probability distribution. The more exact treatment, however, deals with the experiment conditional on the actual outcome sequences that are used.

The operators and the rules for their application are

$$P_{n+1} = \begin{cases} Q_1 p_n = \alpha p_n + 1 - \alpha & \text{if } y_n = 1 \\ Q_2 p_n = \alpha p_n & \text{if } y_n = 0. \end{cases} \quad (12)$$

The vector X_n (Sec. 2) is given by $(y_n, 1 - y_n)$ and the recursive form of Eq. 2 is given by

$$P_{n+1} = \alpha p_n + (1 - \alpha)y_n. \quad (13)$$

Note that in the exact treatment p_n is not a random variable, unlike the case for an experiment with subject control. The operators do not commute, and therefore the cumulative number of E_1 's and E_2 's does not determine p_n uniquely. The explicit formula for p_n , in contrast to the shuttlebox example, includes the entire sequence y_1, y_2, \dots and is given by

$$p_n = F(n, X_1, X_2, \dots, X_{n-1}) = \alpha^{n-1} p_1 + (1 - \alpha) \sum_{j=1}^{n-1} \alpha^{n-1-j} y_j. \quad (14)$$

Equation 14 shows that (when $\alpha < 1$) a recent event has more effect on p_n than an event in the distant past. By contrast, Eq. 10 indicates that for the shuttlebox experiment there is no "forgetting" in this sense: given that the event sequence has a particular number of avoidances, the effect of an early avoidance on p_n is no different from the effect of a late avoidance. As I mentioned earlier, this absence of forgetting is a characteristic of all experiments with commutative events.

The model for the prediction experiment consists of a sequence of independent binomial trials: the p_n -sequence is determined by the y_n -sequence which is independent of all responses.

2.5 Independence from Irrelevant Alternatives: Luce's Beta Response-Strength Model

Stimulus-response theory has traditionally treated response probability as deriving from a more fundamental response-strength variable. For example, Hull (1943, Chapter 18) conceived of the probability of reaction (where the alternative was nonreaction) as dependent, in a complicated way, on a reaction-potential variable that is more fundamental in his system than the probability itself. The momentary reaction potential was thought to be the sum of an underlying value and an error (behavioral oscillation) that had a truncated normal distribution: the response would occur if the effective reaction potential exceeded a threshold value. The result of these considerations was that reaction probability was related to reaction potential by means of a cumulative normal distribution which was altered to make possible zero probabilities. The alteration implied that a range of reaction potentials could give rise to the same (zero) probability. Such a threshold mechanism has not been explicitly embodied in any of the modern stochastic learning models.

The reaction potential variable was more fundamental partly because it changed in a simple way in response to experimental events and partly because the state of the organism was more completely described by the reaction potential than by the probability. The last observation is clearer if we turn from the single-response situation considered by Hull¹¹ to an experiment with two symmetric responses. The viewpoint to be considered is that such an experiment places into competition two responses, each of which may vary independently in its strength. Let us suppose that response strengths, symbolized by $v(1)$ and $v(2)$, are associated with each of the two responses and that the probability of a response is given by the ratio of its strength to the sum of the two strengths: $p\{1\} = v(1)/[v(1) + v(2)]$. Then, although the two strengths determine the probability uniquely, knowledge of the probability can tell us only the ratio of the strengths, $v(1)/v(2)$. Multiplying both strengths by the same constant does not alter the response probability, but it might, for example, correspond to the change from an avoidance-avoidance to an approach-approach conflict and might be revealed by response times or amplitudes. The response strengths therefore provide a more basic description of the state of the organism than the response probabilities. This is the sort of thinking that might lead one to favor a learning model whose underlying

¹¹ For the symmetric two-choice experiment a response-strength analysis that is considerably different from the one discussed here is given by Hull (1952).

variables are response strengths. A critical question regarding this viewpoint is whether there are aspects of the behavior in choice learning experiments that can be accounted for by changes in response strengths but that are not functions of response probabilities alone.

An invariance condition concerning multiple response alternatives, but different from the combining-classes condition, is used by Luce (1959) in arriving at his beta response-strength model. Path independence of the sequence of response strengths is also assumed, and within the model this entails path independence of the sequence of probability vectors. The invariance condition (Luce's Axiom 1) states that, in an experiment in which one of a set of responses is made, the ratio of the probabilities of two alternatives is invariant with respect to changes in the set of remaining alternatives from which the subject can select. As stated, the condition applies to choice situations in which the probabilities of choice from a constant set of alternatives are unchanging. Nonetheless, by assuming that the condition holds during any instant of learning, we can use it to restrict the form of a learning model. The condition implies that a positive response-strength function, $v(j)$, can be defined over the set of alternatives $\{A_j\}$ with the property that

$$\Pr \{A_k\} = \frac{v(k)}{\sum_j v(j)}. \quad (15)$$

The subsequent argument rests significantly on the fact that $v(j)$ is a ratio scale and that the scale values are determined by the choice probabilities only up to multiplication by a positive constant; that is, the unit of the response-strength scale is arbitrary.

The argument begins with the idea that in a learning experiment the effects of an event on the organism can be thought of as a transformation of the response strengths. Two steps in the argument are critical in restricting the form of this transformation. First, it is observed that because the unit of response strength is arbitrary the transformation f must be invariant with respect to changes in this unit: $f[kv(j)] = kf[v(j)]$. Second, it is assumed that the scale of response strength is unbounded and that therefore any real number is a possible scale value. The independence-of-unit condition, together with the unboundedness of the scale, leads to the powerful conclusion that the only admissible transformation is multiplication by a constant.¹² The requirement that response strengths

¹² As suggested by Violet Cane (1960), it is not possible to have both an unbounded response-strength scale and choice probabilities equal to zero or unity ("perfect discrimination"); for, if choice probabilities can take on all values in the closed unit interval, then the v -scale must map onto this closed interval and must therefore itself extend over a closed, and thus bounded, interval. But the unboundedness of the scale

be positive implies that the multiplying constant must be positive. Path independence implies that the constant depends only on the event and not on the trial number or the response strength. This argument completely defines the form of a learning model—called the “beta model”—for experiments with two alternative responses. The model defines a stochastic process on response strengths, which in turn determines a stochastic process on the choice probabilities.

When event E_k occurs, let $v(1)$ and $v(2)$ be transformed into $a_k v(1)$ and $b_k v(2)$. The new probability is then

$$\Pr \{1\} = \frac{a_k v(1)}{a_k v(1) + b_k v(2)} = \frac{(a_k/b_k)[v(1)/v(2)]}{1 + (a_k/b_k)[v(1)/v(2)]}.$$

If we let $v = v(1)/v(2)$ be the ratio of response strengths and $\beta_k = a_k/b_k$ be the ratio of constants, the original probability is $v/(1 + v)$ and the transformed probability is $\beta_k v/(1 + \beta_k v)$. The ratio β_k and the relative response strength v are sufficient to determine p and its transformed value. Because response strengths are important in this chapter only insofar as they govern probabilities, the simplified notation is adequate. We let v_n be the relative response strength $v(1)/v(2)$ on trial n , let β_k be the multiplier of v_n that is associated with the event E_k , and let p_n be $\Pr \{A_1 \text{ on trial } n\}$. Then

$$p_n = \frac{v_n}{1 + v_n} = \frac{1}{1 + v_n^{-1}} \quad \text{and} \quad v_n = \frac{p_n}{1 - p_n}.$$

Moreover, if event E_k occurs on trial n ,

$$p_{n+1} = \frac{\beta_k v_n}{1 + \beta_k v_n} = \frac{\beta_k p_n}{(1 - p_n) + \beta_k p_n} = Q_k p_n, \tag{16}$$

which gives the corresponding nonlinear transformation on response probability.

A number of implications of the model can be seen immediately from the form of the probability operator. If E_k has an incremental (decremental) effect on $\Pr \{A_1\}$, then $\beta_k > 1 (< 1)$. An identity operator results

is an important feature of the argument that forces the learning transformation to be multiplicative. It therefore appears that Luce’s axiom leads to a multiplicative learning model only when it is combined with the assumption that response probabilities can never be exactly zero or unity. In practice, this assumption is not serious, since a finite number of observations do not allow us to distinguish between a probability that is exactly unity and one that is arbitrarily close to that value. The fact that the additional assumption is needed, however, makes it difficult to disprove the axiom on the basis of a failure of the learning model, since the fault may lie elsewhere.

when $\beta_k = 1$. The only limit points possible are $p = 0$ and $p = 1$, which are obtained, respectively, when $\beta_k < 1$ and $\beta_k > 1$. This follows because

$$Q^m p = \frac{\beta^m v}{1 + \beta^m v} = \frac{v}{v + \beta^{-m}}$$

and, when $\beta \neq 1$, either β^m or β^{-m} approaches zero. These properties imply that the effect of an event on p_n must always be in the same direction; all operators are unidirectional in contrast to operators in the linear model, which may have zero points other than zero and unity. The restriction to extreme limit points appears not to be serious in practice, however; as noted in the Sec. 2.4, most experiments seem to call for unidirectional operators.

Perhaps more important from the viewpoint of applications is the fact that operators in the beta response-strength model must always commute; the model requires that events in learning experiments have commutative effects. That nonlinear probability operators of the form given by Eq. 16 commute can be shown directly, or can be seen more simply by noting the commutativity of the multiplicative transformations of v_n , to which such operators correspond. Whether or not commutativity is realistic, it is a desirable simplifying feature of the model. A final property is that the model cannot produce learning when $p_1 = 0$ because this requires that v_1 , hence all v_n , be zero.

Let us consider applications of the beta model to the experiments discussed in Sec. 2.4. When applicable, the same definitions are used as in that section.

ESCAPE-AVOIDANCE SHUTTLEBOX. It is convenient to let p_n be the probability of escape (response A_2 , event E_2) and to let v_n be the ratio of escape strength to avoidance strength. Both events reduce the probability of escape and therefore both $\beta_1 < 1$ and $\beta_2 < 1$. The binary random variable x_n is defined as in Eq. 7. The operators and the rules for their application are

$$\mathbf{P}_{n+1} = \begin{cases} Q_1 \mathbf{P}_n = \frac{\beta_1 \mathbf{P}_n}{(1 - \mathbf{P}_n) + \beta_1 \mathbf{P}_n} & \text{if } x_n = 0 \\ & \text{(i.e., with probability } 1 - \mathbf{P}_n) \\ Q_2 \mathbf{P}_n = \frac{\beta_2 \mathbf{P}_n}{(1 - \mathbf{P}_n) + \beta_2 \mathbf{P}_n} & \text{if } x_n = 1 \\ & \text{(i.e., with probability } \mathbf{P}_n). \end{cases} \quad (17)$$

The recursive form of Eq. 2 is given by

$$\mathbf{P}_{n+1} = f(\mathbf{P}_n; \mathbf{X}_n) = \frac{\beta_2^{x_n} \beta_1^{(1-x_n)} \mathbf{P}_n}{(1 - \mathbf{P}_n) + \beta_2^{x_n} \beta_1^{(1-x_n)} \mathbf{P}_n}. \quad (18)$$

Both expressions are cumbersome. More light is shed by the explicit formula

$$p_n = F(W_n) = \frac{1}{1 + \beta_2^{-s_n} \beta_1^{-t_n} v_1^{-1}}. \tag{19}$$

Redefining the parameters simplifies Eq. 19. Let $v_1 = e^a$, $\beta_1 = e^b$ and $\beta_2 = e^c$. Then the expression becomes

$$p_n = \frac{1}{1 + \exp [-(a + b t_n + c s_n)]}. \tag{20}$$

It is instructive to compare this explicit formula with Eq. 11 for the linear model. In the usual experiment a would be positive and b and c would be negative, unlike the coefficients in Eq. 11, all of which are positive. (Recall that as t_n, s_n increase p_n decreases. These definitions are awkward, but they will facilitate matters later on.) Again it should be noted that t_n and s_n are random variables whose behavior is governed by p -values earlier in the sequence and that the model therefore defines a dependent sequence of trials.

PREDICTION EXPERIMENT. Experimenter-controlled events are assumed here, as in Sec. 2.4. The $\{p_n\}$ and $\{y_n\}$ are defined as in that section. The complementarity of events demands that $\beta_1 = \beta_2^{-1} = \beta > 1$. This can be seen by noting that if E_1 transforms $v(1)/v(2)$ into $\beta v(1)/v(2)$ then for operators to be complementary E_2 must transform $v(2)/v(1)$ into $\beta v(2)/v(1)$. The operators and the rules for their application therefore are

$$p_{n+1} = \begin{cases} Q_1 p_n = \frac{\beta p_n}{(1 - p_n) + \beta p_n} & \text{if } y_n = 1 \\ Q_2 p_n = \frac{p_n}{\beta(1 - p_n) + p_n} & \text{if } y_n = 0. \end{cases} \tag{21}$$

The recursive expression is given by

$$p_{n+1} = \frac{\beta^{y_n} p_n}{\beta^{(1-y_n)}(1 - p_n) + \beta^{y_n} p_n}. \tag{22}$$

Because of the universal commutativity of the beta model, the explicit formula is simple in contrast to Eq. 14. We have $W_n = (l_n, r_n)$, where

$$l_n = \sum_{j=1}^{n-1} y_j$$

is the number of left-light outcomes before trial n and

$$r_n = \sum_{j=1}^{n-1} (1 - y_j)$$

is the corresponding number of right-light outcomes. We define $d_n = l_n - r_n$ to be the difference between these numbers. The explicit formula is then

$$p_n = F(W_n) = \frac{1}{1 + \beta^{a_n} v_1^{-1}}. \quad (23)$$

For this model commutativity seems to be of more use than path independence in simplifying formulas. Again we can define new parameters $v_1 = e^a$, $\beta = e^{-b}$ to obtain

$$p_n = \frac{1}{1 + \exp[-(a + b d_n)]}. \quad (24)$$

Equation 24 indicates that the response probability is expressed in terms of d_n by means of the well-known logistic function. Equation 20 is a generalized form of this function. All of the two-alternative beta models have explicit formulas that are (generalized) logistics. Because the logistic function is similar to the cumulative normal distribution, the relation in the beta model between response strength and probability is reminiscent of Hull's treatment of this problem.

2.6 Urn Schemes and Explicit Forms

The treatment of examples in Secs. 2.4 and 2.5 illustrates two of the alternative ways of regarding a stochastic learning model. One approach is to specify the change in \mathbf{p}_n that is induced by the event (represented by \mathbf{X}_n) on trial n . This change in probability depends, in general, on $\mathbf{X}_1, \dots, \mathbf{X}_{n-1}$ as well as on \mathbf{X}_n . In the general recursive formula for response probability we therefore express \mathbf{p}_{n+1} as a function of \mathbf{p}_n and the events through trial n :

$$\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{X}_n, \mathbf{X}_{n-1}, \dots, \mathbf{X}_1).$$

If the model is path-independent, then \mathbf{p}_{n+1} is uniquely specified by \mathbf{p}_n and \mathbf{X}_n , and the expression may be simplified to give the recursive formula of Sec. 2.1,

$$\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{X}_n),$$

and its corresponding operator expressions. The second approach is to specify the way in which \mathbf{p}_n depends on the entire sequence of events through trial $n - 1$. This is done by the explicit formula in which \mathbf{p}_n is expressed as a function of the event sequence:

$$\mathbf{p}_n = F(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots, \mathbf{X}_1).$$

A model may have a more "natural" representation in one of these forms than in the other. In this section I discuss models for which the explicit form is the more natural.

URN SCHEMES. Among the devices traditionally used in fields other than learning to represent the aftereffects of events on probabilities are games of chance known as urn schemes (Feller, 1957, Chapter V). An urn contains different kinds of balls, each kind representing an event. The occurrence of an event is represented by randomly drawing a ball from the urn. After-effects are represented by changes in the urn's composition. Schemes of this kind are among the earliest stochastic models for learning (Thurstone, 1930; Gulliksen, 1934) and are still of interest (Audley & Jonckheere, 1956; Bush & Mosteller, 1959). The stimulus-sampling models discussed in Chapter 10 may be regarded as urn schemes whose balls are interpreted as "elements" of the stimulus situation. In contrast, Thurstone (1930) suggested that the balls in his scheme be interpreted as elements of response classes. In all of these examples two kinds of balls, corresponding to two events, are used. An urn scheme is introduced to help make concrete one's intuitive ideas about the learning process. Except in the case of stimulus-sampling theory, the interpretation of the balls as psychological entities has not been pressed far.

The general scheme discussed by Audley and Jonckheere (1956) encompasses most of the others as special cases. It is designed for experiments with two subject-controlled events. On trial 1 the urn contains w_1 white balls and r_1 red balls. A ball is selected at random. If it is white, event E_1 occurs ($x_1 = 0$), the ball is replaced, and the contents of the urn are changed so that there are now $w_1 + w$ white balls and $r_1 + r$ red balls. If the chosen ball is red, event E_2 occurs ($x_1 = 1$), the ball is replaced, and the new numbers are $w_1 + w'$ and $r_1 + r'$. This process is repeated on each trial. The quantities w, w', r and r' have fixed integral values that may be positive, zero, or negative, but, if any of them are negative, then arrangements must be made so that the urn always contains at least one ball and so that the number of balls of either color is never negative.

Let $t_n = \sum_{j=1}^{n-1} (1 - x_j)$ be the number of occurrences of E_1 and $s_n = \sum_{j=1}^{n-1} x_j$ be the number of E_2 occurrences before trial n . Let w_n and r_n be the number of white and red balls in the urn before the n th trial. Then

$$w_n = w_1 + wt_n + w's_n, \quad r_n = r_1 + rt_n + r's_n$$

and

$$\begin{aligned} p_n &= \Pr \{ \text{event } E_2 \text{ on trial } n \} \\ &= \frac{r_n}{r_n + w_n} \\ &= \frac{r_1 + rt_n + r's_n}{(r_1 + w_1) + (r + w)t_n + (r' + w')s_n}. \end{aligned} \tag{25}$$

Equation 25 gives the explicit formula for \mathbf{p}_n . It demonstrates the most important property of these models—their commutativity.

If w_n and r_n are interpreted as response strengths, the model can be regarded as a description of additive (rather than multiplicative) transformations of these strengths.¹³

The recursive formula and corresponding operators are unwieldy and are not given. Suffice it to say that the operators are nonlinear and depend on the trial number (that is, on the path length) but not on the particular sequence of preceding events (the content of the path). The model, therefore, is only quasi-independent of path (Sec. 2.1). This is the case because the change induced by an event in the proportion of red balls depends on p_n , on the numbers of reds and whites added (which depend only on the event), and on the total number of balls in the urn before the event occurred (which can in general be inferred only from knowledge of both p_n and n).

Two special cases of the urn scheme that are exceptions to the foregoing statement and produce path-independent models are (1) those for which $r + w = r' + w' = 0$, so that the total number of balls is constant, and (2) those for which either $r = r' = 0$ or $w = w' = 0$, so that the number of balls of one color is constant. The first condition is met by Estes' model, which, however, departs in another respect from the general scheme: its additive increments vary with the changing composition of the urn instead of being constant. (This modification sacrifices commutativity, but it is necessary if the probability operators are to be linear. The modification follows from the identification of balls with stimulus elements, and so is less artificial than it sounds.)

The second condition is assumed in the urn scheme that Bush and Mosteller (1959) apply to the shuttlebox experiment. They assume that $r = r' = 0$, so that only white balls are added to the urn; neither escape nor avoidance alters the "strength" of the escape response. The model is modified so that w and w' are continuous parameters rather than discrete numbers, as they would have to be in a strict interpretation as numbers of balls. The result may be expressed in simple form by defining $a = (r_1 + w_1)/r_1$. To be consistent with Eqs. 11 and 20, we let $b = w/r_1$ and $c = w'/r_1$ and obtain

$$\mathbf{p}_n = \frac{1}{a + bt_n + cs_n}. \quad (26)$$

¹³ The model is also appropriate if it is thought that strength $v(j)$ is transformed multiplicatively but that response probability depends on logarithms of strengths: $p(1) = \log v(1)/\log [v(1)v(2)]$. In such a case w_n and r_n are interpreted as logarithms of response strengths.

The operators are given by

$$\mathbf{p}_{n+1} = \begin{cases} Q_1 \mathbf{p}_n = \frac{\mathbf{p}_n}{1 + b \mathbf{p}_n} & \text{if } \mathbf{x}_n = 0 \text{ (i.e., with probability } 1 - \mathbf{p}_n) \\ Q_2 \mathbf{p}_n = \frac{\mathbf{p}_n}{1 + c \mathbf{p}_n} & \text{if } \mathbf{x}_n = 1 \text{ (i.e., with probability } \mathbf{p}_n). \end{cases} \quad (27)$$

LINEAR MODELS FOR SEQUENTIAL DEPENDENCE. It has been indicated earlier that the responses produced by learning models consist of stochastically dependent sequences, except for the case of experimenter-controlled events. Moreover, insofar as experimenter control is present, the sequence of responses will be dependent on the sequence of outcomes. The autocorrelation of responses and the correlation of responses with outcomes are interesting in themselves, whether in learning experiments or, for example, in trial-by-trial psychophysical experiments in which there is no over-all trend in response probability. Several models have arisen directly from hypotheses about repetition or alternation tendencies that perturb the learning process and produce a degree or kind of response-response or response-outcome dependence that is unexpected on the basis of other learning models. The example to be mentioned is neither path-independent nor commutative.

The one trial perseveration model (Sternberg, 1959a,b) is suggested by the following observation: in certain two-choice experiments with symmetric responses the probability of a particular response is greater on a trial after it occurs and is rewarded than on a trial after the alternative response occurs and is not rewarded. There are several possible explanations. One is that reward has an immediate and lasting effect on \mathbf{p}_n that is greater than the effect of nonreward. This hypothesis attributes the observed effect to a differential influence of outcomes in the cumulative learning process. One of the models already discussed could be used to describe this mechanism: for example, the model given by Eq. 8 (Sec. 2.4) with $\alpha_1 < \alpha_2$.

A second hypothesis is that the two outcomes are equally effective (i.e., they are symmetric) but that there is a short-term one-trial tendency to repeat the response just made. This hypothesis, when applied to an experiment with 100:0 reward¹⁴ leads to the one-trial perseveration model.

Without the repetition tendency, the assumption of outcome symmetry leads to a model with experimenter-controlled events of the kind that was applied in Sec. 2.4 to the prediction experiment. The 100:0 reward

¹⁴ The term “ $\pi_1:\pi_2$ reward” describes a two-choice experiment in which one choice is rewarded with probability π_1 and the other with probability π_2 .

schedule implies that $y_n = 0$ on all trials and therefore that the same operator, Q_2 in Eq. 12, is applied on every trial. Equation 14 shows the explicit formula to be

$$p_n = \alpha^{n-1}p_1. \quad (28)$$

This single-operator model was discussed by Bush and Sternberg (1959). It may also be regarded as a special case of the subject-controlled model used for the shuttlebox (Eq. 8) with $\alpha_1 = \alpha_2 = \alpha$.

In developing the perseveration model, the single-operator model is taken to represent the "underlying" learning process. Define \mathbf{x}_n so that $\Pr\{\mathbf{x}_n = 1\} = \mathbf{p}_n$ and $\Pr\{\mathbf{x}_n = 0\} = 1 - \mathbf{p}_n$. We note that the strongest possible tendency to repeat the previous response can be described by the model $\mathbf{p}_n = \mathbf{x}_{n-1}$. This is the effect that perturbs the learning process.

To combine the underlying and perturbing processes, we take a weighted combination of the two, with nonnegative weights $1 - \beta$ and β . This gives the explicit formula¹⁵ for the subject-controlled model:

$$\mathbf{p}_n = F(n, \mathbf{x}_{n-1}) = (1 - \beta)\alpha^{n-1}p_1 + \beta\mathbf{x}_{n-1}, \quad (n \geq 2). \quad (29)$$

Knowledge of the trial number and of only the last response is needed to determine the value of \mathbf{p}_n . The two possible values that \mathbf{p}_n can have on a particular trial differ by the (constant) value of β ; \mathbf{p}_n takes on the higher of the two values when $\mathbf{x}_{n-1} = 1$ and the lower when $\mathbf{x}_{n-1} = 0$. The extent to which the learning process is perturbed by the repetition tendency is greater with larger β .

That the model is path-dependent is shown by the form of its recursive expression:

$$\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{x}_n, \mathbf{x}_{n-1}) = \alpha\mathbf{p}_n + \beta\mathbf{x}_n - \alpha\beta\mathbf{x}_{n-1}, \quad (n \geq 2). \quad (30)$$

Knowledge of the values of \mathbf{p}_n and \mathbf{x}_n alone is insufficient to specify the value of \mathbf{p}_{n+1} . For this model, in contrast to most others, more past history is needed in order to specify \mathbf{p}_n by the recursive form than by the explicit form. Data from a two-armed bandit experiment have been fruitfully analyzed with the perseveration model.

The development of the perseveration model illustrates a technique that is of general applicability and is occasionally of interest. A tendency to alternate responses may be represented by a similar device. Linear equations may also be used to represent positive or negative correlation between outcome and subsequent response—for example, a tendency in a

¹⁵ A trivial modification in this expression is made by Sternberg (1959a) based on considerations about starting values.

prediction experiment to avoid predicting the event that most recently occurred.

LOGISTIC MODELS. The most common approach to the construction of models begins with an expression for trial-to-trial probability changes, an expression that seems plausible and that may be buttressed by more general assumptions. An alternative approach is to consider what features of the entire event sequence might affect p_n and to postulate a plausible expression for this dependence in terms of an explicit formula. The second approach is exemplified by the perseveration model and also by a suggestion by Cox based on his work on the regression analysis of binary sequences (1958).

In many of the models we have considered the problem arises of containing p_n in the unit interval, and it is solved by restrictions on the parameter values, restrictions that are occasionally complicated and interdependent. The problem is that although p_n lies in the unit interval the variables on which it may depend, such as total errors or the difference between the number of left-light and right-light onsets, may assume arbitrarily large positive or negative values. The probability itself, therefore, cannot depend linearly on these variables. If a linear relationship is desired, then what is needed is a transformation of p_n that maps the unit interval into the real line.¹⁶

Such a transformation is given by $\text{logit } p = \log [p/(1 - p)]$. Suppose that this quantity depends linearly on a variable, x , so that $\text{logit } p = a + bx$. Then the function that relates p to x is the logistic function that we have already encountered in Eq. 24 and is represented by

$$p = \frac{1}{1 + \exp [-(a + bx)]}. \quad (31)$$

As was mentioned in Sec. 2.5, the logistic function is similar in form to the normal ogive and therefore it closely resembles Hull's relation between probability and reaction potential. One advantage of the logistic transformation is that no constraints on the parameters are necessary. A second advantage, to be discussed later, is that good estimates of the parameter values are readily obtained.

Cox (1958) has observed that many studies utilizing stochastic learning models,

... have led to formidable statistical problems of fitting and testing. When these studies aim at linking the observations to a neurophysiological mechanism, it is reasonable to take the best model practicable and to wrestle as vigorously

¹⁶ When the dependent variables are nonnegative, the unit interval needs to be mapped only into the positive reals. This can be achieved, for example, by arranging that the transformations p^{-1} or $\log(p^{-1})$ depend linearly on the variables, as illustrated by Eq. 11 and Eq. 25.

as possible with the resulting statistical complications. If, however, the object is primarily the reduction of data to a manageable and revealing form, it seems fair to take for the probability of a success . . . as simple an expression as possible that seems to be the right general shape and which is flexible enough to represent the various possible dependencies that one wants to examine. For this the logistic seems a good thing to consider.

The desirable features of the logistic function carry over into its generalized form, in which logit p is a linear function of several variables. When these variables are given by the components of \mathbf{W}_n (the cumulative number of times each of the events E_1, E_2, \dots, E_t has occurred in the first $n - 1$ trials), then the logistic function is exactly equivalent to Luce's beta model, so that the same model is obtained from quite different considerations. An example of the logistic function generalized to two dependent variables is given by Eq. 20 for the shuttlebox experiment.

A second example of a generalized logistic function, one that does not follow from Luce's axioms, is given by the analogue of the one-trial perseveration model (Eq. 29) in which

$$\text{logit } \mathbf{p}_n = a + bn + c\mathbf{x}_{n-1}$$

or

$$\mathbf{p}_n = \frac{1}{1 + \exp [-(a + bn + c\mathbf{x}_{n-1})]}. \quad (32)$$

2.7 Event Effects and Their Invariance

The magnitude of the effect of an event is usually represented by the value of a parameter that, ideally, depends only on constant features of the organism and of the apparatus and therefore does not change during the course of an experiment. There are some experiments or phases of experiments in which the ideal is clearly not achieved, at least not in the context of the available models and the customary definitions of events. Magazine training, detailed instructions, practice trials, and other types of pretraining are some of the devices used to overcome this difficulty. Probably few investigators believe that the condition is ever exactly met in actual experiments, but the principle of parameter invariance within an experiment is accepted as a working rule with the hope that it will at least approximate the truth. (It is also desirable that event effects be invariant from experiment to experiment; this principle provides one test of a model.)

A careful distinction should be drawn between invariance of parameter values and equality of an event's effects in the course of an experiment.

All of the models that have thus far been mentioned imply that the probability change induced by an event *varies* systematically in the course of an experiment; different models specify different forms for this variation. It is therefore only in the context of a particular model that the question of parameter invariance makes sense. Insofar as changes in event effects are in accord with the model, parameters will appear invariant, and we would be inclined to favor the model.

In most models, event effects, defined as probability differences, change because $p_{n+1} - p_n$ depends on at least the value of p_n . This dependence arises in part from the already mentioned need to avoid using transition rules that may take p_n outside the unit interval. But the simple fact that most learning curves (of probability, time, or speed versus trials) are not linear first gave rise to the idea that event effects change.

Gulliksen (1934) reviewed the early mathematical work on the form of the learning curve, and he showed that most of it was based on one of two assumptions about changes in the effect of a trial event. Let t represent time or trials and let y represent a performance measure. Models of Type A begin with the assumption that the improvement in performance induced by an event is proportional to the amount of improvement still possible. The chemical analogy was the monomolecular reaction. This assumption led to a differential equation approximation, $dy/dt = a(b - y)$ whose solution is the exponential growth function $y = b - c \exp(-at)$. Models of Type B begin with the assumption that the improvement in performance induced by an event is proportional to the product of the improvement still possible and the amount already achieved. The chemical analogy was the monomolecular autocatalytic reaction. The assumption led to a differential equation approximation, $dy/dt = ay(b - y)$, whose solution is a logistic function $y = b/[1 + c \exp(-At)]$.

The modern versions of these two models are, of course, Bush and Mosteller's linear-operator models and Luce's beta models. In the linear models the quantity $p_{n+1} - p_n$ is proportional to $\lambda_k - p_n$, the magnitude of the change still possible on repeated occurrences of the event E_k . If all events change p_n in the same direction, let us say toward $p = 0$, then their effects are greatest when p_n is large. In contrast, the effect of an event in the beta-model is smallest when p_n is near zero and unity and greatest when p_n is near 0.5; these statements are true whether the event tends to increase or decrease p_n . The sobering fact is that in more than forty years of study of learning curves and learning a decision has not been reached between these two fundamentally different conceptions.

There is one exception to the rule that no model has the property that the increment or decrement induced in p by an event is a constant. In the middle range of probabilities the effects vary only slightly in many models,

and Mosteller (1955) has suggested an additive-increment model to serve as an approximation for this range. The transitions are of the form $\mathbf{p}_{n+1} = Q_k \mathbf{p}_n = \mathbf{p}_n + \delta_k$, where δ_k is a small quantity that may be either positive or negative. This model is not only path-independent and commutative, it is also " p_n -independent."

The foregoing discussion is restricted to path-independent models. In other models the magnitude of the effect of an event depends on other variables in addition to the p -value.

2.8 Simplicity

In model construction appeal is occasionally made to a criterion of simplicity. Because this criterion is always ambiguous and sometimes misleading, it must be viewed with caution: simplicity in one respect may carry with it complexity in another. The relevant attributes of the model are the form of its expressions and the number of variables they contain. Linear forms are thought to be simpler than nonlinear forms (and are approximations to them), which suggests that models with linear operators are simpler than those whose operators are nonlinear. Path-independent models have recursive expressions containing fewer variables than those in path-dependent models, and so they may be thought to be simpler. Classification becomes difficult, however, when other aspects of the models are considered, as a few examples will show.

If we consider the explicit formula, our perspective changes. Commutativity is more fruitful of simplicity than path independence. A conflict arises when we find in the context of an urn (or additive response-strength) model that we can have one only at the price of losing the other (see Sec. 2.6). Also, in such a model even the path-independence criterion taken alone is somewhat ambiguous: one must choose between path independence of the numbers of balls added and path independence of probability changes. Among models with more than one event, the greatest reduction of the number of variables in the explicit formula is achieved by sacrificing both commutativity and path independence, as illustrated by the one-trial perseveration model (Sec. 2.6).

To avoid complicated constraints on the parameters, it appears that nonlinear operators are needed. On the other hand, by using the complicated logistic function, we are assured of the existence of simple *sufficient statistics* for the parameters.

These complications in the simplicity argument are unfortunate: they suggest that simplicity may be an elusive criterion by which to judge models.

3. DETERMINISTIC AND CONTINUOUS APPROXIMATIONS

The models we have been dealing with are, in a sense, doubly stochastic. Knowledge of starting conditions and parameters is not only insufficient to allow one to predict the future response sequence exactly, but, in general, it does not allow exact prediction of the future behavior of the underlying probability or response-strength variable. Even if all subjects, identical in initial probability and other parameter values, behave exactly in accordance with the model, the population is characterized by a distribution of p -values on every trial after the first. For a single subject both the sequence of responses and the sequence of p -values are governed by probability laws.

The variability of behavior in most learning experiments is undeniable, and probably few investigators have ever hoped to develop a mathematical representation that would describe response sequences exactly. Early students of the learning curve, such as Thurstone (1930), acknowledged behavioral variability in the stochastic basis of their models. This basis is obscured by the deterministic learning curve equations which they derived, but these investigators realized that the curves could apply only to the average behavior of a large number of subjects. Stimulus-response theorists, such as Hull, have dealt somewhat differently with the variability problem. In such theories the course of change of the underlying response-strength variable (effective reaction potential) is governed deterministically by the starting values and parameters. Variability is introduced through a randomly fluctuating error term which, in combination with the underlying variable, governs behavior.

Although the stochastic aspect of the learning process has therefore usually been acknowledged, it is only in the developments of the last decade or so that its full implications have been investigated and that probability laws have been thought to apply to the aftereffects of a trial as well as to the performed response.¹⁷

A second distinguishing feature of recent work is its exact treatment of the discrete character of many learning experiments. This renders the models consistent with the trial-to-trial changes of which learning experiments consist. In early work the discrete trials variable was replaced by a

¹⁷ This change parallels comparable developments in the mathematical study of epidemics and population growth. For discussions of deterministic and stochastic treatments of these phenomena, see Bailey (1955) on epidemics and Kendall (1949) on population growth.

continuous time variable, and the change from one trial to the next was averaged over a unit change in time. The difference equations, representing a discrete process, were thereby approximated by differential equations. The differential equation approximations mentioned in Sec. 2.7 are examples.

Roughly speaking, then, a good deal of early work can be thought of as dealing in an approximate way with processes that have been treated more exactly in recent years. Usually the exact treatment is more difficult, and modern investigators are sometimes forced to make continuous or deterministic approximations of a discrete stochastic process. Occasionally these approximations lead to expressions for the average learning curve, for example, that agree exactly with the stochastic process mean obtained by more difficult methods, but sometimes the approximations are considerably in error. In general, the stochastic treatment of a model allows a greater richness of implications to be drawn from it.

It is probably a mistake to think of deterministic and stochastic treatments of a stochastic model as dichotomous. Deterministic approximations can be made at various stages in the analysis of a model by assuming that the probability distribution of some quantity is concentrated at its mean. A few examples will illustrate ways in which approximations can be made and may also help to clarify the stochastic-deterministic distinction.

3.1 Approximations for an Urn Model

In Sec. 2.6 I considered a special case of the general urn scheme, one that has been applied to the shuttlebox experiment. A few approximations will be demonstrated that are in the spirit of Thurstone's (1930) work with urn schemes. Red balls, whose number, r , is constant, are associated with escape; white balls, whose number, w_n , increases, are associated with avoidance. $\text{Pr}\{\text{escape on trial } n\} = p_n = r/(r + w_n)$. An avoidance trial increases w_n by an amount b ; an escape trial results in an increase of c balls. Therefore, if D_k represents an operator that acts on w_n ,

$$w_{n+1} = \begin{cases} D_1 w_n = w_n + b & \text{with probability } 1 - p_n = \frac{w_n}{r + w_n} \\ D_2 w_n = w_n + c & \text{with probability } p_n = \frac{r}{r + w_n} \end{cases} \quad (33)$$

Consider a large population of organisms that behave in accordance with the model and have common values of r , w_1 , b , and c . On the first trial all subjects have the same probability $p_1 = p_1$ of escape. Some will escape and the rest will avoid. If $b \neq c$, there will be two subsets on the

second trial, one for which $w_2 = w_1 + b$ and another for which $w_2 = w_1 + c$. Each of these subsets will divide again on the second trial, but because of commutativity there will be three, not four, distinct values of w_3 : $w_1 + 2b$, $w_1 + b + c$, and $w_1 + 2c$. Each distinct value of w_n corresponds to a distinct p -value. On every trial after the first there is a distribution of p -values.

With two events there are, in general, 2^{n-1} distinct p -values on the n th trial, each corresponding to a distinct sequence of events on the preceding $n - 1$ trials. If the events commute, as in this case, then the number of distinct p -values is reduced to n , the trial number.

Our problem for the urn model is to determine the mean probability of escape on the n th trial, the average being taken over the population.

We let $1 \leq \nu \leq n$ be the index for the n subsets with distinct p -values on trial n . Let $P_{\nu n}$ be the proportion of subjects in the ν th subset on trial n and let $p_{\nu n}$ be the p -value for this subset. Then the m th raw moment of the distribution on trial n is defined by

$$V_{m,n} = E(\mathbf{p}_n^m) = \sum_{\nu} p_{\nu n}^m P_{\nu n}. \tag{34}$$

We use this definition later.

Because w_n , but not \mathbf{p}_n , is transformed linearly, it is convenient to determine $E(w_n) = \bar{w}_n$ first. The increment in numbers of white balls, $\Delta w_n = w_{n+1} - w_n$, is either b or c , and its conditional expectation, conditional on the value of w_n , is given by

$$E_b(\Delta w_n | w_n) = b \left(\frac{w_n}{r + w_n} \right) + c \left(\frac{r}{r + w_n} \right) = \frac{b w_n + cr}{w_n + r}, \tag{35}$$

where E_b denotes the operation of averaging over the binomial distribution of the increment. The unconditional expectation of the increment is obtained by averaging Eq. 35 over the distribution of w_n -values. Using the expectation operator E_w to represent this averaging process, we have

$$E(\Delta w_n) = E_w E_b(\Delta w_n | w_n) = E_w \left(\frac{b w_n + cr}{w_n + r} \right). \tag{36}$$

Note that the right-hand member of this expression is not in general expressible as a simple function of \bar{w}_n .

Now we perform two steps of deterministic approximation. First, we replace Δw_n , which has a binomial distribution, by its average value. From Eq. 35 the increment in w_n (conditional on the value of w_n) can then be written

$$\Delta w_n \simeq \bar{\Delta w}_n = \frac{b w_n + cr}{w_n + r}. \tag{37}$$

Second, we act as if the distribution of w_n is entirely concentrated at its mean value \bar{w}_n . The expectation of the ratio in Eq. 36 is then the ratio itself, and we have

$$\Delta w_n \simeq \bar{\Delta} \bar{w}_n = \frac{b\bar{w}_n + cr}{\bar{w}_n + r}. \quad (38)$$

These two steps accomplish what Bush and Mosteller (1955) call the *expected-operator approximation*. In this method, the change in the distribution of p -values (or w -values) on a trial is represented as the mean p -value (or w -value) acted on by an "average" operator (that is, subject to an average change). Two approximations are involved: the first replaces a distribution of quantities by its mean value and the second replaces a distribution of changes by the mean change. The average operator \bar{D} is revealed in this example if we rewrite Eq. 38 as

$$\bar{w}_{n+1} \simeq \bar{w}_n + \bar{\Delta} \bar{w}_n \equiv \bar{D} \bar{w}_n = \bar{w}_n + b \left(\frac{\bar{w}_n}{\bar{w}_n + r} \right) + c \left(\frac{r}{\bar{w}_n + r} \right)$$

and compare it to Eq. 33. The increments b and c are weighted by their approximate probabilities of being applied. In general,

$$V_{1,n} = E_w \left(\frac{r}{r + w_n} \right) \neq \left(\frac{r}{r + E_w(w_n)} \right) = \frac{r}{r + \bar{w}_n}.$$

But notice that our approximation, which assumes that every w_n -value is equal to \bar{w}_n , leads to this simple relationship.

The discrete stochastic process given by the urn scheme has surely been transformed by virtue of the approximations—but transformed into what? There are at least two interpretations. The first is that the approximate process defines a determined sequence of approximate probabilities for a subject. Like the original process the approximation is stochastic, but on only one "level": the response sequence is governed by probability laws but the response probability sequence is not. According to this interpretation, the approximate model is not deterministic, but it is "more deterministic" than the original urn scheme.

The second interpretation is that the approximate process defines a determined sequence of proportions of white balls, \bar{w}_n , for a population of subjects, and thereby defines a determined sequence of proportions of correct responses, that is, the mean learning curve. According to this interpretation the approximate model is deterministic and it applies only to groups of subjects.

However we think of the approximate model, it is defined by means of a nonlinear difference equation for w_n (Eq. 38). Solution of such equations is difficult and a continuous approximation is helpful. We assume that the trials variable n is continuous and that the growth of \bar{w}_n is gradual

rather than step-by-step. The approximate difference equation given by Eq. 38 can thus itself be approximated by a differential equation:

$$\frac{dw}{dn} = \frac{bw + cr}{w + r}. \quad (39)$$

Integration gives a relation between w and n and therefore between $V_{1,n}$ and n . For the special case of $b = c$ and $w_1 = 0$ the relation has the simple form $w/c = n - 1$, giving

$$V_{1,n} = \frac{r}{r + (n - 1)c}. \quad (40)$$

Equation 40 is an example of an approximation that is also an exact result. In this example it occurs for an uninteresting reason: equating the values of b and c transforms the urn scheme into a single-event model in which the approximating assumption, namely, that all p -values are concentrated at their mean, is correct.

3.2 More on the Expected-Operator Approximation

The expected-operator approximation is important because results obtained by this method are, unfortunately, the only ones known for certain models. Because the approximation also generates a more deterministic model, it is discussed here rather than in Sec. 5 on methods for the analysis of models.

Suppose that a model is characterized by a set $\{Q_k\}$ of operators on response probabilities, where Q_k is applied on trial n with probability $P_k = P_k(p_n)$. This discussion is confined to path-independent models, and therefore P_k can be thought of as a function of at most the p -value on the trial in question and fixed parameters. Because the p -value on a trial may have a distribution over subjects, the probability P_k may also have a probability distribution. The expected operator \bar{Q} is defined by the conditional expectation

$$\bar{Q}\mathbf{p}_n = E_k(Q_k\mathbf{p}_n \mid \mathbf{p}_n) = \sum_k P_k(\mathbf{p}_n)Q_k\mathbf{p}_n. \quad (41)$$

The expectation operator E_k represents the operation of averaging over values of k . The first deterministic approximation in the expected-operator method is the assumption that the same operator—the expected operator—is applied on every trial. Therefore $\mathbf{p}_{n+1} \simeq \bar{Q}\mathbf{p}_n$ for all n .

What is of interest is the average probability on the $(n + 1)$ st trial. This can be obtained by removing the condition on the expectation in Eq. 41 by averaging again, this time over the distribution of p_n -values.

Symbolizing this averaging process by E_p , we have $V_{1,n+1} \simeq E_p(\bar{Q}\mathbf{p}_n)$. The second approximation is to replace $E_p(\bar{Q}\mathbf{p}_n)$ by $\bar{Q}[E_p(\mathbf{p}_n)] = \bar{Q}(V_{1,n})$. This approximation is equivalent to the (deterministic) assumption that the \mathbf{p}_n -distribution is concentrated at its mean. In cases in which $\bar{Q}\mathbf{p}$ is linear in \mathbf{p} , however, $E_p[\bar{Q}\mathbf{p}] = \bar{Q}[E_p(\mathbf{p})]$ is exact and therefore no assumption is needed. (In nonlinear models the method does not seem to give exact results. Of course, it is just for these models that exact methods are difficult to apply.) Applying the second approximation to Eq. 41, we get

$$V_{1,n+1} \simeq \bar{Q}V_{1,n} = \sum_k P_k(V_{1,n})Q_k V_{1,n} \quad (42)$$

as an approximate recursive formula for the mean of the p -value distribution on the n th trial.

EXPECTED OPERATOR FOR TWO EXPERIMENTER-CONTROLLED EVENTS. Consider the model in Sec. 2.4 for the prediction experiment. The operators and the rules for their application are given by Eq. 12:

$$P_{n+1} = \begin{cases} Q_1 p_n = \alpha p_n + 1 - \alpha & \text{if } y_n = 1 \\ Q_2 p_n = \alpha p_n & \text{if } y_n = 0. \end{cases}$$

Recall that the y_n -sequence, and thus the sequence of operators, can be predetermined. Therefore the probability p_n is known exactly from Eq. 14. Often the event sequence is generated by a random device, and, as mentioned in Sec. 2.4, an approximation for p_n can be developed by assuming that the subject experiences the average of all the sequences that the random device generates.

Because the subject has, in fact, experienced one particular event sequence, the approximation may be a poor one. An alternative interpretation of the approximation is that, like many deterministic models, it applies to the average behavior of a large group of subjects. This interpretation is reasonable only if the event sequences are independently generated for each of a large number of subjects. In many experiments to which the approximation has been applied this proviso has unfortunately not been met.

Suppose that the $\{y_n\}$ are independent and that $\Pr\{y_n = 1\} = \pi$ and $\Pr\{y_n = 0\} = 1 - \pi$. Then $P_1(p) = \pi$ and $P_2(p) = 1 - \pi$ are independent of the p -value, as is always the case with experimenter control. Equation 42 gives the recursive relation

$$V_{1,n+1} = \alpha V_{1,n} + (1 - \alpha)\pi. \quad (43)$$

Unlike Eq. 38 for the urn model, Eq. 43 is easily solved and a continuous approximation is not necessary. The solution has already been given in

Sec. 2.4 for the repeated application of the same linear operator. The approximate learning curve equation is

$$V_{1,n} = \alpha^{n-1}V_{1,1} + (1 - \alpha^{n-1})\pi = \pi - \alpha^{n-1}(\pi - V_{1,1}). \quad (44)$$

In this example the result of the approximation is exact in a certain sense: if we average the explicit equation for the model, Eq. 14, over the binomial event distributions, then the result obtained for $V_{1,n}$ will be the same as that given by Eq. 44.

EXPECTED OPERATOR AND THE ASYMPTOTIC PROBABILITY FOR EXPERIMENTER-SUBJECT EVENTS. If one is reluctant to assume for the prediction experiment that reward and nonreward have identical effects, then changes in response probability may depend on the response performed, and the model events are under experimenter-subject control. Let us assume that the outcomes are independent and that $\Pr \{O_j = O_1\} = \pi, \Pr \{O_j = O_2\} = 1 - \pi$. If we assume response-symmetry and outcome-symmetry and use the symbols given in Table 1, the appropriate Bush-Mosteller model is given as follows:

| Event | Operator, Q_k | $P_k(\mathbf{p}_n)$ | |
|------------|---|-------------------------------|------|
| A_1, O_1 | $Q_1\mathbf{p}_n = \alpha_1\mathbf{p}_n + 1 - \alpha_1$ | $\mathbf{p}_n\pi$ | (45) |
| A_2, O_1 | $Q_2\mathbf{p}_n = \alpha_2\mathbf{p}_n + 1 - \alpha_2$ | $(1 - \mathbf{p}_n)\pi$ | |
| A_1, O_2 | $Q_3\mathbf{p}_n = \alpha_2\mathbf{p}_n$ | $\mathbf{p}_n(1 - \pi)$ | |
| A_2, O_2 | $Q_4\mathbf{p}_n = \alpha_1\mathbf{p}_n$ | $(1 - \mathbf{p}_n)(1 - \pi)$ | |

The expected operator approximation (Eq. 42) gives

$$V_{1,n+1} \simeq (1 - \alpha_2)\pi + [\pi + (\alpha_1 - \alpha_2)(1 - 2\pi)]V_{1,n} + (\alpha_1 - \alpha_2)(1 - 2\pi)V_{1,n}^2. \quad (46)$$

This quadratic difference equation is difficult to solve, and Bush and Mosteller approximate it by a differential equation which can then be integrated to give an approximate value for $V_{1,n}$.

The continuous approximation is not necessary if we confine our attention to the asymptotic behavior of the process. At the asymptote the moments of the p -value distribution are no longer subject to change, and therefore $V_{1,n+1} = V_{1,n} = V_{1,\infty}$. Using these substitutions in Eq. 46, we get a quadratic equation whose solution is

$$V_{1,\infty} \simeq \frac{2\pi(1 - \gamma) - 1 + \sqrt{2(\pi - 1)^2 + 4\pi(1 - \pi)\gamma^2}}{2(2\pi - 1)(1 - \gamma)}, \quad (47)$$

where $\gamma = (1 - \alpha_2)/(1 - \alpha_1) \neq 1$ (Bush & Mosteller, 1955, p. 289). For the model defined by Eq. 45 no expression for the asymptotic proportion

of A_1 responses is known other than the approximation of Eq. 47. There is little evidence concerning its accuracy. Our ignorance about this model is especially unfortunate because of the considerable recent interest in asymptotic behavior in the prediction experiment.

Several conclusions about the use of the expected-operator approximation are illustrated in Figs. 1 and 2. Each figure shows average proportions of A_1 responses for 20 artificial subjects behaving in accordance with the model of Eq. 45. All 20 subjects in each group experienced the same event sequence. For both sets of subjects, $\alpha_1 = 0.90$, $\alpha_2 = 0.95$, and $p_1 = V_{1,1} = 0.50$. Reward, then, had twice the effect of nonreward. For the subjects of Fig. 1, $\pi = 0.9$; for those of Fig. 2, $\pi = 0.6$. In both examples the expected operator estimate for the asymptote (Eq. 47) seems too high. Also shown in each figure are exact and approximate (Eq. 44)

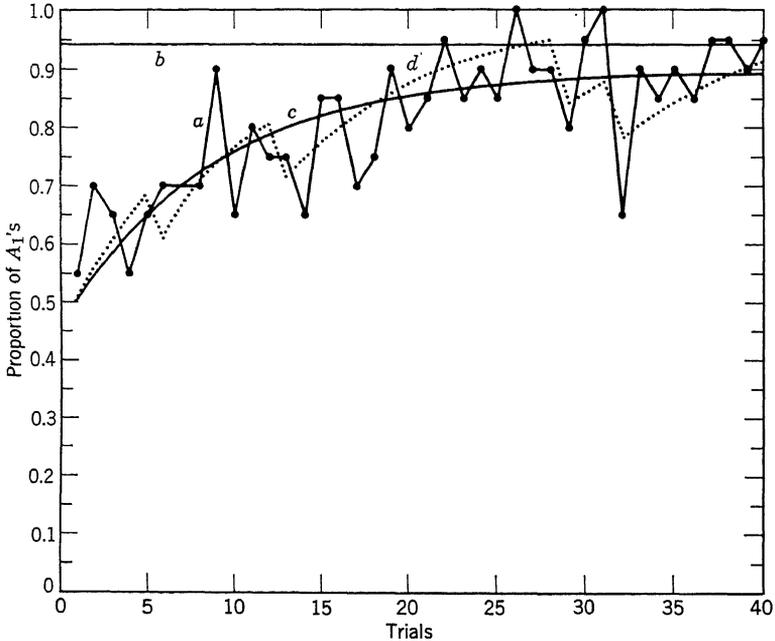


Fig. 1. *a*. The jagged solid line gives the mean proportion of A_1 responses of 20 stat-organisms behaving in accordance with the four-event model with experimenter-subject control (Eq. 45) with $\alpha_1 = 0.90$, $\alpha_2 = 0.95$, $p_1 = 0.50$ and $\pi = 0.90$. *b*. The horizontal line gives the expected-operator approximation of the four-event model asymptote (Eq. 47). *c*. The smooth curve gives the approximate learning curve (Eq. 44) for the two-event model with experimenter control (Eq. 12) with $\hat{\alpha} = 0.8915$ estimated from stat-organism "data," and $p_1 = 0.50$. *d*. The dotted line gives the exact learning curve (Eq. 14) for the two-event model.

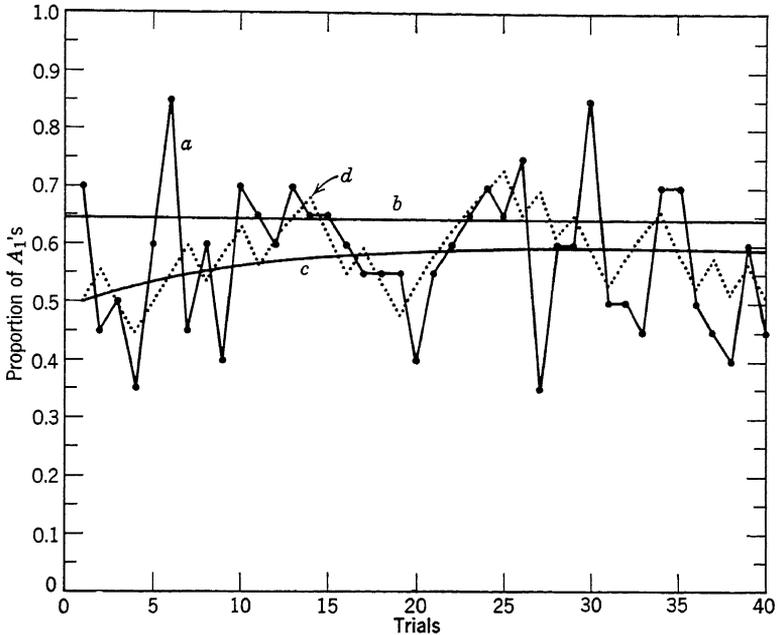


Fig. 2. *a*. The jagged solid line gives the mean proportion of A_1 responses of 20 stat-organisms behaving in accordance with the four-event model with experimenter-subject control (Eq. 45) with $\alpha_1 = 0.90$, $\alpha_2 = 0.95$, $p_1 = 0.50$, and $\pi = 0.60$. *b*. The horizontal line gives the expected-operator approximation of the four-event model asymptote (Eq. 47). *c*. The smooth curve gives the approximate learning curve (Eq. 44) for the two-event model with experimenter-control (Eq. 12) with $\hat{\alpha} = 0.8960$ estimated from stat-organism "data," and $p_1 = 0.50$. *d*. The dotted line gives the exact learning curve (Eq. 14) for the two-event model.

learning curves of the model with two experimenter-controlled events, which has been fitted to the data. The superiority of the exact curve is evident. I shall discuss later the interesting fact that even though the data were generated by a model (Eq. 45) in which reward had more effect than nonreward a model that assumes equal effects (Eq. 12) produces learning curves that are in "good" agreement with the data.

3.3 Deterministic Approximations for a Model of Operant Conditioning

Examples of approximations that transform a discrete stochastic model into a continuous and completely deterministic model are to be found in

treatments of operant conditioning (Estes, 1950, 1959; Bush & Mosteller, 1951). To demonstrate the flavor of these treatments and the approximations used, a sketch of a model along the lines of Estes' is given. In applying a choice-experiment analysis to a free-operant situation, each interresponse period is thought of as a sequence of short intervals of constant length h . It is these intervals that are identified as "trials." During each interval the subject chooses either to press (A_1) or not to press (A_2) the lever; pressing occurs with some probability and is rewarded. The probability is assumed to be unchanged by trials (intervals) on which A_2 occurs and increased by trials (intervals) on which A_1 occurs. The problem is to describe the resulting sequence of interresponse times. To define the model completely it is necessary to consider the way in which $\Pr\{A_1\}$ is increased by reward. We do this in the context of an urn scheme.

An urn contains x white balls (which correspond to A_1) and $b - x$ red balls (which correspond to A_2). At the beginning of a trial a sample of balls is randomly selected from the urn. Each ball has the same fixed probability of being included in the sample, which is of size s . The proportion of white balls in the sample defines the probability $\mathbf{p} = \Pr\{A_1\}$ for the trial in question. At the end of an interval in which A_2 occurs the sample of balls is retained and used to define \mathbf{p} for the interval that follows. At the end of an interval in which A_1 occurs all the red balls in the sample [there are $s(1 - \mathbf{p})$ of them] are replaced by white balls and the sample is returned to the urn. The number of trials (intervals of length h) from one press to the next, including the one on which the lever is pressed, is \mathbf{m} . The interresponse time thus defined is $\tau = \mathbf{m}h$.

The deterministic approximations are as follows:

1. The sample size s is binomially distributed. It is replaced by its mean \bar{s} .

2. Conditional on the value of s , the proportion of white balls \mathbf{p} is binomially distributed. It is replaced by its mean x/b .

3. Conditional on the value of \mathbf{p} , the number of intervals \mathbf{m} is distributed geometrically, with $\Pr\{\mathbf{m} = m\} = \mathbf{p}(1 - \mathbf{p})^{m-1}$. It is replaced by its mean, $1/\mathbf{p}$. By combining the other approximations with this one, we can approximate the number of intervals in the interresponse period by $\mathbf{m} \simeq b/x$ and therefore the interresponse time is approximated by $\tau \simeq hb/x$.

4. Finally, $s(1 - \mathbf{p})$, which is the increase in x (the number of white balls in the urn) that results from reward, is replaced by the product of the means of s and $1 - \mathbf{p}$, and becomes $\bar{s}(b - x)/b$.

The result of this series of approximations is a deterministic process. Given a starting value of x/b and a value for the mean sample size \bar{s} , the

approximate model generates a determined sequence of increasing values of x and of decreasing latencies.

The final approximation is a continuous one. The discrete variables x and τ are considered to be continuous functions of time, $x(t)$ and $\tau(t)$, and $n = n(t)$ is the cumulative number of lever presses. A first integration of an approximate differential equation gives the rate of lever pressing as a (continuous) function of time; integration of this rate gives $n(t)$.

Little work has been done on this model or its variants in a stochastic form. We therefore have little knowledge as to which features of the stochastic process are obscured by the deterministic approximations. One feature that definitely is obscured depends on sampling fluctuations of the proportion of white balls \mathbf{p} . When \mathbf{p} has a high value, then the interresponse time will tend to be short and the increment in x small; when the value of \mathbf{p} happens to be low, then the interresponse time will tend to be above its mean and the increment in x large. One consequence is that interresponse times constitute a dependent sequence such that the variance of the cumulated interresponse time will be less than the sum of the variances of its components.

4. CLASSIFICATION AND THEORETICAL COMPARISON OF MODELS

A good deal of work has been devoted to the mathematical analysis of various model types, but less attention has been paid to the development of systematic criteria by which to characterize or compare models. What are the important features that distinguish one model from another? More pertinent, in what aspects or statistics of the data do we expect these features to be reflected? The need to answer these questions arises primarily in comparative and "baseline" applications of models to data.

Comparative studies, in which we seek to determine which of several models is most appropriate for a set of data, require us to discover *discriminating statistics* of the data: these are statistics that are sensitive to the important differences among the models and that should therefore help us to select one of several models as best.

Once a model is selected as superior, the statistician may be satisfied but the psychologist is not; the data presumably have certain properties that are responsible for the model's superiority, properties that the psychologist wants to know about.

Finally, a model is occasionally used as a baseline against which data are compared in order to discover where the discrepancies lie. Again, a study is incomplete if it leads simply to a list of agreeing and disagreeing statistics;

what is needed as well is an interpretation of these results that suggests which of the model's features seem to characterize the data and which do not.

Analysis of the distinctive features of model types and how they are reflected in properties of the data is useful in the discovery of discriminating statistics, in the interpretation of a model's superiority to others, and in the interpretation of points of agreement and disagreement between a model and data. Some of the important features of several models were indicated in passing as the models were introduced in Sec. 2. A few examples of more systematic methods of comparison are given in this section. Where possible, they are illustrated by reference to one of the comparative studies that have been performed on the Solomon-Wynne shuttlebox data (Bush & Mosteller, 1959; Bush, Galanter, & Luce, 1959), on the Goodnow two-armed bandit data (Sternberg, 1959b), and on some T-maze data (Galanter & Bush, 1959; Bush, Galanter, & Luce, 1959).

4.1 Comparison by Transformation of the Explicit Formula

A comparable form of expression can be used for all of the path-independent commutative-operator models that were introduced in Sec. 2. By suitably defining new parameters in terms of the old, we can write the explicit formula for \mathbf{p}_n as a function of an expression that is linear in the components of \mathbf{W}_n . (Recall that \mathbf{W}_n is the vector whose t components give the cumulative number of occurrences of events E_1, \dots, E_t prior to the n th trial). Suppose $\mathbf{W}_n = (\mathbf{t}_n, \mathbf{s}_n)$, as in the shuttlebox experiment, where \mathbf{t}_n is the total number of avoidances and \mathbf{s}_n the total number of shocks before the n th trial. Let \mathbf{p}_n be the probability of error (nonavoidance) which decreases as \mathbf{s}_n and \mathbf{t}_n increase. Then for the linear-operator model (Eq. 11) we have

$$\mathbf{p}_n = \exp [-(a + b\mathbf{t}_n + c\mathbf{s}_n)],$$

and thus

$$\log \mathbf{p}_n = -(a + b\mathbf{t}_n + c\mathbf{s}_n). \quad (48)$$

For Luce's beta model (Eq. 20)

$$\mathbf{p}_n = \frac{1}{1 + \exp(a + b\mathbf{t}_n + c\mathbf{s}_n)},$$

and thus

$$\text{logit } \mathbf{p}_n = -(a + b\mathbf{t}_n + c\mathbf{s}_n). \quad (49)$$

For the special case of the urn scheme (Eq. 26)

$$\mathbf{p}_n = \frac{1}{a + b\mathbf{t}_n + c\mathbf{s}_n},$$

and thus

$$\frac{1}{p_n} = a + bt_n + cs_n. \tag{50}$$

Finally, for Mosteller's additive-increment-approximation model (Sec. 2.7), applied to the two-event experiment,

$$p_n = -(a + bt_n + cs_n), \tag{51}$$

For each model some transformation, $g(p)$, of the response probability is a linear function of t_n and s_n . The models differ only in the transformations they specify.

The behavior of models for which this type of expression is possible can be described by a simple nomogram. Examples for the first three models above are given in Fig. 3, in which the transformation $x = dg(p) + e$ is plotted for each model (d and e are constants). The units on the abscissa are arbitrary. To facilitate comparisons, the coefficients d and e are

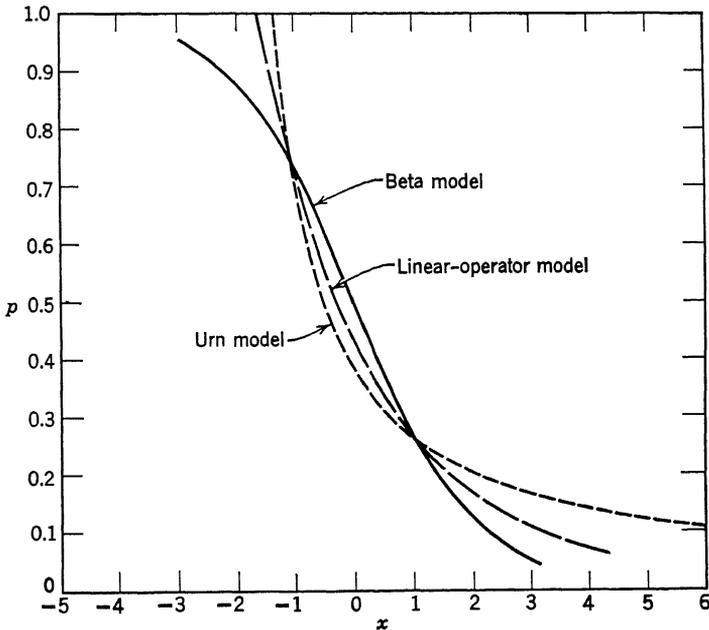


Fig. 3. Nomograms for three models. The linear-operator model (Eq. 48) is represented by $\log_e p = -0.500x + 0.837$. The beta model (Eq. 49) is represented by $\text{logit } p = x$. The urn scheme (Eq. 50) is represented by $p^{-1} = 1.214x + 2.667$. Constants were so chosen that curves would agree at $p = 0.25$ and $p = 0.75$. The units on the abscissa are arbitrary.

chosen so that values of x agree at $p = 0.25$ and $p = 0.75$. The additive-increment model is represented by a straight line passing through the two common points.

The nomogram is interpreted as follows: a subject's state (the value of the linear expression $a + bt_n + cs_n$) is represented by a point on the abscissa, and his error probability is given by the corresponding point on the p -axis. The occurrence of an event corresponds to a displacement along the abscissa whose magnitude depends only on the event and not on the starting position. For this example all displacements are to the right and correspond to reductions in the error probability. An avoidance corresponds to a displacement b units to the right, and a shock to a displacement c units to the right. If events have equal effects, then we have a single-event model, and each trial corresponds to the same displacement.

Although a displacement along the abscissa is a constant for a given event, the corresponding displacement on the probability axis depends on the slope of the curve at that point. Because the slope depends on the p -value (except for the additive-increment model), the probability change corresponding to an event depends on that value. Because the probability change induced by an event depends only on the p -value, this type of nomogram is limited to path-independent models. Its use is also limited to models in which the operators commute. For the additive-increment, path-independent urn, and beta models it can be used when there are more than two events. For these models events that increase p_n correspond to displacements to the left.

A number of significant features of the models can be seen immediately from Fig. 3. First, the figure indicates that in the range of probabilities from 0.2 to 0.8 the additive-increment model approximates each of the other three models fairly well. This supports Mosteller's (1955) suggestion that estimates for the additive model based on data from this range be used to answer simple questions such as which of two events has the bigger effect. Caution should be exercised, however, in applying the model to data from a subsequence of trials that begins after trial 1. Even if we had reason to believe that all subjects have the same p -value on trial 1, we would probably be unwilling to assume that the probabilities on the first trial of the subsequence are equal from subject to subject. Therefore the estimation method used should not require us to make this assumption.

A second feature disclosed by Fig. 3 concerns the rate of learning (rate of change of p_n) at the early stages of learning. When p_n is near unity, events in the urn and linear operator models have their maximum effects. In contrast, the beta model requires that p_n change slowly when it is near unity. If the error probability is to be reduced from its initial value to, let us say, 0.75 in a given number of early trials, then for the beta model

to accomplish this reduction it must start at a lower initial value than the other models or its early events must correspond to larger displacements along the x -axis than they do in the other models. Early events tend predominantly to be errors, and therefore the second alternative corresponds to a large error-effect.

A third feature has to do with the behavior of the models at low values of p_n . The models differ in the rates at which the error probability approaches zero. Especially notable is the urn model which, after a translation, is of the form $p = x^{-1}$ ($x \geq 1$). Unlike the situation in the other models, the area under the curve of p versus x diverges, so that in an unlimited sequence of trials we expect an unlimited number of errors. (The expected number of trials between successive errors increases as the trial number increases but not rapidly enough to maintain the total number of errors at a finite value). The urn model, then, tends to produce more errors than the other models at late stages in learning.

This analysis of the three models helps us to understand some of the results that have been obtained in applications to the Solomon-Wynne avoidance-learning data. The analyses have assumed common values over subjects for the initial probability and other parameters. The relevant results are as follows:

1. A "model-free" analysis, which makes only assumptions that are common to the three models, shows that an avoidance response leads to a greater reduction in the escape probability than an escape response. This analysis is described in Sec. 6.7.

2. The best available estimate of p_1 in this experiment is 0.997 and is based on 331 trials, mainly pretest trials (Bush & Mosteller, 1955).

3. The linear-operator model is in good agreement with the data in every way that they have been compared (Bush & Mosteller, 1959). The estimates are $\hat{p}_1 = 1.00$, $\hat{\alpha}_1$ (avoidance parameter) = 0.80, and $\hat{\alpha}_2$ (escape parameter) = 0.92. According to the parameter values, for which approximately the same estimates are obtained by several methods (Bush & Mosteller, 1955), escape is less effective than avoidance in reducing the escape probability.

4. There is one large discrepancy between the urn model and the data. Twenty-five trials were examined for each of 30 subjects. The last escape response occurred, on the average, on trial 12. The comparable figure for the urn model is trial 20. As in the linear-operator model, the estimates suggest that escape is less potent than avoidance (Bush & Mosteller, 1959).

5. One set of estimates for the beta model is given by $\hat{p}_1 = 0.94$, $\hat{\beta}_1$ (avoidance parameter) = 0.83, and $\hat{\beta}_2$ (escape parameter) = 0.59 (Bush, Galanter, & Luce, 1959). With these estimates, the model differs from the

data in several respects, notably producing underestimates of intersubject variances of several quantities, such as total number of escapes. As discussed later, this probably occurs because the relative effects of avoidance and escape trials are incorrectly represented by the model.

6. The approximate maximum-likelihood estimates for the beta model are given by $\hat{p}_1 = 0.86$, $\hat{\beta}_1 = 0.74$, and $\hat{\beta}_2 = 0.81$. In contrast to the inference made by Bush, Galanter, and Luce, these estimates imply that escape is less effective than avoidance. The estimate of the initial probability of escape is lower than theirs, however.

These results strongly favor the linear-operator model. The results of analysis of the data with the other models are intelligible in the light of our study of the nomogram. The first set of estimates for the beta model gives an absurdly low value of p_1 . In addition, the relative magnitudes of escape and avoidance effects are reversed. The second set of estimates, which avoids attributing to error trials an undue share of the learning, underestimates p_1 by an even greater amount. Apparently, if the beta model is required to account for other features of the data as well, it cannot describe the rapidity of learning on the early trials of the experiment. The major discrepancy between the urn model and data is in accord with the exceptional behavior of that model at low p -values; the average trial of the last error is a discriminating statistic when the urn model is compared to others.

Before we leave this type of analysis, it is instructive to consider Hull's model (1943) in the same way. This model was discussed briefly in Sec. 2.5. It is intended to describe the change in probability of reactions of the all-or-none type, such as conditioned eyelid responses and barpressing in a discrete-trial experiment. If we assume that incentive and drive conditions are constant from trial to trial, then the model involves the assumptions that (1) reaction potential (${}_sE_R$) is a growth function of the number of reinforcements and (2) reaction probability (q) is a (normal) ogival function of the difference between the reaction potential and its threshold (${}_sL_R$) when this difference is positive; otherwise the probability is zero.

The assumptions can be stated formally as follows:

$$1. \quad {}_sE_R = M(1 - e^{-AN}), \quad (A, M > 0). \quad (52)$$

The quantity N is defined to be the number of reinforcements, not the total number of trials. Unrewarded trials correspond to the application of an identity operator to ${}_sE_R$.

$$2. \quad \text{logit } q = \begin{cases} B + C({}_sE_R - {}_sL_R), & ({}_sE_R > {}_sL_R) \\ -\infty, & ({}_sE_R \leqslant {}_sL_R). \end{cases} \quad (53)$$

For uniformity and ease of expression the logistic function is substituted for the normal ogive. When these equations are combined, we obtain for the probability p of not performing the response

$$\log(\text{logit } p + c) = \begin{cases} -(a + bN), & (N > k) \\ -\infty, & (N \leq k). \end{cases} \quad (54)$$

This is to be compared to Eq. 48 and Eq. 49.

From Hull's figures a rough estimate of $c = 5$ is obtained. This allows us to construct a nomogram, again choosing constants so that the curve agrees with the other models at $p = 0.25$ and $p = 0.75$. The result is displayed in Fig. 4, with nomograms for the linear and beta models. The curve for Hull's model falls in between those of the other two. The existence of a threshold makes the model difficult to handle mathematically, but, in contrast to the beta model, it allows learning to occur in a finite number of trials even with an initial error-probability of unity.

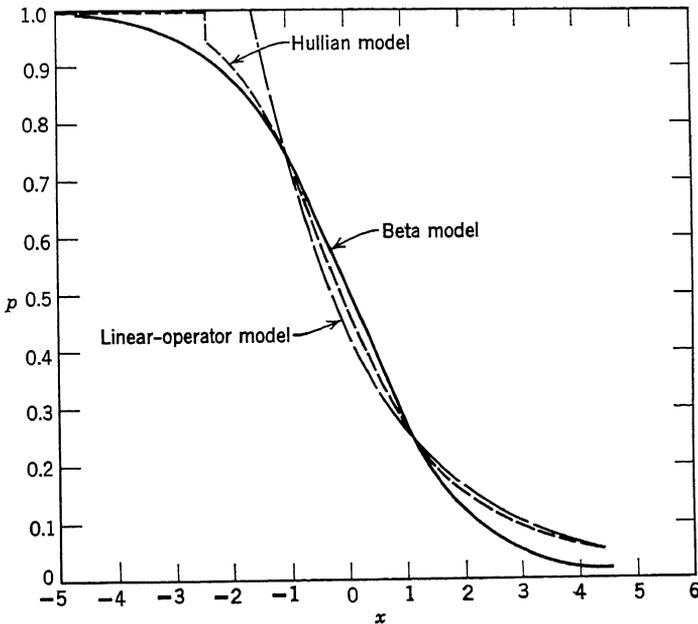


Fig. 4. Nomogram for Hullian model (Eq. 48) and two other models. The Hullian model is represented by $\log_e(\text{logit } p + 5) = -0.204x + 1.585$. The nomograms for linear-operator and beta models are those presented in Fig. 3.

4.2 Note on the Classification of Operators and Recursive Formulas

The classification of operators and recursive formulas is probably more relevant to the choice of mathematical methods for the analysis of models than it is to the direct appreciation of their properties. (Two exceptions considered later are the implications of commutativity and of the relative magnitudes of the effects of rewarded and nonrewarded trials.) The classification described here is based on the arguments that appear in the recursive formula. Let us consider models with two subject-controlled events, in which $\mathbf{x}_n = 1$ if E_2 occurs on trial n , $\mathbf{x}_n = 0$ if E_1 occurs on trial n , and $\mathbf{p}_n = \Pr\{\mathbf{x}_n = 1\}$. A rough classification is given by the following list:

1. $p_{n+1} = f(p_n)$. Response-independent, path-independent. *Example:* single-operator model (Eq. 28).
2. $p_{n+1} = f(n, p_n)$. Response-independent, quasi-independent of path. *Example:* urn scheme (Eq. 25) with equal event-effects.

Classes 1 and 2 produce sequences of independent trials and, if there are no individual differences in initial probabilities and other parameters, they do not lead to distributions of p -values.

3. $\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{x}_n)$. Response-dependent, path-independent. *Example:* linear commutative operator model (Eq. 8).
4. $\mathbf{p}_{n+1} = f(n, \mathbf{p}_n; \mathbf{x}_n)$. Response-dependent, quasi-independent of path. *Example:* general urn scheme (Eq. 25).
5. $\mathbf{p}_{n+1} = f(\mathbf{p}_n; \mathbf{x}_n, \mathbf{x}_{n-1})$. Path-dependent. *Example:* one-trial perseveration model (Eq. 30).

4.3 Implications of Commutativity for Responsiveness and Asymptotic Behavior

In Sec. 2.2 I pointed out that in a model with commutative events there is no "forgetting": the effect of an event on \mathbf{p}_n is the same whether it occurred on trial 1 or on trial $n - 1$. The result is that models with commutative events tend to respond sluggishly to changes in the experiment.

As an example to illustrate this phenomenon we take the prediction

experiment and, for the moment, consider it as a case of experimenter-controlled events. We use the linear model (Eq. 12) to illustrate non-commutative events and the beta model (Eq. 21) to illustrate commutative events. The explicit formulas are revealing. The quantity d_n is defined as before as the number of left-light outcomes less the number of right-light outcomes, cumulated through trial $n - 1$. The beta model is then represented by Eq. 23 which is reproduced here:

$$p_n = \frac{1}{1 + \beta^{d_n} v_1^{-1}}.$$

In this model all trials with equal d_n -values also have equal p_n -values. The response probability can be returned to its initial value simply by introducing a sequence of trials that brings d_n back to its initial value of zero. The response of the model to successive reversals is illustrated in Fig. 5 with the event sequence $E_1 E_1 E_1 E_1 E_2 E_2 E_2 E_2 E_1 E_1$. Despite the fact that on the ninth trial, on which d_9 is zero, the most recent outcomes have been right-light onsets, the probability of predicting the left light is no lower than it was initially.

The behavior of the commutative model is in contrast to that of the linear model, whose explicit formula (Eq. 14) is reproduced here:

$$p_n = \alpha^{n-1} p_1 + (1 - \alpha) \sum_{j=1}^{n-1} \alpha^{n-1-j} y_j.$$

The formula shows that when $\alpha < 1$ more recent events are weighted more heavily and that equal d_n -values do not in general imply equal probabilities. The response of this model to successive reversals is also illustrated in Fig. 5. Parameters were chosen so that the two models would agree on the first and fifth trials. This model is more responsive to the reversal than the beta model; not only does the curve of probability versus trials change more rapidly, but its direction of curvature is also altered by the reversal.

At first glance the implications of commutativity for responsiveness of a model seem to suggest crucial experiments or discriminating statistics that would allow an easy selection to be made among models. The question is more complicated, however. The contrast shown in Fig. 5 is clear-cut only if we are willing to make the dubious assumption that events in the prediction experiment are experimenter-controlled. Matters become complicated if we allow reward and nonreward to have different effects. The relative effectiveness of reward and nonreward trials is then another factor that determines the responsiveness of a model.

To show this, we shift attention to models with experimenter-subject control of events. To make the conditions extreme, we compare equal-parameter models (experimenter-control) with models in which the

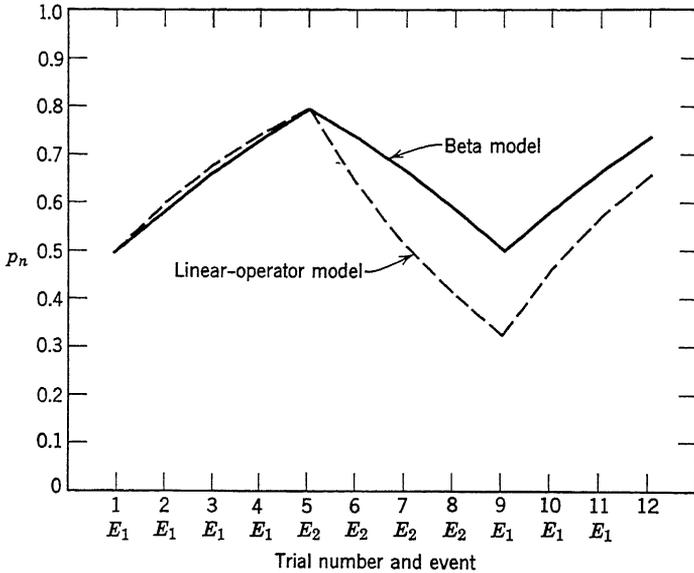


Fig. 5. Comparison of models with commutative and noncommutative experimenter-controlled events. The broken curve represents p_n for the linear-operator model (Eq. 12) with $\alpha = 0.8$ and $p_1 = 0.5$. The continuous curve represents p_n for the beta model with $\beta = 0.713$ and $p_1 = 0.5$. Parameter values were selected so that curves would coincide at trials 1 and 5.

identity operator is associated with either reward or nonreward. The results are shown in Figs. 6 and 7. Parameter values are chosen so that the models agree approximately on the value of $V_{1,5}$. In Fig. 6 the “equal alpha” linear model (Eq. 12) with $\alpha = 0.76$ is compared with the two models defined in (55):

| Response | Outcome | Model with Identity Operator for Nonreward ($\alpha = 0.60$) | Model with Identity Operator for Reward ($\alpha = 0.40$) | |
|----------|---------|--|---|------|
| A_1 | O_1 | $p_{n+1} = p_n + 1 - \alpha$ | $p_{n+1} = p_n$ | (55) |
| A_2 | O_1 | $p_{n+1} = p_n$ | $p_{n+1} = \alpha p_n + 1 - \alpha$ | |
| A_1 | O_2 | $p_{n+1} = p_n$ | $p_{n+1} = \alpha p_n$ | |
| A_2 | O_2 | $p_{n+1} = \alpha p_n$ | $p_{n+1} = p_n$ | |

In Fig. 7 the “equal beta” model (Eq. 21) with $\beta = 0.68$ is compared with the two models defined in (56):

| Response | Outcome | Model with Identity Operator for Nonreward ($\beta = 0.50$) | Model with Identity Operator for Reward ($\beta = 0.30$) |
|----------|---------|---|--|
| A_1 | O_1 | $p_{n+1} = \frac{\beta p_n}{(1 - p_n) + \beta p_n}$ | $p_{n+1} = p_n$ |
| A_2 | O_1 | $p_{n+1} = p_n$ | $p_{n+1} = \frac{\beta p_n}{(1 - p_n) + \beta p_n}$ |
| A_1 | O_2 | $p_{n+1} = p_n$ | $p_{n+1} = \frac{p_n}{\beta(1 - p_n) + p_n}$ |
| A_2 | O_2 | $p_{n+1} = \frac{p_n}{\beta(1 - p_n) + p_n}$ | $p_{n+1} = p_n$ |

(56)

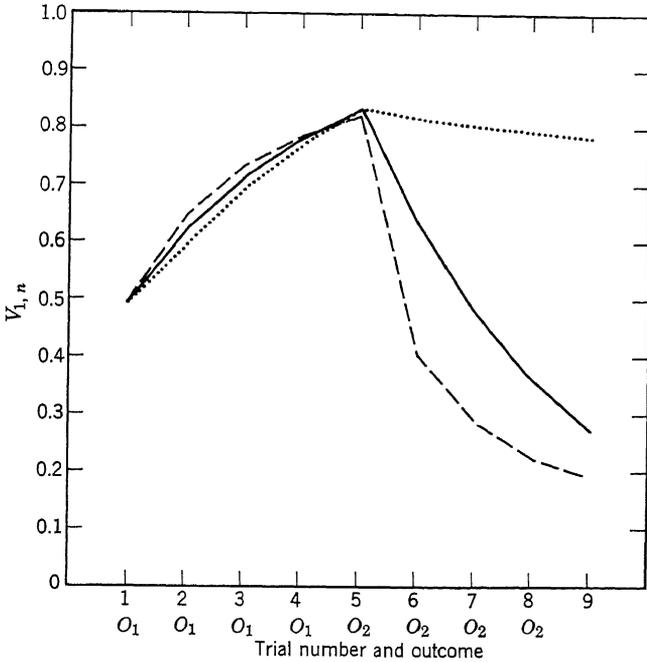


Fig. 6. Responsiveness of the linear-operator model (Eq. 55) depends on the relative effectiveness of reward and nonreward. The solid curve represents $V_{1,n}$ for the linear-operator model with equal reward and nonreward parameters ($\alpha_1 = \alpha_2 = 0.757, p_1 = 0.5$). The broken curve represents $V_{1,n}$ for the linear-operator model with an identity operator for reward ($\alpha_1 = 1.0, \alpha_2 = 0.4, p_1 = 0.5$). The dotted curve represents $V_{1,n}$ for the linear-operator model with an identity operator for nonreward ($\alpha_1 = 0.6, \alpha_2 = 1.0, p_1 = 0.5$). Parameter values were selected so that the models would agree approximately on the values of $V_{1,1}$ and $V_{1,5}$.

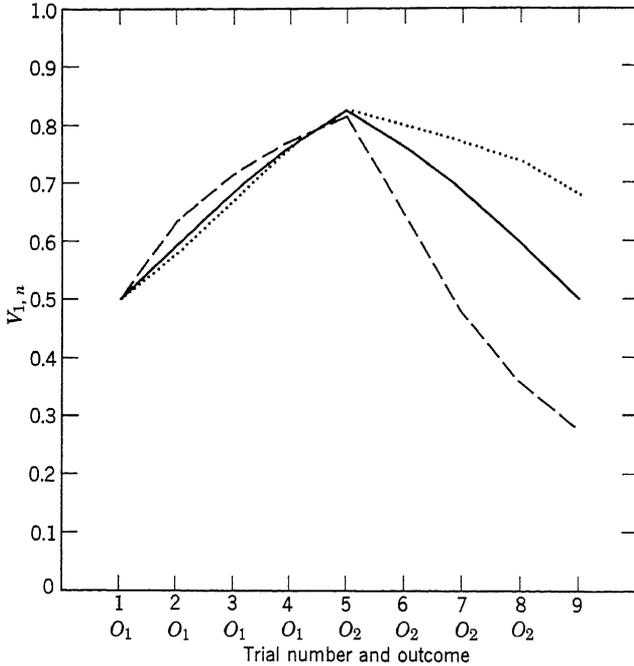


Fig. 7. Responsiveness of the beta model (Eq. 56) depends on the relative effectiveness of reward and nonreward. The solid curve represents $V_{1,n}$ for the beta model with equal reward and nonreward parameters ($\beta_1 = \beta_2 = 0.68, p_1 = 0.5$). The broken curve represents $V_{1,n}$ for the beta model with an identity operator for reward ($\beta_1 = 1.0, \beta_2 = 0.3, p_1 = 0.5$). The dotted curve represents $V_{1,n}$ for the beta model with an identity operator for nonreward ($\beta_1 = 0.5, \beta_2 = 1.0, p_1 = 0.5$). Parameter values were selected so that the models would agree approximately on the values of $V_{1,1}$ and $V_{1,5}$.

Roughly the same pattern appears for both models. When nonreward is less effective than reward, the response to a change in the outcome sequence that leads to a higher probability of nonreward is sluggish. When reward is less effective than nonreward, the response is rapid. From these examples it appears that the influence on responsiveness of changing the relative effects of reward and nonreward is less marked in the commutative-operator beta model than in the linear model.

Responsiveness alone, then, is not useful in helping us to choose between the models. We must use it in conjunction with knowledge about the

relative effects of reward and nonreward. This situation is typical in working with models: observation of a single aspect of the data is often insufficient to lead to a decision. If, in this example, we observe that subjects' behavior is highly responsive, this might imply that a model with commutative operators is inappropriate, but, alternatively, it might mean that the effect of nonreward is relatively great. Also, if we examined such data under the hypothesis that events in the prediction experiment are experimenter-controlled, then the increased rate of change of $V_{1,n}$ after the reversal would probably lead us to conclude, perhaps in error, that a change in the value of a learning rate parameter had occurred. This example indicates how delicate are the conclusions one draws regarding event invariance (Sec. 2.7) and illustrates how the apparent failure of event invariance may signify that the wrong model has been applied.

In a number of studies the prediction experiment has been analyzed by the experimenter-controlled event model of Eq. 12 (Estes & Straughan, 1954; Bush & Mosteller, 1955). This model also arises from Estes' stimulus sampling theory. One of the findings that has troubled model builders is that estimates of the learning-rate parameter α tend to vary systematically from experiment to experiment as a function of the outcome probabilities. It is not known why this occurs, but the phenomenon has occasionally been interpreted as indicating that event effects are not invariant as desired. Another interpretation, which has not been investigated, is that because reward and nonreward have different (but possibly invariant) effects the estimate of a single learning-rate parameter is, in effect, a weighted average of reward and nonreward parameters. Variation of outcome probabilities alters the relative number of reward-trials and thus influences the weights given to reward and nonreward effects in the over-all estimate. The estimation method typically used depends on the responsiveness of the model, which, as we have seen, depends on the extent to which rewarded trials predominate.

4.4 Commutativity and the Asymptote in Prediction Experiments

One result of two-choice prediction experiments that has interested many investigators is that when $\Pr \{y_n = 1\} = \pi$ and $\Pr \{y_n = 0\} = 1 - \pi$ then for some experimental conditions the asymptotic mean probability $V_{1,\infty}$ with which human subjects predict $y_n = 1$ appears to "match"

$\Pr \{y_n = 1\}$, that is, $V_{1,\infty} \simeq \pi$.¹⁸ (The artificial data in Figs. 1 and 2 illustrate this phenomenon.) The phenomenon raises the question of which model types or model families are capable of mimicking it. No answer even approaching completeness seems to have been proposed, but a little is known. Certain linear-operator models are included in the class, and we shall see that models with commutative events can be excluded, at least when the events are assumed to be experimenter-controlled and symmetric. [Feldman and Newell (1961) have defined a family of models more general than the Bush-Mosteller model that displays the matching phenomenon.]

Figures 1 and 2 suggest that a linear-operator model with experimenter-subject control can approximate the matching effect. As already mentioned, an exact expression for the asymptotic mean of this model is not known. It is easy to demonstrate that the linear model with experimenter-controlled events can produce the effect exactly; indeed, a number of investigators have derived confidence in the adequacy of this particular model from the "probability matching" phenomenon (e.g., Estes, 1959; Bush & Mosteller, 1955, Chapter 13). The value of $V_{1,\infty}$ for the experimenter-controlled model when the $\{y_n\}$ are independent binomial random variables can easily be determined from its explicit formula (Eq. 14), which is reproduced here:

$$p_n = \alpha^{n-1} p_1 + (1 - \alpha) \sum_{j=1}^{n-1} \alpha^{n-1-j} y_j.$$

We take the expectation of both sides of this equation with respect to the independent binomial distributions of the $\{y_j\}$, making use of the fact that $E(y_j) = \pi$. Performing the summation, we obtain

$$V_{1,n} = \alpha^{n-1} p_1 + (1 - \alpha^{n-1})\pi. \quad (57)$$

Note that this is the same result given by the expected-operator approximation in Eq. 44. The final result, $V_{1,\infty} = \pi$, is obtained by letting $n \rightarrow \infty$.

As an example of a model with experimenter-controlled events that cannot produce the effect, we consider Luce's model (Eq. 23), with

¹⁸ The validity of this finding and the particular conditions that lead to it have been the subjects of considerable controversy. Partial bibliographies may be found in Edwards (1956, 1961), Estes (1962), and Feldman & Newell (1961). The reader should also consult Restle (1961, Chapter 6) and, for work with several nonhuman species, the papers of Bush & Wilson (1956) and of Bitterman and his colleagues (e.g., Behrend & Bitterman, 1961). The general conclusions to be drawn are that the phenomenon does not occur under all conditions or for all species, that when it seems to occur the response probability may deviate slightly but systematically from the outcome probability, that matching may characterize a group average although it occurs for only a few of the individuals within the group, and that an asymptote may not have been reached in many experiments.

$0 < \beta < 1$. We restrict our attention to experiments in which $\pi \neq \frac{1}{2}$; without loss of generality we can restrict it further to $\pi > \frac{1}{2}$. Because \mathbf{p}_n is governed entirely by the value of \mathbf{d}_n , we must concern ourselves with the behavior of

$$\mathbf{d}_n = \sum_{j=1}^{n-1} (2y_j - 1).$$

Roughly speaking, because the number of left-light outcomes (E_1) increases faster than the number of right-light outcomes (E_2), the difference between their numbers increases, and with an unlimited number of trials this difference \mathbf{d}_n becomes indefinitely large. More precisely, we note that

$$E(\mathbf{d}_n) = \sum_{j=1}^{n-1} (2\pi - 1)$$

and that therefore $E(\mathbf{d}_n) \rightarrow \infty$ as $n \rightarrow \infty$. From the law of large numbers (Feller, 1957, Chapter X) we conclude that with probability one $\mathbf{d}_n \rightarrow \infty$ as $n \rightarrow \infty$. Using Eq. 23, it follows that for this model $\mathbf{p}_n \rightarrow 1$ when $\pi > 0.50$.

The asymptotic properties of other examples of the beta model for the prediction experiment have been studied by Luce (1959) and Lamperti and Suppes (1960). They find that there are special conditions, determined by the values of π and model parameters, under which $V_{1,n} = E(\mathbf{p}_n)$ does not approach a limiting value of either zero or unity. Therefore it should not be inferred from the foregoing example that the beta model is incapable of producing probability matching. (In view of the fact that the phenomenon does not occur regularly or in all species, one might consider a model that invariably produces it to be more suspect than one in which its occurrence depends on conditions or parameter values.)

As an illustration of the present state of knowledge, we consider the beta model with experimenter-subject control for the prediction experiment. In this model, absorption at a limiting probability of zero or unity does not always occur. The outcomes are $O_1 : y = 1$ and $O_2 : y = 0$. The responses are A_1 : predict O_1 , and A_2 : predict O_2 . We assume that the pairs of events, $\{A_1O_1, A_2O_2\}$ and $\{A_1O_2, A_2O_1\}$ are complementary. The transformations of the response-strength ratio $v = v(1)/v(2)$ are therefore as follows:

| Event | Transformation |
|----------|------------------------------------|
| A_1O_1 | $v \rightarrow \beta v$ |
| A_2O_1 | $v \rightarrow \beta' v$ |
| A_1O_2 | $v \rightarrow \frac{1}{\beta'} v$ |
| A_2O_2 | $v \rightarrow \frac{1}{\beta} v$ |

The parameters β and β' correspond to reward and nonreward, respectively; both are greater than one.

For this model the results of Lamperti and Suppes (1960, Theorem 3) imply that the asymptotic value of $\mathbf{p}_n = \Pr \{A_1 \text{ on trial } n\}$ is either zero or unity *except* when the following inequality is satisfied:

$$\frac{\log \beta}{\log \beta'} < \frac{\pi}{1 - \pi} < \frac{\log \beta'}{\log \beta}.$$

Luce has shown (1959, Chapter 4, Theorem 17) that when the inequality is satisfied the asymptotic value of $V_{1,n}$ is given by

$$V_{1,\infty} = \pi + \frac{2\pi - 1}{(\log \beta')/(\log \beta) - 1}.$$

(It is interesting to note that the value of $V_{1,\infty}$ for the corresponding linear-operator model, given by Eq. 47, is known only approximately.)

From these results several conclusions may be drawn. First, if $\beta > \beta'$, then this model *always* produces asymptotic absorption at zero or one; only if nonreward is more potent than reward ($\beta' > \beta$) is a limiting average probability other than zero or one possible. Second, for a fixed pair of parameter values, $\beta' > \beta \geq 1$, absorption at zero or one can be avoided, but only for a limited range of π -values. Third, when $V_{1,\infty}$ is between zero and one, it is equal to π only if $\pi = \frac{1}{2}$; otherwise the asymptote is further from $\frac{1}{2}$ than π is and in the same direction, with the magnitude of the "overshoot" or "undershoot" increasing linearly with $|\pi - \frac{1}{2}|$.

In the experimenter-controlled-events example, which was first discussed, it is the commutativity of the beta model that is responsible for its asymptotic behavior. An informal argument shows that the same asymptotic behavior characterizes any model with two events that are complementary, commutative, and experimenter-controlled and in which repeated occurrence of a particular event leads to a limiting probability of zero or unity. In any such model the response probability returns to its initial value on any trial on which $\mathbf{d}_n = 0$. Moreover, the probability on any trial is invariant under changes in the order of the events that precede that trial. Therefore the probability \mathbf{p}_n after a mixed sequence composed of m E_2 's and $(n - m - 1)$ E_1 's is the same as the value of \mathbf{p}_n after a block of $(n - m - 1)$ E_1 's preceded by a block of m E_2 's. Now let $\mathbf{m}(n)$ be an integral random variable whose value is the number of E_2 -events in n trials. Note that $E[\mathbf{m}(n)] = n(1 - \pi)$. For $\pi > \frac{1}{2}$ we have already seen that as n increases we have $n - \mathbf{m}(n) - 1 > \mathbf{m}(n)$ with probability one. Consider what happens when the order of the events is rearranged so that a block of all the E_2 's precedes a block containing all the E_1 's. On the

m th trial of the second block the probability returns to its initial value. After this trial there are $[n - 2\mathbf{m}(n) - 1]$ E_1 -trials; but

$$E[n - 2\mathbf{m}(n) - 1] = n(2\pi - 1) - 1,$$

which becomes indefinitely large. The behavior of the model is the same as if, starting at the initial probability, an indefinitely long sequence of E_1 -trials occurred. The limiting value of \mathbf{p}_n is therefore unity. A similar result applies when the event-effects have different magnitudes. Without further calculations we know that the urn scheme of Eq. 25 cannot mimic the matching effect.

I have discussed the asymptotic behavior of these models partly because it is of interest in itself but mainly to emphasize the strong implications of the commutativity property. As a final example of the absence of "forgetting," let us consider an experiment in which first a block of E_1 's occurs and then a series in which E_1 's and E_2 's occur independently with probability $\pi = \frac{1}{2}$. In both the beta and linear models for complementary experimenter-controlled events the initial block of E_1 -events will increase the probability to some value, say p' . In the linear-operator model the mixed event series will reduce the probability from p' toward $p = \frac{1}{2}$. In the beta model, on the other hand, the mixed event series will cause the probability to fluctuate indefinitely about p' with, on the average, no decrement. The last statement is true for any model whose events are complementary, experimenter-controlled, and commutative.

4.5 Analysis of the Explicit Formula¹⁹

In this section I consider some of the important features of explicit formulas for models with two subject-controlled events. These models are meant to apply to experiments such as the escape-avoidance shuttlebox and 100:0 prediction, bandit, and T-maze experiments. The event (response) on trial n is represented by the value of \mathbf{x}_n , where $\mathbf{x}_n = 0$ if the rewarded response is made and $\mathbf{x}_n = 1$ if the nonrewarded response (an "error") is made. The probability $\mathbf{p}_n = \Pr \{ \mathbf{x}_n = 1 \}$ decreases over the sequence of trials toward a limiting value of $p = 0$.

EXAMPLES USED. The following models are used as examples:

Model I $p_n = F(n) = \alpha^{n-1}p_1, \quad (0 < \alpha < 1); \quad (58)$

Model II $\mathbf{p}_n = F(n, \mathbf{x}_{n-1})$
 $= \alpha^{n-1}p_1(1 - \beta) + \beta\mathbf{x}_{n-1}, \quad (0 < \beta < 1, n \geq 2); \quad (59)$

¹⁹ Much of this discussion is drawn from Sternberg (1959b).

$$\begin{aligned} \text{Model III } \mathbf{p}_n &= F(n, \mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \dots, \mathbf{x}_1) \\ &= \alpha^{n-1} p_1 + \beta \sum_{j=1}^{n-1} \alpha^{n-1-j} \mathbf{x}_j, \quad (0 < \alpha, \beta < 1); \end{aligned} \quad (60)$$

$$\text{Model IV } \mathbf{p}_n = F(n, \mathbf{s}_n) = \exp [-(a + bn + cs_n)], \quad (0 < a, b). \quad (61)$$

We have seen Models I, II, and IV before. Model I is the single-operator model of Eq. 28 (Bush & Sternberg, 1959) and is an example of the family of single-event models. Model II is the one-trial perseveration model of Eq. 29. (An analogous nonlinear model is given by the generalized logistic in Eq. 32.) Model IV is the Bush-Mosteller model of Eqs. 10 and 11, rewritten by using the fact that $\mathbf{t}_n = n - 1 - \mathbf{s}_n$. The quantity $\mathbf{s}_n = \sum_{j=1}^{n-1} \mathbf{x}_j$ is the number of errors before trial n , and $c = -\log(\alpha_2/\alpha_1)$. If the effect of reward is greater than the effect of nonreward ($\alpha_1 < \alpha_2$), then $c < 0$ and more errors (larger \mathbf{s}_n) imply a higher probability of error (larger \mathbf{p}_n); if $\alpha_1 > \alpha_2$, then $c > 0$ and the converse holds. This model has been studied by Tatsuoka and Mosteller (1959). (Analogous beta and urn models are given by Eqs. 19 and 26.)

Almost all the models that have been applied to data involve either identity operators or operators with limit points of zero or unity. One exception is Model III, whose operators are given by

$$\mathbf{p}_{n+1} = \begin{cases} \alpha \mathbf{p}_n & \text{if } \mathbf{x}_n = 0 \\ \alpha \mathbf{p}_n + \beta & \text{if } \mathbf{x}_n = 1. \end{cases}$$

Referred to as the "many-trial perseveration model," this model has been applied to two-armed bandit data by Sternberg (1959b). The explicit formula is similar in form to Eq. 14 for the linear model for two experimenter-controlled events; more recent events are weighted more heavily. **DIRECT RESPONSE EFFECTS.** Consider first the *direct effect* of a response, \mathbf{x}_j , on the probability \mathbf{p}_n . By "direct effect" is meant the influence of \mathbf{x}_j on the magnitude of \mathbf{p}_n when intervening responses $\mathbf{x}_{j+1}, \dots, \mathbf{x}_{n-1}$ are held fixed. Response \mathbf{x}_j has a direct effect on \mathbf{p}_n if it appears as an argument of the explicit formula F . The effect is *positive* if $\mathbf{x}_j = 1$ results in a larger value of \mathbf{p}_n than does $\mathbf{x}_j = 0$; otherwise the effect of \mathbf{x}_j is *negative*. Models II and III show positive response effects, achieved by associating an additive constant with \mathbf{x}_j in the explicit formula. In Model IV the direct effects can be positive or negative, depending on the sign of c . The effect is achieved by adding a constant to $\log \mathbf{p}_n$ when $\mathbf{x}_j = 1$; this is equivalent to applying a multiplicative constant to \mathbf{p}_n . When

response effects occur in one of these models, they are all of the same sign; the direction of the effect of an event does not depend on when the event occurred. Let us confine our discussion to this type of model.

If none of the x_j appears in F , then there are no response effects and the model is response-independent. This is a characteristic of all single-event models. Model I is an example.

If any of the x_j appear in F , there are direct response effects. If only x_{n-1} appears, then p_n is directly affected only by the immediately preceding response, as in Model II. Because the p_m for $m > n$ are not affected by x_{n-1} we say that the direct effect is *erased* as the process advances. If several, say k , of the x_j appear in F , then the direct effect of a response continues for $k - 1$ trials and is then erased. [Audley and Jonckheere (1956) have considered a special case of their urn scheme that has this property.] If all the x_j ($j = n - 1, n - 2, \dots, 1$) appear in F , the effect of a response is never erased and continues indefinitely. This last condition must hold for any response-dependent model that is also path-independent. Models III and IV are examples.

When more than one x_j appears in F , we can ask two further questions concerned with the way in which the arguments x_j appear in F . The first is whether there is *damping* of the continuing effects. We define the *magnitude* of the effect of x_j on p_n to be the change in the value of p_n when the value of x_j in $F(n, 0, 0, 0, \dots)$ is increased from 0 to 1. When the magnitude of the effect of x_j is smaller for earlier x_j , then we say that direct response effects are damped. If the magnitudes are equal, then the effects are *undamped*. (Direct effects might also be augmented with trials; this could occur in a model in which the full effect of a response took more than one trial to appear. No such models have been studied, however. In what follows we assume that effects are either damped or undamped.)

The second question we can ask, when two or more of the x_j appear in F , is whether their effects *accumulate*. If so, then the effect on p_n when two of the x_j are errors is greater than the effect when either one of them alone is an error. In all of the models mentioned in this chapter for which effects continue they also accumulate.

If a model exhibits damped response effects, the cumulative number of errors alone is not sufficient to tell us the value of p_n ; we must also know on which trials the errors occurred. Therefore, the events in a model with damped effects cannot commute; and, conversely, if a model with commutative events shows response effects, then these effects cannot be damped. Models III and IV provide examples of the foregoing statement and its converse. In Model III the response effects are damped and events do not commute; in Model IV, a commutative event model, response effects are undamped. These two models are analogous to the linear and beta models

for experimenter-controlled events (Sec. 4.3). In the linear model outcome effects are damped (there is "forgetting") and events do not commute; in the beta model there is no damping and we have commutativity.

By means of these ideas models can be roughly ordered in terms of the extent to which direct response effects occur. First is the response-independent, single-event model in which there is no effect at all (Model I). Then we have a model in which an effect occurs but is erased (Model II). Next is a model in which the effect continues but is damped (Model III); and finally we have a model with an undamped, continuing effect (Model IV).

INDIRECT RESPONSE EFFECTS. One of the most important properties of models with subject-control of events is the fact that the responses in a sequence are not independent. This is the property that causes subjects' response probabilities to differ even when they have common parameter values and are run under identical reinforcement schedules. One result is that, in contrast to models in which only the experimenter controls events, we must deal with distributions rather than single values of the response probabilities. A second implication is that events (responses) have indirect as well as direct effects on future responses, effects that are transmitted by the intervening trials. In contrast, experimenter-controlled events have only direct effects.

Until now we have been considering only the direct effect of a response \mathbf{x}_j on \mathbf{p}_n . If $j < n - 1$, so that trials intervene between responses \mathbf{x}_j and \mathbf{x}_n , there also may be *indirect effects* mediated by the intervening responses. For example, whether or not \mathbf{x}_{n-2} has a direct effect on \mathbf{p}_n , it may have an indirect effect, mediated through its direct effect on \mathbf{p}_{n-1} and the relation of \mathbf{p}_{n-1} to the value of \mathbf{x}_{n-1} . Therefore, even if the direct effect of \mathbf{x}_j on \mathbf{p}_n is erased, the response may influence the probability. Model II provides an example. In this model the p -value on a trial is determined uniquely by the trial number and the preceding response, so that, conditional on the value of \mathbf{x}_{n-1} , \mathbf{x}_n is independent of all the \mathbf{x}_m , $m < n - 1$. On the other hand, if the value of \mathbf{x}_{n-1} is not specified, then \mathbf{x}_n depends on any one of the \mathbf{x}_m , $m < n - 1$, that may be selected. Put another way, the conditional probability $\Pr \{ \mathbf{x}_n = 1 \mid \mathbf{x}_{n-1} \}$ is uniquely determined, whatever the $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-2}$ sequence is. But, given any trial at all before the n th, the unconditional ("absolute") probability $\Pr \{ \mathbf{x}_n = 1 \}$ depends on the response that is made on that trial. A more familiar example is the one-step Markov chain, in which the higher-order conditional probabilities are not the same as the corresponding "absolute probabilities" (Feller, 1957), despite the fact that the direct effects extend only over a single trial. Because \mathbf{x}_{n-1} has no effect at all on \mathbf{p}_n in a response-independent model, it cannot have any indirect effect on \mathbf{p}_m ($m > n$).

The *total effect* of a response \mathbf{x}_j on the probability p_n can be represented by the difference between two conditional probabilities:²⁰

$$\Pr \{x_n = 1 \mid \mathbf{x}_j = 1\} - \Pr \{x_n = 1 \mid \mathbf{x}_j = 0\}.$$

When direct effects are positive, the total effect of \mathbf{x}_j on p_n cannot be less than its direct effect alone. The extent to which the total effect is greater depends in part on whether there is accumulation of the direct effects of \mathbf{x}_j and the intervening responses and in part on whether and how effects are damped. When direct effects are negative, the situation is more complicated, and the relation between total and direct effects depends on whether the number of intervening trials is even or odd as well as on accumulation and damping.

SUBJECT-CONTROLLED EVENTS AS A PROCESS WITH FEEDBACK. In most of the foregoing discussion we have been considering the effects of responses on probabilities. The altered probabilities influence their associated responses, and these responses in turn have effects on future probabilities. Thus the effects we have been considering "feed back" the "output" of the p_n -sequence so as to influence that sequence. Insofar as two response sequences have different p_n -values on some trial, the nature of the response effects determines whether this probability difference will be enhanced, maintained, reduced, or reversed in sign on the next trial.

Each of the p_n -sequences produced by a model is an *individual learning curve*, and the "area" under this curve represents the expected number of errors associated with that sequence. In a large population of response sequences (subjects) the model specifies a proportion of the population that will be characterized by each of the possible individual learning curves. The mean learning curve is the average of these individual curves. If there are no response effects, there is, of course, only one individual curve.

When response effects exist and are positive, we may speak of a *positive feedback of probability differences* and determine measures of its magnitude. With more positive feedback of p_n -differences, individual learning curves have a greater tendency to deviate from their mean curve as n increases. The *negative feedback* of p_n -differences, which may occur if response effects exist and are negative, may cause the opposite result: p_n -differences that arise among sequences may be neutralized or reversed in sign. Thus an individual curve that deviated from the mean curve would tend to return to it or to cross it and therefore to compensate for the deviation. A rough idea of the magnitude of the feedback can be obtained by comparing an assumed p_n -difference of Δp_n on trial n with the associated expected difference of $\bar{\Delta p}_{n+1}$ on the next trial.

²⁰ For a binary-event sequence that is generated by a stationary stochastic process this expression gives the autocorrelation function with lag $n - j$.

Also relevant to the feedback question is the range of the p_n -values that a model can produce on a given trial. For example, in a model with positive response effects the maximum possible value of p_n is attained when all the responses have been errors and the minimum is attained when they have all been successes. For a model with negative effects the reverse holds. Therefore the p_n -range is given by the absolute value of

$$F(n, 1, 1, 1, \dots) - F(n, 0, 0, 0, \dots).$$

Whatever the sign or magnitude of any feedback of probability differences that may occur, it cannot lead to p_n -differences larger than the p_n -range. The p_n -values corresponding to the extremes of the p_n -range produce the pair of individual learning curves that differ maximally in area. In general, the p_n -range imposes a limit on all the response effects discussed in this section.

For Model I the p_n -range is zero. For Model II it is a constant. For Models III and IV the range increases with n .

DISCRIMINATING STATISTICS: SEQUENTIAL PROPERTIES. The analysis of response effects presented above is useful, first in suggesting statistics of the data that may discriminate among Models I to IV and second in helping us to interpret the results of applications of these models. To illustrate these uses, let us consider results of Sternberg's (1959b) application of these models to data collected by Goodnow in a two-armed bandit experiment with 100:0 reward.

The analysis tells us that fundamental differences among the four models lie in the extent to which response effects occur and are erased or damped. This suggests that the models differ in their sequential properties and that it is among the sequential features of the data that we should find discriminating statistics. This suggestion is confirmed by the following results that were obtained in application of the models to the Goodnow data:

1. Parameter values can be chosen for all four models so that they produce mean learning curves in good agreement with the observed curve of trial-by-trial proportions of errors. The observed and fitted curves are shown in Fig. 8. Despite the differences among the models, the mean learning curve does not discriminate one from another.

2. Now we begin to examine sequential properties. First we consider the mean number of runs of errors. The parameters in Model I cannot be adjusted so that it will retain its good agreement with the learning curve and at the same time produce few enough runs of errors; this model can be immediately disqualified. In contrast, parameters in Models II, III, and IV can be chosen so that these models will agree with both the learning curve and the number of error runs. (This difference is not altogether

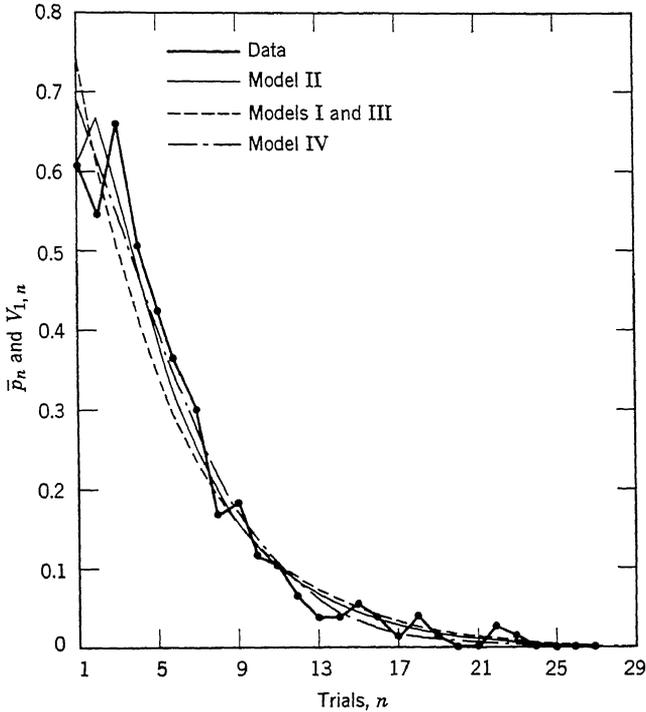


Fig. 8. Observed trial-by-trial proportions \hat{p}_n of errors in the Goodnow experiment and theoretical means $V_{1,n}$ for the four models of Eqs. 58 to 61.

surprising because Model I has one less free parameter than the others.) A finer analysis of error runs, considering the average number of runs of each length, j , $1 \leq j \leq 7$, does not help; Models II, III, and IV produce equally good agreement with the distribution of error-run lengths. This is shown by Fig. 9.

3. Finally, we examine the serial autocovariance of errors at lags 1 to 10. This statistic is defined to be the mean value of

$$c_k = \sum_n x_n x_{n+k}$$

where the lag is given by the value of k . (Models II, III, and IV all agree with the observed value of \bar{c}_1 , but this tells us nothing new, since their agreement follows automatically from agreement with the learning curve and the number of runs.) What is of interest is the behavior of \bar{c}_k as k increases: its observed value falls rapidly, and only Model II is able to produce so rapid a decrease. The slowest decrease is produced by Model IV, with Model III a close second. The results are illustrated in Fig. 10.

Our analysis of the four models in terms of response effects aids the interpretation of these results. Figure 8 suggests, as did Figs. 1 and 2, that interesting and important differences among models and between models and data may be totally obscured if we restrict our attention to the learning curve. The results regarding runs of errors reflect the fundamental difference between Model I, which is response-independent, and the others. Responses in this experiment were clearly not independent of each other; when errors occurred they tended to occur in clusters, suggesting a positive response effect. This suggestion is confirmed by the values of the estimated parameters for the other models. The behavior of \bar{c}_k is sensitive to the extent of erasing or damping the positive response effect. Its value drops most rapidly with k in Model II, in which the effect is erased after a single

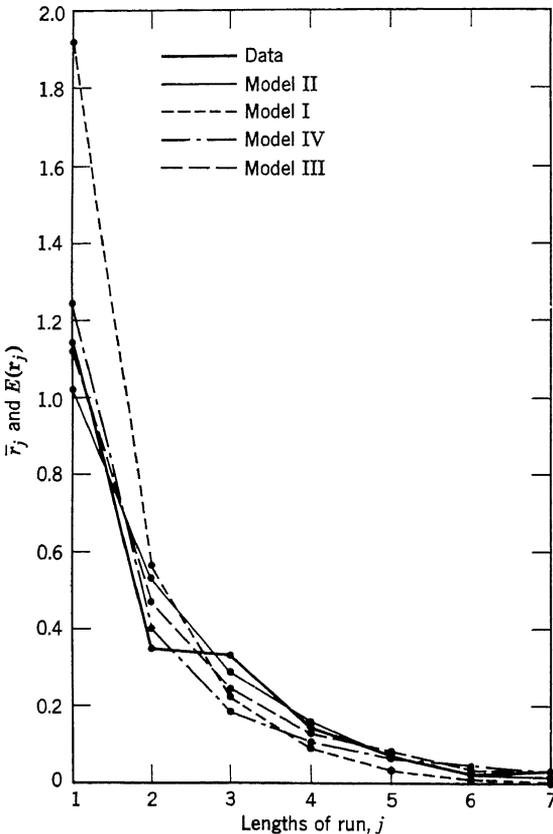


Fig. 9. Observed mean number of error runs \bar{r}_j of length j in the Goodnow experiment and theoretical values $E(\mathbf{r}_j)$ for the four models of Eqs. 58 to 61.

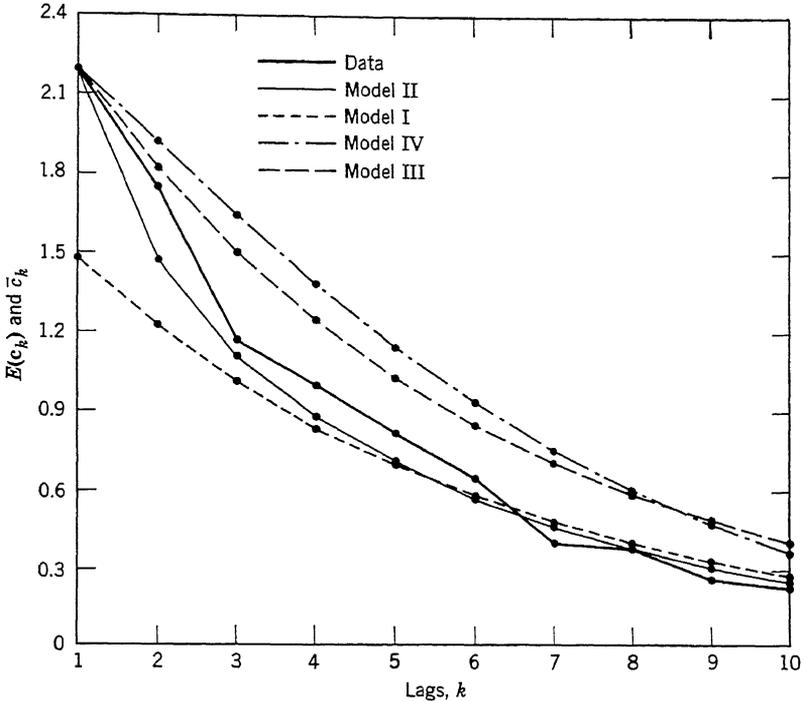


Fig. 10. Observed mean values \bar{c}_k of the serial autocovariance of errors in the Goodnow experiment at lags 1 to 10 and theoretical values $E(c_k)$ for the four models of Eqs. 58 to 61.

trial. Its value drops most slowly in Model IV, in which the effect continues and is undamped. Model III, with a continuing but damped effect, is intermediate.

The interpretation of these results, based on our analysis of response effects, leads us to conclude that Goodnow's data exhibit short-lived, positive response effects. If a model for these data is to reproduce the learning curve, a degree of positive sequential dependence is required for it to match the number of error runs as well. But this response effect must be radically damped or erased if the model is to describe the autocovariance of errors.

DISCRIMINATING STATISTICS: REWARD AND NONREWARD AND THE VARIANCE OF TOTAL ERRORS. One of the questions that has interested investigators concerns the relative magnitudes of the effects of reward and nonreward in situations in which their effects have the same sign. For example, one plausible interpretation of the events in a T-maze with 100:0 reward is that both rewarded and nonrewarded trials increase

the probability of making the rewarded response. Similarly, in the shuttle-box, both the escape and avoidance responses might be thought to increase the probability of avoidance. How would differences in the relative magnitudes of these effects manifest themselves in the data? One answer is suggested if we translate the question into the language of positive and negative response effects. If the effect of reward is the greater, then the error probability after an error is higher than the error probability after a success, and we have a positive response effect. If the effect of nonreward is the greater, a negative response effect exists.

We have already considered the relation between response effects and the feedback of probability differences. When response effects are positive, feedback is positive, individual learning curves tend to diverge from the mean learning curve, and subjects with more early errors tend to have more errors late in learning. When response effects are negative, feedback is negative, differences between individual learning curves tend to be neutralized or reversed, and subjects with more early errors tend to have fewer later errors. Areas under individual learning curves tend to be more variable with positive effects than they are with negative effects.

Because the area under an individual learning curve represents the total number of errors for that individual, this informal argument suggests that there should be a relation between response effects and the variance of the total number of errors. Roughly speaking, if two experiments produce the same average number of total errors, then the experiment in which reward is the more effective should produce a greater variance of total errors than the experiment in which nonreward is the more effective. Moreover, there should be a positive correlation between the extent of positive response effects (magnitude, continuation, damping) and the variance of total errors.

Both conclusions are borne out by studies of experiments and models. One experiment that provides clear evidence of a negative response effect is a study of reversal after overlearning in a T-maze (Galanter & Bush, 1959; Bush, Galanter, & Luce, 1959). (A "model-free" demonstration that the effect is negative is discussed in Sec. 6.7.) In Table 3 this experiment is compared with five others in which analyses have suggested that there is a positive response effect. The coefficient of variation behaves appropriately. The second conclusion is supported by the theoretical values of the variance of total errors for Models I to IV when their parameters are selected to produce the same learning curve and, when possible, the same number of error runs as observed in the Goodnow data. The figures for the variance correspond roughly to the extent of positive response effects: Model I: 2.53; Model II: 4.04; Model III: 10.78; Model IV: 10.42. (The corresponding figure for the data is 5.17.)

Table 3 Positive and Negative Response Effects and the Variance of Total Errors in Several Experiments

| Experiment | Response Effect | Mean Total Errors, \bar{U}_1 | S. D. of Total Errors, $S(U_1)$ | Coefficient of Variation $C = S(U_1)/\bar{U}_1$ |
|---|-----------------|--------------------------------|---------------------------------|---|
| T-Maze reversal after overlearning (rats) (Galanter & Bush, 1959) | Negative | 24.68 | 2.24 | 0.09 |
| T-Maze reversals (rats) (Galanter & Bush, 1959) | | | | |
| Period 2 | Positive | 14.10 | 3.50 | 0.24 |
| Period 3 | Positive | 9.50 | 3.04 | 0.32 |
| Period 4 | Positive | 12.70 | 3.85 | 0.30 |
| Solomon-Wynne Shuttlebox (dogs) (Bush & Mosteller, 1959) | Positive | 7.80 | 2.52 | 0.32 |
| Goodnow two-armed bandit (Sternberg, 1959b) | Positive | 4.32 | 2.27 | 0.53 |

5. MATHEMATICAL METHODS FOR THE ANALYSIS OF MODELS

In the preceding sections I have considered informal and approximate methods of analyzing and comparing models. A good many of the conclusions have been qualitative, and, although we should not belittle their usefulness in guiding research and in aiding the interpretation of results, it is in the quantitative properties of learning models that the core of our knowledge lies. No unified methods of analysis exist, however. Various devices have been used, of which only a few samples are illustrated here. Other examples are contained in numerous references, including Bush & Mosteller (1955), Bush & Estes (1959), Karlin (1953), Lamperti & Suppes (1960), Bush (1960), and Kanal (1962a,b).

In principle, only the investigator's imagination limits the number of different statistics of response sequences, short of individual responses, that his model can be coaxed to describe. Examples, some of which we have already come across, are the mean learning curve, the number of trials before the k th success, the number of runs of errors of a particular length, the autocovariance of errors, the number of occurrences of a

particular outcome-response pair, and the trial on which the last error occurs. The entire distribution for a statistic or, if desired, just its mean and variance can be described by a model. In practice, analytic methods are limited, often severely so, and a good many of these statistics must be estimated from Monte Carlo sampling experiments.

Usually the expectation of a statistic produced by a model is a function of parameter values, and therefore the problem of estimating these values cannot be bypassed in the analysis of data. On the other hand, as we shall see later, the dependence of statistics on parameter values has been exploited a good deal in estimation procedures. Occasionally a model makes a parameter-free prediction; examples are the asymptotes of the beta and linear models that were discussed in Sec. 4.3. When this occurs, we can often dismiss or get favorable evidence for a model type without bothering to narrow it down to a particular model by estimating its parameters.

In addition to their use in estimation, model statistics are, of course, used in evaluating goodness of fit. This is often done by enumerating the statistics thought to be pertinent and comparing their expected values with those observed. The general point of view has been that the observed response sequences constitute a sample from a population of sequences, and the question is which model type describes the population.

5.1 The Monte Carlo Method

The Monte Carlo method is the generation of an artificial realization of a stochastic process by a sampling procedure that satisfies the same probability laws. A random device, usually a table of random numbers, is substituted for the behaving organism and is used to select responses with the appropriate probabilities. Once we have selected values for the initial probability and other parameters of a model we can generate as large a sample of artificial response sequences as we wish. Any feature of these artificial data or "stat-organisms" can be compared with the corresponding feature of the real response sequences. The method is therefore extremely versatile for testing models whose parameters we are able to estimate.²¹

If the outcome sequence in the real experiment is generated by a probability mechanism, as it often is in the prediction experiment, for example, there is a choice between generating new sequences for the Monte Carlo experiment or generating the artificial data conditional on the actual

²¹ The Monte Carlo method is discussed in Bush & Mosteller (1955, Chapter 6), and in Chapters 7 and 8 of Vol. I of this *Handbook*. Examples and references are given by Barucha-Reid (1960, Appendix C).

sequences used in the experiment. In this instance, most workers agree on using conditional Monte Carlo calculations.

If the model involves subject-control of events, then a similar choice is available between letting p_n in the Monte Carlo experiment be determined by the preceding sequence of artificial responses and letting it be determined by the real response sequences. In other words, the value of \mathbf{p}_n is specified by an explicit formula whose arguments consist of parameters and variables that represent the responses on trials 1 through $n - 1$. The responses used can be those in the real data, thus conditioning the Monte Carlo experiment by those data, or artificial responses can be used. Most workers have used artificial data that are not conditioned by the observed responses, but the relative merits of these methods for learning-model research have not yet been assessed.

Sampling experiments are extremely inefficient for handling the estimation problem, and here analytic methods have a considerable practical advantage as well as their usual greater elegance. We turn now to a few examples of analytic methods.

5.2 Indicator Random Variables

We have already used random variables whose values indicate which of two responses or which of two outcomes occurs on a trial. By appropriately choosing the two possible values, generally as zero and unity, many of the statistics in which we are interested can be represented easily by products and sums of these random variables. This type of representation facilitates the calculation of expectations and variances.

Let $\mathbf{x}_n = 1$ if the response on trial n is an error and $\mathbf{x}_n = 0$ if it is a success. Then the number of errors in a sequence of N trials is defined by

$$\mathbf{u}_{1,N} = \sum_{n=1}^N \mathbf{x}_n. \tag{62}$$

When we consider an infinite sequence of trials, we drop the subscript N ; for example, \mathbf{u}_1 denotes the number of errors in an infinite sequence. (This number may or may not be finite.) One approximation often used, when the probability of error approaches zero as n increases, replaces error statistics for finite sequences by their infinite counterparts.

The number $\mathbf{r}_{T,N}$ of runs of errors during N trials is expressed in terms of the $\{\mathbf{x}_j\}$ by noting that every error run, except the one that terminates a sequence, is followed by a success. Therefore,

$$\mathbf{r}_{T,N} = \sum_{n=1}^{N-1} \mathbf{x}_n(1 - \mathbf{x}_{n+1}) + \mathbf{x}_N = \sum_{n=1}^N \mathbf{x}_n - \sum_{n=1}^{N-1} \mathbf{x}_n \mathbf{x}_{n+1}. \tag{63}$$

For an infinite sequence the upper limit of the summations becomes infinite and we use the symbol r_T .

If we define u_j , the number of “ j -tuples” of errors in an infinite sequence, by

$$u_j = \sum_{n=1}^{\infty} \mathbf{x}_n \mathbf{x}_{n+1} \cdots \mathbf{x}_{n+j-1}, \quad (64)$$

then r_k , the number of runs of errors of a particular length k , can be expressed in terms of the $\{u_j\}$ by

$$r_k = u_k - 2u_{k+1} + u_{k+2}$$

(Bush, 1959).

In experiments in which both outcomes and responses may vary, similar expressions can be used for various response-outcome patterns. On trial n for subject i ($i = 1, 2, \dots, I$), let $x_{i,n} = 1$ if response A_1 occurs, $x_{i,n} = 0$ if response A_2 occurs, and let $y_{i,n} = 1$ if outcome O_1 follows, $y_{i,n} = 0$ if outcome O_2 follows. The number of subjects for which O_1 occurred on trial n and A_1 on trial $n + 1$ (a measure of the correlation of responses with prior outcomes) is given by

$$\sum_{i=1}^I y_{i,n} x_{i,n+1}. \quad (65)$$

5.3 Conditional Expectations

Partly because of the “doubly stochastic” nature of most learning models (Sec. 3) in which both the responses and the p -values have probability distributions, it is often convenient when finding the expectation of a statistic to determine first the expectation conditional on, say, the p -value and then to average the result over the distribution of p -values. A few examples will illustrate this use of conditional expectations. We let $\mathbf{p}_n = \Pr \{\mathbf{x}_n = 1\}$. It will be convenient to let $V_{1,k}(p)$ denote the first moment of the p -value distribution of a process that started k trials ago at probability p ; that is,

$$V_{1,k}(p) = \Pr \{\mathbf{x}_{n+k} = 1 \mid \mathbf{p}_{n+1} = p\} = E(\mathbf{x}_{n+k} \mid \mathbf{p}_{n+1} = p). \quad (66)$$

It is also useful to let E_x denote an average over the binomial distribution of responses, E_y an average over a binomial distribution of outcomes, and E_p an average over a p -value distribution. Recall that when the response probability is considered to be a random variable it is written \mathbf{p}_n ; a particular value is p_n .

Suppose we wish to evaluate $E(\mathbf{r}_T)$ for an experiment with subject-controlled events.

$$E(\mathbf{r}_T) = \sum_{n=1}^{\infty} E(\mathbf{x}_n) - \sum_{n=1}^{\infty} E(\mathbf{x}_n \mathbf{x}_{n+1}).$$

$$E(\mathbf{x}_n) = E_p E_x(\mathbf{x}_n \mid \mathbf{p}_n) = E_p(\mathbf{p}_n) = V_{1,n}.$$

$$E(\mathbf{x}_n \mathbf{x}_{n+1}) = E_p E_x(\mathbf{x}_n \mathbf{x}_{n+1} \mid \mathbf{p}_n) = E_p[\mathbf{p}_n \Pr \{ \mathbf{x}_{n+1} = 1 \mid \mathbf{x}_n = 1, \mathbf{p}_n \}].$$

To evaluate the last expression further requires us to specify the model. Consider the commutative linear-operator model that we discussed in connection with the shuttlebox experiment (Eq. 8). Then $\Pr \{ \mathbf{x}_{n+1} = 1 \mid \mathbf{x}_n = 1, \mathbf{p}_n \} = \alpha_2 \mathbf{p}_n$ and therefore

$$E(\mathbf{x}_n \mathbf{x}_{n+1}) = E_p(\alpha_2 \mathbf{p}_n^2) = \alpha_2 V_{2,n},$$

where $V_{2,n}$ is the second (raw) moment of the p -value distribution on trial n . We thus have

$$E(\mathbf{r}_T) = \sum_{n=1}^{\infty} V_{1,n} - \alpha_2 \sum_{n=1}^{\infty} V_{2,n}, \tag{67}$$

and the sums can be evaluated in terms of the model parameters (Bush 1959). For the single-operator model ($\alpha_1 = \alpha_2 = \alpha$), $V_{1,n} = \alpha^{n-1} p_1$ and $V_{2,n} = \alpha^{2(n-1)} p_1^2$, and Eq. 67 gives

$$E(\mathbf{r}_T) = \frac{p_1}{1 - \alpha} - \frac{\alpha p_1^2}{1 - \alpha^2},$$

which function is illustrated, for $p_1 = 1$, in Fig. 12, p. 91.

As a second example, suppose we wish to evaluate the expectation of the statistic

$$\mathbf{c}_k = \sum_{n=1}^{\infty} \mathbf{x}_n \mathbf{x}_{n+k},$$

which we encountered in Sec. 4.5. We have

$$\begin{aligned} E(\mathbf{x}_n \mathbf{x}_{n+k}) &= E_p E_x(\mathbf{x}_n \mathbf{x}_{n+k} \mid \mathbf{p}_n) \\ &= E_p[\mathbf{p}_n \Pr \{ \mathbf{x}_{n+k} = 1 \mid \mathbf{x}_n = 1, \mathbf{p}_n \}] \\ &= E_p(\mathbf{p}_n V_{1,k} [\Pr \{ \mathbf{x}_{n+1} = 1 \mid \mathbf{x}_n = 1, \mathbf{p}_n \}]). \end{aligned}$$

Again we use the commutative linear-operator model as our example. The conditional probability is $\alpha_2 \mathbf{p}_n$ and therefore

$$E(\mathbf{c}_k) = \sum_{n=1}^{\infty} E_p[\mathbf{p}_n V_{1,k} (\alpha_2 \mathbf{p}_n)].$$

Turning to experiments in which outcomes may vary from trial to trial, let us consider the evaluation of the expectation of

$$\mathbf{t} = \frac{1}{NI} \sum_{n=m}^{m+N-1} \sum_{i=1}^I \mathbf{y}_{i,n} \mathbf{x}_{i,n+1},$$

which is the proportion of outcome-response pairs in the indicated block of trials for which A_1 on trial $n + 1$ follows O_1 on trial n . Statistics of this type were considered by Anderson (1959) and are examples of aspects of the data that are of interest even after the average response probability has stabilized. We assume a linear operator model with experimenter control and let $\Pr \{y_{i,n} = 1\} = \pi$.

First let us consider $E(\mathbf{t})$ conditional on the particular $\{y_{i,n}\}$ sequences used. Let E_n denote an average taken over trials and E_i an average over subjects.

$$\begin{aligned} E(\mathbf{t} \mid \{y_{i,n}\}) &= E_n E_i E_x(y_{i,n} \mathbf{x}_{i,n+1} \mid p_{i,n}) \\ &= E_n E_i (\alpha_1 y_{i,n} p_{i,n} + a_1 y_{i,n}) \\ &= \alpha_1 E_n E_i (y_{i,n} p_{i,n}) + a_1 \bar{y}, \end{aligned} \quad (68)$$

where \bar{y} is the average value of $y_{i,n}$ for the sequences used. The corresponding equation in terms of statistics of the data is

$$\frac{\sum_n \sum_i y_{i,n} x_{i,n+1}}{NI} = \alpha_1 \frac{\sum_n \sum_i y_{i,n} x_{i,n}}{\sum_n \sum_i y_{i,n}} + a_1 \bar{y}. \quad (69)$$

The expectation in Eq. 68 can be evaluated if parameters are known, or Eq. 69 can be used for estimation or testing.

By using the fact that the $y_{i,n}$ are generated by a probability mechanism, we can arrive at an approximation that is easier to work with. We do this by evaluating $E(\mathbf{t})$ for the "average" $y_{i,n}$ -sequence produced by the probability mechanism rather than for the particular sequences used in the experiment. The approximation is obtained by applying to Eq. 68 the expectation operator E_y , which averages over the binomial outcome distribution of $y_{i,n}$. Let $V_1 = E_n(V_{1,n})$, where the expectation is taken over the indicated trial block. Then,

$$\begin{aligned} E_y E(\mathbf{t} \mid \{y_{i,n}\}) &= \alpha_1 E_n E_i E_y (y_{i,n} p_{i,n}) + a_1 E_y (\bar{y}) \\ &= \alpha_1 \pi E_n E_i (p_{i,n}) + a_1 \pi, \end{aligned}$$

and so

$$E_y E(\mathbf{t}) = \alpha_1 \pi V_1 + a_1 \pi. \quad (70)$$

The corresponding equation in terms of statistics of the data is

$$\frac{\sum_n \sum_i y_{i,n} x_{i,n+1}}{NI} = \alpha_1 \pi \frac{\sum_n \sum_i x_{i,n}}{NI} + a_1 \pi, \tag{71}$$

which is to be compared to the more exact Eq. 69.

5.4 Conditional Expectations and the Development of Functional Equations

Conditional expectations are useful also in establishing functional equations for interesting model properties. Let $G_1, G_2, \dots, G_k, \dots$ be a set of mutually exclusive and exhaustive events, and let \mathbf{h} be a statistic whose expectation is desired. Then the property used is

$$E(\mathbf{h}) = \sum_k \Pr \{G_k\} E(\mathbf{h} \mid G_k), \tag{72}$$

and we consider two examples of its application to path-independent models with two subject-controlled events. Let $\mathbf{x}_n = 1$ if there is an error on trial n and $\mathbf{x}_n = 0$ if there is a success; let the operator for error be Q_2 and for success, Q_1 . As our first example we let $\mathbf{h} = \mathbf{u}_1$, the total number of errors in an infinite sequence. The conditioning events are the possible responses on trial 1, so that G_1 corresponds to $\mathbf{x}_1 = 0$ and G_2 corresponds to $\mathbf{x}_1 = 1$. Equation 72 becomes

$$E(\mathbf{u}_1 \mid p_1 = p) = \Pr \{ \mathbf{x}_1 = 0 \} E(\mathbf{u}_1 \mid \mathbf{x}_1 = 0, p_1 = p) + \Pr \{ \mathbf{x}_1 = 1 \} E(\mathbf{u}_1 \mid \mathbf{x}_1 = 1, p_1 = p).$$

Now we note that $E(\mathbf{u}_1 \mid \mathbf{x}_1 = 0, p_1 = p) = E(\mathbf{u}_1 \mid p_1 = Q_1 p)$; that is, we can consider the process as if it began on the second trial with a different initial probability. Similarly, $E(\mathbf{u}_1 \mid \mathbf{x}_1 = 1, p_1 = p) = 1 + E(\mathbf{u}_1 \mid p_1 = Q_2 p)$; in this case we consider the process as beginning on the second trial but we add the error that has already occurred. The result is

$$E(\mathbf{u}_1 \mid p_1 = p) = (1 - p) E(\mathbf{u}_1 \mid p_1 = Q_1 p) + p[1 + E(\mathbf{u}_1 \mid p_1 = Q_2 p)]. \tag{73}$$

For a particular model, the expectation $E(\mathbf{u}_1 \mid p_1 = p)$ depends on the parameters and the value of p ; we can suppress the parameters and write it simply as $f(p)$, a function of p . This function is unknown, but Eq. 73 tells us that it has the property that

$$f(p) = (1 - p)f(Q_1 p) + p[1 + f(Q_2 p)].$$

If $Q_1p = \alpha_1p$ and $Q_2p = \alpha_2p$, then

$$f(p) = (1 - p)f(\alpha_1p) + p[1 + f(\alpha_2p)], \quad (74)$$

with the boundary condition $f(0) = 0$. Equation 74 is an example of a *functional equation*, which defines some property of an unknown function that we seek to specify explicitly. It has been studied by Tatsuoka and Mosteller (1959).

In the preceding example a relation is given among the values of the function at an infinite set of triples of the values of its argument; that is, the set defined $\{p, \alpha_1p, \alpha_2p \mid 0 \leq p \leq 1\}$. A more familiar example of a functional equation is a difference equation; the values of the argument differ only by multiples of some constant. An example of such a set of arguments is $\{p, p + h, p + 2h \mid p = 0, h, 2h, \dots, Nh\}$. Without loss of generality, a difference equation of this kind can be converted into one in which the arguments of the function are a subset of successive integers. A second familiar example of a functional equation is any differential equation. For both of these special types of functional equations, there is a much wider variety of methods of solution—methods of specifying the unknown function—than for the more general equations.

As a second example of the use of Eq. 72 in developing a functional equation let us consider a model with two subject-controlled events in which $x_n = 1$ results in an increase in $\Pr\{x_n = 1\} = p_n$ toward $p_n = 1$ and $x_n = 0$ results in a decrease in p_n toward $p_n = 0$. In such a model, after a sufficient number of trials, any response sequence will consist of either all “errors” or all “successes”; there are two asymptotically absorbing barriers, at $p = 1$ and $p = 0$.

An example of a linear-operator model of this kind is

$$\mathbf{p}_{n+1} = \begin{cases} Q_1\mathbf{p}_n = \alpha_1\mathbf{p}_n + (1 - \alpha_1), & \text{with probability } p_n, \\ Q_2\mathbf{p}_n = \alpha_2\mathbf{p}_n, & \text{with probability } 1 - p_n. \end{cases}$$

One of the interesting questions about such a model is to determine the probability of asymptotic absorption at $p_\infty = 1$. Bush and Mosteller (1955, p. 155) show by an elementary argument that the distribution of p_∞ in this model is entirely concentrated at the two absorbing barriers. Therefore $\Pr\{p_\infty = 1\} = E(p_\infty)$, and it is fruitful to identify the \mathbf{h} in Eq. 72 with p_∞ . As before, we let G_1 correspond to $x_1 = 0$ and G_2 correspond to $x_1 = 1$, and we consider the expectation as a function of the starting probability. Equation 72 thus becomes

$$\begin{aligned} \Pr\{p_\infty = 1 \mid p_1 = p\} &= E(p_\infty \mid p_1 = p) \\ &= \Pr\{x_1 = 0\} E(p_\infty \mid x_1 = 0, p_1 = p) \\ &\quad + \Pr\{x_1 = 1\} E(p_\infty \mid x_1 = 1, p_1 = p). \end{aligned}$$

We note that $E(\mathbf{p}_\infty \mid \mathbf{x}_1 = 0, p_1 = p) = E(\mathbf{p}_\infty \mid p_1 = Q_1p)$ and similarly that $E(\mathbf{p}_\infty \mid \mathbf{x}_1 = 1, p_1 = p) = E(\mathbf{p}_\infty \mid p_1 = Q_2p)$. We then have

$$E(\mathbf{p}_\infty \mid p_1 = p) = pE(\mathbf{p}_\infty \mid p_1 = Q_1p) + (1 - p)E(\mathbf{p}_\infty \mid p_1 = Q_2p).$$

Letting $g(p)$ represent the expectation as a function of the starting probability, we arrive at

$$g(p) = pg(Q_1p) + (1 - p)g(Q_2p) \tag{75}$$

as a functional equation for the probability of absorption at $p = 1$. Boundary conditions are given by $g(1) = 1$ and $g(0) = 0$. The function $g(p)$ is understood to depend on parameters of the model in addition to p_1 . Mosteller and Tatsuoka (1960) and others²² have studied this functional equation for the foregoing linear-operator model. In general, no simple closed solution seems to be available. For the symmetric case of $\alpha_1 = \alpha_2 = \alpha < 1$ the solution is $g(p) = p$.

A similar functional equation can be developed for the beta model with two absorbing barriers and symmetric events. It is convenient to work with logit \mathbf{p}_n instead of \mathbf{p}_n itself. Following Eq. 49, we have, for this model

$$\text{logit } \mathbf{p}_n = -(a + bt_n - bs_n),$$

and the corresponding operator expression is

$$\text{logit } \mathbf{p}_{n+1} = \begin{cases} \text{logit } \mathbf{p}_n + b & \text{if } \mathbf{x}_n = 1 & \text{(i.e., with probability } \mathbf{p}_n), \\ \text{logit } \mathbf{p}_n - b & \text{if } \mathbf{x}_n = 0 & \text{(i.e., with probability } 1 - \mathbf{p}_n). \end{cases}$$

Let $L_n = \text{logit } \mathbf{p}_n$ and let $g(L)$ be the probability of absorption at $L = \infty$ (which corresponds to $\mathbf{p}_\infty = 1$) for a process that starts at $L_1 = L$. Then Eq. 75 becomes the linear difference equation

$$g(L) = pg(L + b) + (1 - p)g(L - b), \tag{76}$$

where $p = \text{antilogit } L = 1/(1 + e^{-L})$. The boundary conditions are $g(-\infty) = 0$ and $g(+\infty) = 1$. This equation has been studied by Bush (1960) and Kanai (1962b).

5.5 Difference Equations

The discreteness of learning models makes difference equations ubiquitous in their exact analysis. The recursive equation for \mathbf{p}_n is a difference equation whose solution is given by the explicit equation for \mathbf{p}_n ; the argument of the difference equation is in this case the trial number n .

²² See Shapiro and Bellman, cited by Bush & Mosteller, 1955.

A simple example is the recursive equation for the single linear operator model $p_{n+1} = \alpha p_n$, whose solution is $p_n = \alpha^{n-1} p_1$. A more interesting case is Eq. 13 for the prediction experiment,

$$p_{n+1} = \alpha p_n + (1 - \alpha) y_n,$$

with the solution

$$p_{n+1} = \alpha^{n-1} p_1 + (1 - \alpha) \sum_{j=1}^{n-1} \alpha^{n-1-j} y_j.$$

There are systematic methods of solution for many linear difference equations such as these (see, for example, Goldberg, 1958). Often a little manipulation yields a conjectured solution whose validity can be proved by mathematical induction.

Partial difference equations occasionally arise in learning-model analysis; they are more difficult to solve. We have seen in Sec. 5.2 that it is often necessary to know the moments $\{V_{m,n}\} = \{E(\mathbf{p}_n^m)\}$ of the p -value distributions generated by a model; properties of a model are often expressed in terms of these moments. The transition rules of linear-operator models lead to linear difference equations or other recurrence formulas for the moments. Occasionally these equations are "ordinary": one of the subscripts of $V_{m,n}$ is constant throughout the equation. An example is the equation for $V_{1,n}$ in the experimenter-controlled events model above, when we consider the $\{y_j\}$ to be random variables and $\Pr\{y_n = 1\} = \pi$; it is given by the ordinary difference equation (Eq. 43)

$$V_{1,n+1} = \alpha V_{1,n} + (1 - \alpha)\pi, \quad (77)$$

whose solution is easily obtained (Eq. 44).

It is more usual for neither m nor n to be constant in the recurrence formula for $V_{m,n}$, and the formula is then a partial difference equation. In this case we cannot ignore the fact that $V_{m,n}$ is a function of a bivariate argument, and methods of solution are correspondingly more difficult. As an example we consider the linear-operator model with two commutative events and subject control, for which

$$\mathbf{p}_{n+1} = \begin{cases} \alpha_1 \mathbf{p}_n & \text{with probability } 1 - \mathbf{p}_n \\ \alpha_2 \mathbf{p}_n & \text{with probability } \mathbf{p}_n. \end{cases}$$

First let us consider how the partial difference equation for $V_{m,n}$ is derived. We assume a population of subjects with common values of p_1 , α_1 , and α_2 . On trial n , after $n - 1$ applications of the operators, the population consists of n distinct subgroups defined by the number of times α_1 has been applied. Let $1 \leq \nu \leq n$ be the index for these subgroups, let $p_{\nu,n}$ be the p -value for the ν th subgroup on trial n , and let $P_{\nu,n}$ be the size of this subgroup, expressed as a proportion of the population. Now let us consider the fate of the ν th subgroup on trial n . A proportion,

$p_{v,n}$, of the subgroup makes an error on that trial, and its p -value becomes $\alpha_2 p_{v,n}$. The remaining proportion of the subgroup, $1 - p_{v,n}$, performs a correct response, and its p -value becomes $\alpha_1 p_{v,n}$. The result is expressed in the following table:

| New p -Values | New Proportions | |
|--------------------|-------------------------|------|
| $\alpha_2 p_{v,n}$ | $p_{v,n} P_{v,n}$ | (78) |
| $\alpha_1 p_{v,n}$ | $(1 - p_{v,n}) P_{v,n}$ | |

Therefore,

$$\begin{aligned}
 V_{m,n+1} &= \sum_{v=1}^{n+1} p_{v,n+1}^m P_{v,n+1} \\
 &= \sum_{v=1}^n (\alpha_2 p_{v,n})^m p_{v,n} P_{v,n} + \sum_{v=1}^n (\alpha_1 p_{v,n})^m (1 - p_{v,n}) P_{v,n} \\
 &= (\alpha_2^m - \alpha_1^m) \sum_{v=1}^n p_{v,n}^{m+1} P_{v,n} + \alpha_1^m \sum_{v=1}^n p_{v,n}^m P_{v,n} \\
 V_{m,n+1} &= (\alpha_2^m - \alpha_1^m) V_{m+1,n} + \alpha_1^m V_{m,n}.
 \end{aligned} \tag{79}$$

One feature of this equation, which is generally true of models with subject control, is that $V_{m,n}$ is expressed in terms of moments higher than the m th moment of the p -value distribution on preceding trials. With experimenter control this complicating feature is absent, as illustrated by Eq. 77.

Equation 79 has been solved by conjecture and inductive proof rather than by any direct method. To illustrate how cumbersome some of the results become in this field, I reproduce the solution here:

$$V_{m,n} = \alpha_1^{m(n-1)} p_1^m + \sum_{k=2}^n \alpha_1^{m(n-k)} p_1^{j+m-1} \prod_{j=m}^{k+m-2} \frac{(\alpha_2^j - \alpha_1^j)(1 - \alpha_1^{n-j+m-1})}{1 - \alpha_1^{j-m+1}} \tag{80}$$

$(m \geq 1, n \geq 1)$,

where the sum is defined to be zero for $n = 1$.

For more examples of the development of recursive formulas for moments, see Bush & Mosteller (1955, Chapter 4), and for some examples of their use see Bush (1959), Estes & Suppes (1959, Sec. 8), and Sternberg (1959b).

5.6 Solution of Functional Equations²³

Two methods by which functional equations have been studied are illustrated here; the first is a power-series expansion and the second is a differential equation approximation.

²³ See Kanai (1962a,b) for the formulation of some functional equations arising in the analysis of the linear and beta models and for methods of solution and approximation.

Tatsuoka and Mosteller (1959) solved Eq. 74 by using a power-series expansion. Assume that $f(p)$ is expressible as a power series in p :

$$f^*(p) = \sum_{k=0}^{\infty} c_k p^k. \quad (81)$$

The boundary condition $f^*(0) = 0$ implies that $c_0 = 0$. By substituting the series expansion into the functional equation and equating coefficients of like powers of p we find

$$c_k = \frac{\prod_{j=1}^{k-1} (\alpha_2^j - \alpha_1^j)}{\prod_{j=1}^k (1 - \alpha_1^j)}, \quad k \geq 1. \quad (82)$$

For certain special cases this expression can be simplified. For example, with $\alpha_2 = 1$ (identity operator for "error") and $0 \leq \alpha_1 < 1$, $c_k = 1/(1 - \alpha_1^k)$ and therefore

$$f^*(p) = \sum_{k=1}^{\infty} \frac{p^k}{1 - \alpha_1^k}. \quad (83)$$

A closed form for this expression has not been found, but there are tables (Bush, 1959) and approximations (Tatsuoka & Mosteller, 1959).

The only solution of Eq. 74 that satisfies the boundary condition and has an expansion in powers of p is $f^*(p)$. To prove this, we assume that there is a second power-series solution, $f^{**}(p) = \sum_{k=1}^{\infty} d_k p^k$ and replace $f(p)$ in Eq. 74 first by $f^*(p)$ and second by $f^{**}(p)$. By subtracting the second resulting equation from the first we obtain an equation for

$$f^*(p) - f^{**}(p) = \sum_{k=1}^{\infty} (c_k - d_k) p^k,$$

whose solution requires that $c_k - d_k = 0$ for $k \geq 1$. In general, however, a functional equation may possess solutions for which a power-series expansion is not possible. For this reason it is necessary either to provide a general proof of the uniqueness of $f^*(p)$ or to show that we are interested only in solutions of Eq. 74 with power-series expansions.

Kanal (1960, 1962a) has shown that $f^*(p)$ is the only solution of Eq. 74 that is continuous at $p = 0$ but no *general* proof of uniqueness is available at present. Fortunately, we can use Eq. 80 to show that for the model in question $E(\mathbf{u}_1)$ has a series expansion in powers of p and that power-series solutions of Eq. 74 are therefore the only ones of interest. To do this, we note that

$$E(\mathbf{u}_1) = E\left(\sum_{n=1}^{\infty} \mathbf{x}_n\right),$$

and that

$$E\left(\sum_{n=1}^{\infty} \mathbf{x}_n\right) = \sum_{n=1}^{\infty} E(\mathbf{x}_n) \tag{84}$$

if the right-hand series converges. Equation 80 provides an expression for $E(\mathbf{x}_n) = V_{1,n}$; it is a polynomial in $p_1 = p$. If $\alpha_1 < 1$ and either $\alpha_2 < 1$ or $p_1 < 1$, $\sum_{n=1}^{\infty} V_{1,n}$ converges. We therefore know, incidentally, that under these conditions $E(\mathbf{u}_1)$ exists. Moreover, because it is the sum of a convergent infinite series of polynomials, it must have a power-series expansion.

For some functional equations the power series that is obtained may not converge, and we cannot apply the foregoing method. As an example of a second method we consider Bush's (1960) solution of Eq. 76. First the equation is written in terms of first differences and $(1 - p)/p$ is replaced by e^{-L} :

$$g(L + b) - g(L) = e^{-L}[g(L) - g(L - b)]. \tag{85}$$

In order to convert (85) into a linear equation, the logarithmic transformation is applied to both sides, and the logarithm of the first difference is defined as a new function, $h(L) \equiv \log [g(L) - g(L - b)]$, to give

$$h(L + b) = h(L) - L. \tag{86}$$

Equation 86 is the difference equation to be solved.

In this case a solution is sought, not by a power series expansion but by a differential equation approximation of the difference equation. We write Eq. 86 in a form symmetric about L :

$$h\left(L + \frac{b}{2}\right) - h\left(L - \frac{b}{2}\right) = -\left(L - \frac{b}{2}\right),$$

divide by b ,

$$\frac{\Delta h}{\Delta L} = -\frac{L}{b} + \frac{1}{2},$$

and treat the result as a derivative,

$$\frac{dh}{dL} = -\frac{L}{b} + \frac{1}{2}.$$

Integration gives

$$h(L) = -\frac{L^2}{2b} + \frac{L}{2} + C \tag{87}$$

as a conjectured solution. The result satisfies Eq. 86, and therefore a particular solution of the complete equation is given by Eq. 87 with $C = 0$.

The homogeneous equation $h(L + b) = h(L)$ has as its general solution

$\tilde{F}(L)$, an arbitrary periodic function of L with period b . For the general solution of the complete equation we then have

$$h(L) = \frac{L}{2} - \frac{L^2}{2b} + \tilde{F}(L). \tag{88}$$

To recover g , we use

$$g(L) - g(L - b) = \exp [h(L)] = \exp \left[\frac{L}{2} - \frac{L^2}{2b} + \tilde{F}(L) \right],$$

and completing the square gives us

$$g(L) - g(L - b) = P(L) \exp \left[-\frac{1}{2b} \left(L - \frac{b}{2} \right)^2 \right], \tag{89}$$

where $P(L)$ is some other periodic function of L , with period b . This new difference equation is simpler than the original (Eq. 85) because it contains one difference instead of two, and therefore routine procedures can be used. We first note the boundary conditions $g(-\infty) = 0$ and $g(\infty) = 1$. Then we replace L by $L - b$, $L - 2b$, and so on, to obtain the semi-infinite system

$$\begin{aligned} g(L) - g(L - b) &= P(L) \exp \left[-\frac{1}{2b} \left(L - \frac{b}{2} \right)^2 \right] \\ g(L - b) - g(L - 2b) &= P(L) \exp \left[-\frac{1}{2b} \left(L - b - \frac{b}{2} \right)^2 \right] \\ g(L - 2b) - g(L - 3b) &= P(L) \exp \left[-\frac{1}{2b} \left(L - 2b - \frac{b}{2} \right)^2 \right] \\ &\dots \end{aligned} \tag{90}$$

Addition of these equations and use of the first boundary condition gives

$$g(L) = P(L) \sum_{k=0}^{\infty} \exp \left[-\frac{1}{2b} \left(L - kb - \frac{b}{2} \right)^2 \right].$$

The sum may be approximated by a normal integral. The periodic function is still arbitrary; to specify it, we write the full infinite system corresponding to Eqs. 90, sum, and use both boundary conditions; the final result is

$$g(L) = \frac{\sum_{k=0}^{\infty} \exp \left[-\frac{1}{2b} \left(L - kb - \frac{b}{2} \right)^2 \right]}{\sum_{k=-\infty}^{\infty} \exp \left[-\frac{1}{2b} \left(L - kb - \frac{b}{2} \right)^2 \right]}, \tag{91}$$

which is roughly of the form of a normal integral. In most practical cases $P(L)$ may be approximated by a constant. When antilogits are taken, the absorption probability as a function of p is no longer a normal integral,

but it is similar in character; the resulting *S*-shaped curve is to be compared to the result for the symmetric linear-operator model for which the absorption probability is equal to the initial probability.

6. SOME ASPECTS OF THE APPLICATION AND TESTING OF LEARNING MODELS

6.1 Model Properties: A Model Type as a Subspace

In the last three sections I have mentioned examples of many of the model properties that have been studied and compared with data. Linear models are the best known: they were studied by Bush and Mosteller (1955) and recent progress, a good deal of which is represented in Bush & Estes (1959), has been considerable. Even so, our knowledge tends to be spotty, concentrated at certain special examples of linear models. Models with extreme or equal limit points and those with equal learning-rate parameters are better understood than the others. The single-operator model and models with experimenter control are the most thoroughly studied (Bush & Sternberg, 1959; Estes & Suppes, 1959); analytic expressions are available for the expectations of a good many statistics of these models. On the other hand, except for some of its asymptotic properties, we have less information about Luce's beta model (Bush, 1960; Lamperti & Suppes, 1960; Kanal, 1962a,b). For this model, and for any more general logistic models, there is a compensating advantage: standard methods of estimation can easily be applied. Once estimates are obtained, Monte Carlo calculations can be used for detailed comparisons. It can be argued, moreover, that the advantages of optimal (maximum-likelihood) estimation methods outweigh the convenience of having analytic expressions for model properties.

To arrive at one view of the properties of a model—a view that is helpful in considering the problems of fitting and testing the model—we begin by considering the *m*-dimensional "property-space" consisting of all values of the vector (s_1, s_2, \dots, s_m) , where s_j denotes a property (the expectation or variance of a statistic) of the model. The corresponding statistic for some observed data sequences is denoted by \bar{s}_j . In general, the properties depend on parameter values, and therefore $s_j = s_j(\Theta)$, where Θ is a vector of parameters corresponding to a point in the parameter space. As the point moves through the entire parameter space, the s_j take on all the combinations of values allowed by the model type. For certain purposes we can now ignore the parameters and consider only these allowed combinations, which define a subspace of the property-space. If there were no

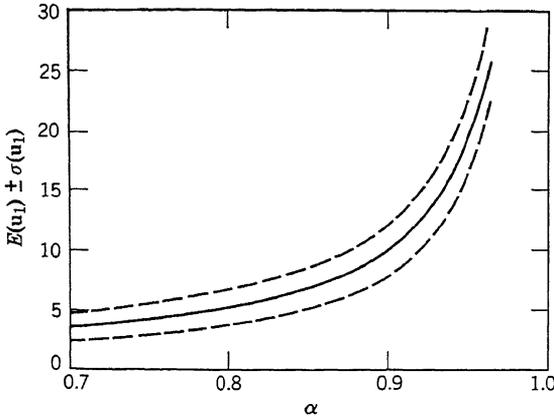


Fig. 11. The solid curve represents the expected number of errors in an infinite sequence of trials, $E(\mathbf{u}_1)$, as a function of the learning-rate parameter α for the single-operator model (Eq. 28) with $p_1 = 1$. The distance between the solid curve and a broken curve represents the standard deviation of the number of errors.

sampling variability, the problem of testing a model type would reduce to the question whether the observed $(\bar{s}_1, \bar{s}_2, \dots, \bar{s}_m)$ is a point in the subspace. The existence of sampling fluctuations means that the question must be modified so that a certain degree of discrepancy is tolerated.

A simple example with a one-dimensional parameter space illustrates this viewpoint. We consider the single-operator model given by Eq. 28 and discussed in Sec. 4.4. A good many properties of this model are known (Bush & Sternberg, 1959), and in Figs. 11, 12, and 13 three are illustrated graphically. The total number of errors (in an infinite sequence of trials) is symbolized by \mathbf{u}_1 , the total number of runs of errors by \mathbf{r}_T , and the number of trials before the first success by \mathbf{f} . These three properties suffice for our purpose, and the property-space we consider is therefore three-dimensional. In order for the model to have only a single free parameter we assume that p_1 , the initial probability of error, is known to be unity. The dependence of $E(\mathbf{f})$, $E(\mathbf{u}_1)$, and $E(\mathbf{r}_T)$ on the value of the learning-rate parameter α is shown by the figures and also by the following equations.²⁴

$$E(\mathbf{u}_1) = \frac{1}{1 - \alpha}, \quad E(\mathbf{r}_T) = \frac{1}{1 - \alpha^2}, \quad E(\mathbf{f}) = \sum_{k=0}^{\infty} \alpha^{k(k+1)/2}. \quad (92)$$

²⁴ The infinite sum for $E(\mathbf{f})$ is tabulated in Bush & Mosteller (1955, Table A) and approximated in Galanter & Bush (1959).

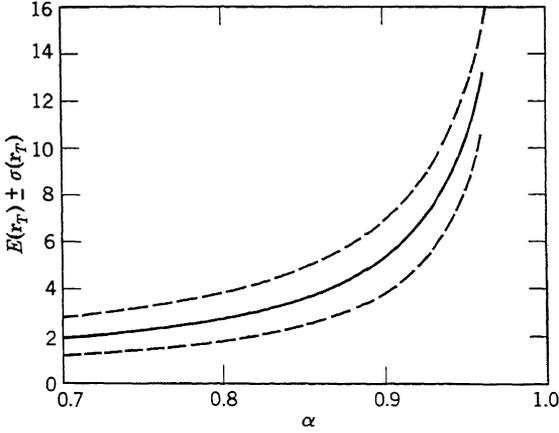


Fig. 12. The solid curve represents the expected number of runs of errors in an infinite sequence of trials, $E(x_T)$, as a function of the learning-rate parameter α for the single-operator model (Eq. 28) with $p_1 = 1$. The distance between the solid curve and a broken curve represents the standard deviation of the number of runs of errors.

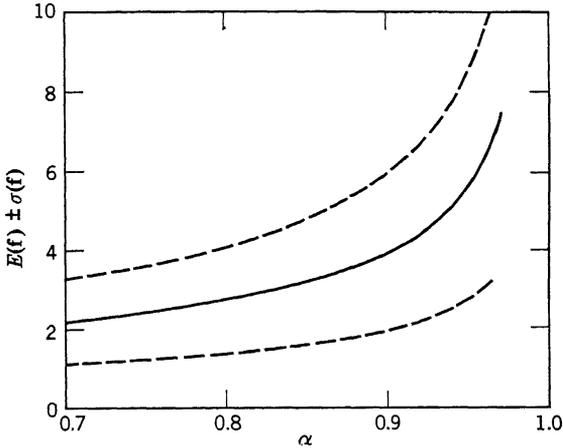


Fig. 13. The solid curve represents the expected number of trials before the first success $E(f)$ as a function of the learning-rate parameter α for the single-operator model (Eq. 28) with $p_1 = 1$. The distance between the solid curve and a broken curve represents the standard deviation of the number of trials before the first success.

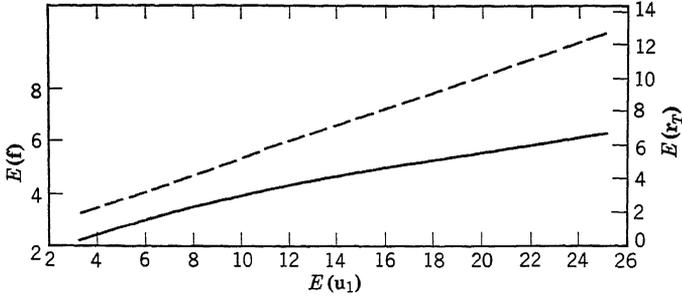


Fig. 14. The solid curve (left-hand ordinate) describes $E(\mathbf{f})$ as a function of $E(\mathbf{u}_1)$ for the single-operator model (Eq. 28) with $p_1 = 0$. The broken curve (right-hand ordinate) describes $E(\mathbf{r}_T)$ as a function of $E(\mathbf{u}_1)$ for this model. Units are chosen so that scales are comparable in standard deviation units.

Also shown for each statistic is the interval defined by twice its standard deviation, for example, $\pm\sqrt{\text{Var}(\mathbf{u}_1)}$. The parameter α can be eliminated from any pair of the functions $s_j(\alpha)$, to define a relation between the two expectations. Two such relations are shown in Fig. 14, in which $E(\mathbf{f})$ and $E(\mathbf{r}_T)$ are plotted against $E(\mathbf{u}_1)$. These two relations are projections of a curve in the three-dimensional property-space with dimensions $E(\mathbf{f})$, $E(\mathbf{r}_T)$, and $E(\mathbf{u}_1)$; this curve is the subspace of the property-space to which the model corresponds. The units are so chosen that the scales are roughly comparable in standard deviation units, that is, a 1-cm discrepancy on the $E(\mathbf{u}_1)$ scale is as serious as a 1-cm discrepancy on either of the other scales.

In Table 4 the average values of the three statistics are given for three

Table 4 Observed Values of u_1, f, r_T in Three Experiments

| Experiment | \bar{u}_1 | \bar{f} | \bar{r}_T |
|--|-------------|-----------|-------------|
| T-Maze reversal after overlearning (Galanter & Bush, 1959) | 24.68 | 13.32 | 6.11 |
| T-Maze reversal (Galanter & Bush, 1959, Period 2) | 14.10 | 5.30 | 6.60 |
| Solomon-Wynne shuttlebox (Bush & Mosteller, 1959) | 7.80 | 4.50 | 3.24 |

experiments in which the assumption that $p_1 = 1$ is tenable. It is instructive to examine these values in conjunction with the four graphs for the single-operator model. In none of the experiments does any of the pairs of statistics satisfy the relations for this simple model. Put another way, in no case does the observed point $(\bar{u}_1, \bar{f}, \bar{r}_T)$ fall within the allowed subspace. How large a discrepancy is tolerable is a statistical problem that is touched on later.

A good deal of the work that has been done on fitting and testing models can be thought of as being analogous to the process exemplified above: mathematical study of a model yields several functions, $s_j(\Theta)$, of properties in terms of parameter values, and the question is asked whether there is a choice of parameter values (a point in the parameter space) for which the observed \bar{s}_j are close to their theoretical values. The process is usually conducted in two stages: first, estimation, in which parameter values $\hat{\Theta}$ are selected so that a subset of the \bar{s}_j agrees exactly with the theoretical values, and, second, testing, in which the remaining \bar{s}_j are compared to their corresponding $s_j(\Theta)$. In the second stage some or all of the theoretical values may be estimated from Monte Carlo calculations. These stages correspond in Fig. 14, for example, to first letting \bar{u}_1 determine a point on the abscissa and, second, comparing the corresponding ordinate values to \bar{f} and \bar{r}_T .

Conclusions from this method are conditional on the choice of properties used in each of the two stages. To assert that "model X cannot describe the observed distribution of error-run lengths" is stronger than is usually warranted. More often the appropriate statement is of the form "when parameters for model X are chosen so that it describes \bar{s}_1 and \bar{s}_2 exactly then it cannot describe \bar{s}_3 ." Occasionally there are exceptions in which some property of a model is independent of its parameter values. For example, we can assert unconditionally that "an S-shaped curve of probability versus trials cannot be described by the single-operator model" or that "the linear-operator model with complementary experimenter-controlled events for the prediction experiment must (if learning occurs at all) produce asymptotic probability-matching for all values of π ." Such parameter-free properties of a model are worthy of energetic search.

6.2 The Estimation Problem

Most model types have one or more free parameters whose values must be estimated from data. Estimates that satisfy over-all optimal criteria, such as maximum likelihood or minimum chi-square, cannot usually be obtained explicitly in terms of statistics of the data. Because the iterative

or other numerical methods that are needed in order to obtain such estimates are inconvenient, they have seldom been used in research with learning models. The more common method has been briefly touched on in Sec. 6.1: parameter values are chosen to equate several of the observed statistics of the data with their expectations as given by the model. The estimates $\hat{\Theta}$ are therefore produced by the solution of a set of equations of the form $s_j(\Theta) = \bar{s}_j$.

Because the properties of estimates so obtained are not well understood, this method may lead to serious errors, as is illustrated by an example. Let us suppose that a learning process behaves in accordance with the single-operator model with $\alpha = 0.90$ and that we do not know this but wish to test the model as a possible description of the data. Suppose that we have a single sequence of responses and that we have reason to assume that $p_1 = 1$. Suppose, further, that because of sampling variability the number of errors at the beginning of the sample sequence is accidentally too large and that the observed values of u_1 and f are inflated, each by an amount equal to its theoretical standard deviation. Figures 12 and 14 then provide us with the following values:

| | u_1 | f |
|-------------------------|-------|------|
| True (population) value | 10.00 | 3.91 |
| Sample value | 12.18 | 5.91 |

We have two properties s_j , namely u_1 and f , and a single parameter, α , to estimate. One property is needed for estimation and the remaining one is available for testing. The choice of which property to use for which purpose involves only two alternatives; but it has the essential character of the more complicated choice usually available to the investigator. One procedure has been to use "gross" features of the data, such as the total number of errors, for estimation, and "fine-grain" features, such as the distribution of error-run lengths, for testing. (For examples, see Bush & Mosteller, 1959, and Sternberg, 1959b). Occasionally the investigator cannot choose; whatever few statistics he is lucky enough to have analytic expressions for are automatically elected for use in estimation, and for testing he must resort to Monte Carlo calculations. (For an example, see Bush, Galanter, & Luce, 1959.) Let us examine, in the case of the two statistics tabulated above, how the choice that is made affects our inference about goodness of fit of the model.

First, suppose that we use f for estimation, choosing α so that $E(f | \alpha) = \bar{f}$. Entering Fig. 13 with the observed value of $\bar{f} = 5.9$, we find the corresponding estimate to be $\hat{\alpha} = 0.957$. Now to test the model we refer to Fig. 11. Corresponding to $\alpha = 0.957$ are the values of $E(u_1) = 23.3$ and $\sigma(u_1) = 3.35$. The difference between $E(u_1)$ and its observed value of

$\bar{u}_1 = 12.18$ is more than three times its theoretical standard deviation, a sizable discrepancy. On these grounds we would be inclined to discard the model. But first let us consider the result if we take the second option. We use u_1 for estimation, and Fig. 11 gives the value of $\hat{\alpha} = 0.917$ for $E(u_1) = 12.18$. To test the model, we enter Fig. 13 with $\alpha = 0.917$; the corresponding theoretical values are $E(f) = 4.3$, $\sigma(f) = 2.15$. The observed value $\bar{f} = 5.91$ is therefore within one standard deviation of the theoretical value. The second option inclines us to accept the model. It is worth noting that, if anything, this example is conservative: had the number of errors been accidentally large, but toward the end of the sequence instead of the beginning, \bar{f} would have been close to its theoretical value, \bar{u}_1 would have been inflated, and the two results would have been still more discrepant.

The reason for the disagreement may be clarified by Fig. 14. Here it can be seen that for this particular model an error in $E(f)$ corresponds to a much larger error in $E(u_1)$, in terms of standard deviation units. The total-errors statistic is the most "sensitive" of the three; therefore, to give the model the best chance, it is the one that should be used for estimation.

The question of the choice of an estimating statistic is therefore a delicate one. In the lucky instance in which a model approximates the data well in many respects it is unimportant how the estimation is carried out. Such an instance is the Bush-Mosteller (1955) analysis of the Solomon-Wynne data, in which several methods gave estimates in very good agreement; indeed, this fact in itself is strong evidence in favor of the model. But this is a rare case, and more often estimates are in conflict.

The question becomes especially important when several models are to be compared in their ability to describe a set of data. It is crucial that the estimation methods be equally "fair" to the models, and the standard procedures do not ensure this. For example, we might be comparing with the single-operator model of Fig. 14 another hypothetical model for which the curve of $E(f)$ versus $E(u_1)$ had a slope greater rather than less than unity. If we then used u_1 in estimation for both models, we would be prejudicing the comparison in favor of the single-operator model. For different models different sets of statistics may be the best estimators: we do not ensure equal fairness by using the same estimating statistics for all the models to be compared. This observation, for which I am indebted to A. R. Jonckheere,²⁵ casts doubt on the results of certain comparative studies, such as those of Bush and Mosteller (1959), Bush, Galanter, and Luce (1959), and Sternberg (1959b).

One possibility for retrieving the situation is to search for aspects of the data that one or more of the competing models are incapable of describing,

²⁵ Personal communication, 1960.

regardless of the values of its parameters. An example arises in the analysis of reversal after overlearning in a T-maze, one of the experiments included by Bush, Galanter, and Luce (1959) in their comparison of the linear and beta models. The observed curve of proportion of successes versus trials starts at zero and is markedly S-shaped, rising steeply in its middle portion. Parameters can be chosen for the beta model so that its curve agrees well with the one observed. But, as Galanter and Bush (1959) show, although the linear model of Eq. 8 is capable of producing an S-shaped curve that starts at zero ($p_1 = 1$), no choice of α_1 and α_2 permits its curve to rise both slowly enough at the beginning and end of learning and steeply enough in its middle portion. As the analysis stands, then, the beta model is to be preferred for these data. The problem is that if we search long enough we may be able to find a property of the data that this model cannot describe and that the linear model can. To choose between the models, we would then have to decide which of the two properties is the more "important," and the problem of being equally fair to the competing models would again face us.

A solution to the problem lies in the use of maximum likelihood (or other "best") estimates (Wald, 1948), despite their frequent inconvenience, and in the comparison of the maximized likelihoods and the use of likelihood-ratio tests to assess relative goodness of fit. Bush and Mosteller (1955) discuss several over-all measures of goodness of fit. The use of such over-all tests has occasionally been objected to on grounds that they may be sensitive to uninteresting differences among models or between models and data and that they may not reveal the particular respects in which a model is deficient. Our example of the f and u_1 statistics shows that the first objection applies to the more usual methods as well. In answer to the second objection, there is no reason why detailed comparison of particular statistics cannot be used as a supplement to the over-all test.

One of the desirable features of the beta model and of more general logistic models is that a simple set of sufficient statistics exists for the parameters and that the standard iterative method (Berkson, 1957) for obtaining the maximum-likelihood estimates is easily generalized for more than two parameters, converges rapidly, and is facilitated by existing tables. Cox (1958) suggests that, in applications to learning, initial estimates be obtained by the minimum-logit χ^2 method (Berkson, 1955; Anscombe, 1956), which does not require iteration.

Examples of the results of these methods applied to the Solomon-Wynne data (Sec. 4.1) are given in Table 5. Both the maximum-likelihood and minimum-logit χ^2 methods can be thought of as ways of fitting the linear regression equation given by Eq. 49:

$$\text{logit } p_n = -(a + bt_n + cs_n).$$

The random variables t_n and s_n are considered to be the independent variables, and the logit of the escape probability is the dependent variable. The equation defines a plane; the observed points to which the plane is fitted are the logits of the proportions of escapes at given values of (t_n, s_n) . A difficulty arises with the minimum-logit χ^2 method when the observed proportion for a (t_n, s_n) pair is zero, as happens often when $s_n + t_n$ is large in the later trials. Most of these zero observations were omitted in obtaining the values in the second row of Table 5, so that this row of values depends principally on early trials. Relations between the values of \hat{p}_1 , $\hat{\beta}_1$, and $\hat{\beta}_2$ and the estimates of a , b , and c are given in Sec. 2.5.

Table 5 Results of Four Procedures for Estimating Parameters of the Beta Model (Eqs. 17, 19) from the Solomon-Wynne Data

| Method | \hat{p}_1 (initial escape probability) | $\hat{\beta}_1$ (avoidance) | $\hat{\beta}_2$ (escape) |
|-----------------------------------|--|-----------------------------|--------------------------|
| Bush-Galanter-Luce (1959) | 0.94 | 0.59 | 0.83 |
| Minimum logit χ^2 | 0.864 | 0.760 | 0.778 |
| One maximum-likelihood iteration | 0.857 | 0.805 | 0.718 |
| Two maximum-likelihood iterations | 0.857 | 0.811 | 0.735 |

When there are only two parameters, as in the case of the beta model for two symmetric experimenter-controlled events (Eq. 24),

$$\text{logit } p_n = -(a + bd_n),$$

a simple graphical method (Hodges, 1958) provides close approximations to the maximum-likelihood estimates; it is probably preferable to minimum-logit χ^2 for obtaining starting values for maximum-likelihood iteration. The minimum-logit χ^2 method should be used with caution; it may occasionally be misleading, perhaps because of the difficulty with zero entries already mentioned. Consider, as an example, the logistic one-trial perseveration model (Eq. 32): $\text{logit } p_n = a + b(n-2) + cx_{n-1}$, ($n \geq 2$). A simple graphical method is the visual fitting of a pair of parallel lines to the proportions that estimate $\Pr(x_n = 1 \mid x_{n-1} = 0)$ and $\Pr(x_n = 1 \mid x_{n-1} = 1)$ when they are plotted against $n \geq 2$ on logistic (or normal probability) paper. These lines then represent $\text{logit } p_n = a + b(n-2)$ and $\text{logit } p_n = (a+c) + b(n-2)$ and provide estimates of the three parameters. Values obtained for the Goodnow data (Sec. 4.5), using the graphical method and then applying one cycle of maximum-likelihood

iteration to its results, are presented in Table 6. For these data, the minimum-logit χ^2 method gave values that departed more from the maximum-likelihood values than the simple graphical procedure.

The advantages of maximum-likelihood estimates are that their variances are known, at least asymptotically, and that their values tend to represent much of the information in the data. When the maximum-likelihood method is not used, alternative methods that have these properties are to be preferred. As an example, let us consider estimation for the linear-operator model with experimenter control (Eq. 12) that has been used for the prediction experiment with $\Pr \{y_n = 1\} = \pi$ and $\Pr \{y_n = 0\} = 1 - \pi$. If

Table 6 Estimates for the Logistic Perseveration Model (Eq. 32) from the Goodnow Data

| Method | \hat{a} | \hat{b} | \hat{c} |
|--|-----------|-----------|-----------|
| Visual fit of parallel lines on logistic paper | 0 | -0.24 | 0.94 |
| One maximum-likelihood iteration | 0.035 | -0.236 | 0.927 |

t_i is the total number of A_1 responses by the i th subject during the first N trials, then for this model

$$E(t_i) = N\pi - (\pi - V_{1,1}) \left(\frac{1 - \alpha^N}{1 - \alpha} \right), \tag{93}$$

and, having determined $\hat{V}_{1,1}$, we can estimate α by setting $E(t_i)$ equal to the value of t for a group of subjects. This is the method used by Bush and Mosteller (1955), Estes and Straughan (1954), and others. Equation 93 is obtained by adding both sides of the approximate equation for the learning curve (Eq. 44) over trials, $1 \leq n \leq N$.

One of the observed features of estimates obtained by this method is that $\hat{\alpha}$ varies with the value of π : the higher π (the "easier the discrimination"), the more potent the learning-rate parameter. Psychological mechanisms have been proposed to explain this effect (e.g., Estes, 1959), but very little is known about the method of estimation itself. For example, is $\hat{\alpha}$ unbiased? If not, how does the bias depend on π ? The estimate depends entirely on preasymptotic data. (This can be seen from the fact that its value is indeterminate if $V_{1,1} = \pi$.) For experiments in which $V_{1,1} \simeq 0.5$, therefore, the higher the value of π , the more data are used in the estimate of α , hence the more reliable the estimate. Other information about this estimation procedure that is not known but that is

vital for the interpretation of the findings is the extent to which perturbations in the process affect the estimate. Examples of such perturbations are intersubject differences in initial probabilities and in values of α and the difference between the effects of nonreward and reward discussed in Sec. 4.3.

The curious fact that no information about α seems to be available from asymptotic data is a result of averaging over the distribution of outcomes and using the resulting approximate equation (Eq. 44). Clearly there is more information in the data than is used for the estimate. This sequential information has been exploited by Anderson (1959) in a variety of procedures for estimation and testing. A simple improvement on the average learning curve procedure arises from the idea that the extent to which responses are correlated with the immediately preceding outcomes depends on the value of α and leads to the use of Eq. 71, or its exact version, Eq. 69, with $\alpha_1 = 1 - \alpha_1$, to estimate α . By this method, estimates can be obtained even from trend-free portions of the data. Such an improvement is in the direction of the use of sufficient statistics, to which the maximum-likelihood method often leads. But it is still inferior to what is possible for the comparable beta model.

6.3 Individual Differences

In most of the discussion in this chapter, and in most applications of learning models, it is assumed that the same values of the initial probability and other parameters characterize all the subjects in an experimental group. When events are subject-controlled, differences in p -values arise on trials after the first, but under this homogeneity assumption these are due entirely to differences between event sequences.

It must be kept in mind, when this assumption is made in the application of a model type, that what is tested by comparisons between data and model is the conjunction of the assumption and the model type and not the model type alone. It is convenience, not theory, that leads to the homogeneity assumption. The question of primary interest is whether each individual subject, with his own parameter values, can be said to behave in accordance with the model type. It is usually thought that if the assumption is not entirely justified then the discrepancy will cause the model to underestimate the intersubject variances of response-sequence statistics. It is hoped (but not known) that the discrepancy will have no other adverse effects. We therefore expect the variances given by a model to be on the small side, and we are not perturbed when this occurs, as it often does (Bush & Mosteller, 1959; Sternberg, 1959b).

On the other hand, unless we are interested specifically in testing the homogeneity assumption, it is probably unwise to use an observed variance as a statistic for estimation, and this is seldom done. One difficulty with the customary procedure, in which the assumption of homogeneity is made, is that estimation and testing methods for different models may be differentially sensitive to deviations from homogeneity. For comparing models, therefore, it is probably preferable to estimate parameters separately for each subject. Audley and Jonckheere (1956) argued for the desirability of this procedure, and Audley (1957) carried it out for a model that describes both choice and choice time.

Estimates for an individual subject that are based on few observations may be unstable. One way of avoiding both the instability of individual estimates and the assumption of homogeneity is to study long sequences of responses from individual subjects. Anderson (1959) favors this method and gives estimation procedures. However, it clearly cannot be applied to experiments in which a single response is perfectly learned in a small number of trials.

Certain types of inference, based on between-subject comparisons, may be misleading if the homogeneity assumption is not met. For example, we might observe a positive correlation between the number of errors before and after some arbitrary trial. One possible cause of the correlation is a positive response effect. A second is the existence of individual differences. Even if a response-independent, single-event model describes the process for each subject, differences in initial probabilities or learning rates will produce a positive correlation of this kind. If there is a negative response effect as well as individual differences, statistics such as the variance of total errors or the correlation of early and late errors might lead us to infer no response effect at all if we assume homogeneity. If, on the other hand, we observe a negative correlation between the number of early and late errors, despite the possible interference of individual differences, we are on sure ground when we infer a negative response effect. By the same token, when homogeneity is assumed and the theoretical variances are too large, we have especially strong evidence against the model type. This last effect has occurred in the analysis of data from both humans (Sternberg, 1959b) and rats (Galanter & Bush, 1959; Bush, Galanter, & Luce, 1959).

When we look at the over-all picture of results from the application of learning models, it is remarkable how weak the evidence against the homogeneity assumption usually appears to be. One is reluctant to believe that individuals are so alike, but the only alternative seems to be that our testing methods are insensitive. N. H. Anderson²⁶ has argued

²⁶ Personal communication, 1962.

that this alternative gains support from the fact that analyses of variance of repeated measurements almost always yield significant individual differences.

Direct tests of the homogeneity assumption have occasionally been performed. Data from a single trial alone are, of course, useless as evidence of any more than the first moment of the p -value distribution. But if the p -value for each subject is approximately constant during a block of m trials, then raw moments from the first to the m th can be estimated from the block. As an example, suppose we use a block containing trials one and two. The method²⁷ depends on the two relations

$$E(\mathbf{x}_1 + \mathbf{x}_2) = E_p E_x(\mathbf{x}_1 + \mathbf{x}_2) = E_p(2\mathbf{p}) = 2V_1$$

$$E[(\mathbf{x}_1 + \mathbf{x}_2)^2] = E_p E_x[(\mathbf{x}_1 + \mathbf{x}_2)^2] = E_p(2\mathbf{p} + 2\mathbf{p}^2) = 2V_1 + 2V_2,$$

where p is the (approximately constant) probability on the two trials and V_1 and V_2 are the (approximate) first and second moments of its distribution. The expectations on the left are replaced by the averages of $x_1 + x_2$ and $(x_1 + x_2)^2$ over subjects, and then the equations are solved for \hat{V}_1 and \hat{V}_2 .

The homogeneity assumption requires that on the first trial $V_2 = V_1^2$. In his analysis of the Goodnow two-armed bandit data Bush estimated these two quantities by using three-trial blocks, drawing a smooth curve through the estimates and extrapolating back to the first trial. The result was $\hat{V}_{2,1} = 0.13$ and $\hat{V}_{1,1}^2 = 0.11$, making the homogeneity assumption tenable for initial probability.

In another test of the assumption Bush and Wilson (1956) examined the number of A_1 responses in the first 10 trials of a two-choice experiment for each of 49 paradise fish. The distribution of number of choices had more spread than could be accounted for by a common probability for all subjects. The assumption was therefore rejected and a distribution of initial probabilities was used. Instead of one initial probability parameter, two were then needed, one giving the mean and the other giving the variance of the distribution of initial probabilities, whose form was assumed to be that of a beta distribution.

Even less work has been done in which variation in the learning-rate parameters is allowed. One example appears in Bush and Mosteller's (1959) analysis of the Solomon-Wynne data: the linear single-operator model was used with a distribution of α -values. In certain respects this generalization improved the agreement between model and data.

²⁷ This "block-moment" method was developed by Bush, as a general estimation scheme, in an unpublished manuscript, 1955.

6.4 Testing a Single Model Type

In a good deal of the work with learning models a single model is used in the analysis of a set of data. Estimates are obtained, and then several properties of the model are compared with their counterparts in the data. There is little agreement as to which properties should be examined or how many. Informal comparisons, sometimes aided by the theoretical variances of the statistics considered, are used in order to decide on the model's adequacy. Values of the parameter estimates may be used as descriptive statistics of the data.

As with any theory, a stochastic learning model can be more readily discredited than it can be accepted. Two reasons, however, lead investigators to expect and allow some degree of discrepancy between model and data. One reason is the view, held by some, that a model is intended only as an approximation to the process of interest. A second is the fact that today's experimental techniques probably do not prevent processes other than the one described by the model from affecting the data. The matter is one of degree: how good an approximation to the data do we desire and to which of their properties? And how deeply must we probe for discrepancies before we can be reasonably confident that there are no important ones? Recent work that reveals how difficult it may be to select among models (e.g., Bush, Galanter, & Luce, 1959; Sternberg, 1959b) suggests that some of our testing methods for a single model may lack power with respect to alternatives of interest to us and that we may be accepting models in error.

One finding is that the learning curve is often a poor discriminator among models. Two examples have already been illustrated in this chapter. In Figs. 1 and 2 a model with experimenter-controlled events provides an excellent description of learning curves generated by a process with a high degree of subject control. In Fig. 9 four models that differ fundamentally in the nature of their response effects produce equally good agreement with an observed learning curve.

It would be an error to conclude from these examples that the learning curve can never discriminate between models; this is far from true. Occasionally it provides us with strong negative evidence. We have already seen (Sec. 4.4) that the beta model with experimenter control cannot account for the asymptote of the learning curve in prediction experiments with $\pi \neq \frac{1}{2}$, in those experiments in which probability matching occurs. On the other hand, the linear experimenter-controlled event model can be eliminated for a T-maze experiment (Galanter & Bush, 1959) in which its

theoretical asymptote is exceeded. The shape of the preasymptotic learning curve may also occasionally discriminate between models; for example, as mentioned in Sec. 6.2, the linear-operator model cannot produce a learning curve that is steep enough to describe the T-maze data of Galanter and Bush (1959) on reversal after overlearning, whereas the beta model provides a curve that is in reasonable agreement with these data. The important point, however, is that agreement between an observed learning curve and a curve produced by a model cannot, alone, give us a great deal of confidence in the model.

More surprising, perhaps, is that the distribution of error-run lengths also seems to be insensitive. In Fig. 9 it can be seen that three distinctly different models can be made to agree equally well with an observed distribution. As another example, let us consider the fourth period of a T-maze reversal experiment of Galanter and Bush (1959, Experiment III). In this experiment three trials were run each day, and by the fourth period there appeared a marked daily "recovery" effect: on the first trial of each day there was a large proportion of errors. Needless to say, this effect was not a property of the path-independent model used for the analysis. Despite the oscillating feature of the learning curve, a feature that one might think would have grave consequences for the sequential aspects of the data, the agreement between model and data, as regards the run-length distribution, was judged to be satisfactory. Again, as for the learning curve, there are examples in which the run-length distribution can discriminate. In Fig. 9 it can be seen that one of the four models cannot be forced into agreement with it. And Bush and Mosteller (1959) show that a Markov model and an "insight" model, when fitted to the Solomon-Wynne data, produce significantly fewer runs than the other models studied.

We do not wish to be limited to negative statements about the agreement between models and data, yet we have evidence that some of the usual tests are insensitive, and we have no rules to tell us when to stop testing. In comparative studies of models this situation is somewhat ameliorated: we continue making comparisons until all but one of the competing models is discredited. Another possible solution is to use over-all tests of goodness of fit. As already mentioned, these tests suffer from being powerful with respect to uninteresting alternatives: such a test, for example, might lead us to discard a model type under conditions in which only the homogeneity assumption is at fault. In contrast, the usual methods seem to suffer from low power with respect to alternatives that may be important.

One role proposed for estimates of the parameters of a model is that they can serve as descriptive statistics of the data. Such descriptive statistics are useful only if the model approximates the data well and if the values are not strongly dependent on the particular method used for

their estimation. I have already discussed how an apparent lack of parameter invariance from one experiment to another may be an artifact of applying the wrong model. This has been recognized in recent suggestions that the invariance of parameter estimates from experiment to experiment be used as an additional criterion by which to test a model type.

6.5 Comparative Testing of Models

As I have already suggested, the comparative testing of several models improves in some ways on the process of testing models singly. The investigator is forced to use comparisons sensitive enough so that all but one of the models under consideration can be discredited. Attention is thereby drawn to the features that distinguish the models used, and this allows a more detailed characterization of the data than might otherwise be possible.

As an example, let us take the Bush-Mosteller (1959) comparison of eight models in the analysis of the Solomon-Wynne data. Examination of several properties eliminates all but two of the models. The remaining two are the linear-operator model (Eq. 8) and a model of the kind developed by Restle (1955). A theoretical comparison of the two models is necessary to discover a potential discriminating statistic. The "Restle model" is an example of a single-event model; under the homogeneity assumption all subjects have the same value of p_n . The linear-operator model, on the other hand, involves two subject-controlled events, and parameter estimates suggest a positive response effect. This difference should be revealed in the magnitude of the correlation between number of early and late errors (shocks). The linear-operator model calls for a positive correlation; Restle's model (together with homogeneity) calls for a zero correlation. The observed correlation is positive, and the linear-operator model is selected as the better. (This inference exemplifies the type discussed in Sec. 6.3 that depends critically on the validity of the homogeneity assumption.)

As a second example of a comparative study let us take the Bush-Wilson study (1956) of the two-choice behavior of paradise fish. On each trial the fish swam to one of two goalboxes. On 75% of the trials food was presented in one (the "favorable" goalbox); on the remaining 25% the food was presented in the other. For one group of subjects food in one goalbox was visible from the other. Two models were compared, each of which expressed a different theory about the effects of nonfood trials. The first theory suggests that on these trials the performed response is weakened, giving a model that has commonly been applied to the prediction experiment with humans:

Information Model

| Event | P_{n+1} |
|------------------------------|---------------------------|
| Favorable goalbox, food | $\alpha p_n + 1 - \alpha$ |
| Favorable goalbox, no food | αp_n |
| Unfavorable goalbox, food | αp_n |
| Unfavorable goalbox, no food | $\alpha p_n + 1 - \alpha$ |

The second theory suggests that on nonfood trials the performed response is strengthened:

Secondary Reinforcement Model

| Event | P_{n+1} |
|------------------------------|-------------------------------|
| Favorable goalbox, food | $\alpha_1 p_n + 1 - \alpha_1$ |
| Favorable goalbox, no food | $\alpha_2 p_n + 1 - \alpha_2$ |
| Unfavorable goalbox, food | $\alpha_1 p_n$ |
| Unfavorable goalbox, no food | $\alpha_2 p_n$ |

We have already seen that the information model produces "probability-matching behavior": if this model applied, each fish would tend to divide its choices in the ratio 75:25, and no individual would consistently make one choice. In regard to the proportion of "favorable" choices averaged over subjects, probability-matching can also be produced by the secondary reinforcement model with the appropriate choice of parameters. (Again we have a case in which the average learning curve does not help us to discriminate between models.) There is a clear difference, however, if we examine properties, other than the mean, of the distribution over subjects of asymptotic choice proportions. The secondary reinforcement model implies that a particular fish will stabilize on either one choice or the other in the long run and that some will consistently choose the favorable goalbox, others the unfavorable. (If the proportion of fish that stabilized on the favorable side were 0.75, then the average learning curve would suggest probability matching.)

The observed distribution of the proportion of choices to the favorable side on the last 49 trials of the experiment is U-shaped, with most fish either at very low values or very high values, giving support to the secondary reinforcement model. One merit of this study that should be mentioned is that the decision between the two models can be made without having to estimate values for the parameters.

A third example of a comparative study that makes use of data from a two-armed bandit experiment was discussed in Sec. 4.5.

Occasionally we wish to compare models that contain different numbers of free parameters. When this occurs, a new problem is added to that of equal "fairness" of the estimation and testing procedures discussed in Sec. 6.2: the model with fewer degrees of freedom will be at a disadvantage. This difficulty can be overcome if one model is a special case of another. If so, we can apply the usual likelihood-ratio test procedure, which takes into account the difference in the number of free parameters.

Suppose, for example, that we wish to decide between the single-event beta model and the beta model with two subject-controlled events (Eq. 20) in application to an experiment such as the escape-avoidance shuttlebox. The second model is the more general and can be written

$$\text{logit } \mathbf{p}_n = -(a + bt_n + cs_n). \quad (94)$$

The first model is given by the same equation, but with $c = b$, so that

$$\text{logit } p_n = -[a + b(t_n + s_n)] = -(a + bn). \quad (95)$$

The test is equivalent to the question: are the magnitudes of the two event effects equal? It is performed by obtaining maximum-likelihood estimates of a , b , and c in Eq. 94 and of a and b in Eq. 95 and calculating the maximized likelihood for each model. These steps are straightforward for logistic models. Because Eq. 94 has an additional degree of freedom, the likelihood associated with it will generally be greater than the likelihood associated with the first model. The question of how much greater it must be in order for us to reject the first model and decide that the two events have unequal effects can be answered by making use of the (large sample) distribution of the likelihood ratio λ under the hypothesis that the equal event model holds (Wilks, 1962).²⁸

In an alternative procedure a statistic that behaves monotonically with the likelihood ratio is used, and its (small sample) distribution under the hypothesis of equal event effects can be obtained analytically or, if this is difficult, from a Monte Carlo experiment based on Eq. 95. A comparable test for a generalized linear-operator model has been developed and applied by Hanania (1959) in the analysis of data from a prediction experiment. She concludes for those data that the effect of reward is significantly greater than the effect of nonreward.

6.6 Models as Baselines and Aids to Inference

Most of our discussion to this point has been oriented towards the question whether a model can be said to describe a set of data. Usually a model entails several assumptions about the learning process, and

²⁸ For this example $-2 \log \lambda$ is distributed as chi-square with one degree of freedom.

therefore in asking about the adequacy of a model we are testing the set of assumptions taken together. Models are also occasionally useful when a particular assumption is at stake or when a particular feature of the data is of interest. Occasionally, as with the "null hypothesis" in other problems, it is the discrepancy between model and data that is of interest, and the analysis of discrepancies may reveal effects that a model-free analysis might not disclose. A few examples may make these ideas clear.

In Sec. 6.5 the choice between the two models of Eqs. 94 and 95 was equivalent to the question whether the effects of reward and nonreward are different. We might, on the other hand, start with this question and perform the same analysis, not being concerned with whether either of the models fitted especially well but simply with whether one (Eq. 94) was significantly preferable to the other (Eq. 95).

With the same kind of question in mind we could estimate reward and nonreward parameters for some model and compare their values. One difficulty with this procedure is that unless we know the sampling properties of the estimator it is difficult to interpret such results. A second difficulty is that if a model is not in accord with the data the estimates may depend strongly on the method used to obtain them. An example of this dependence is shown in Table 5; different estimates lead to contradictory answers to the question which of the two events has the larger effect. That the parameters in a model properly represent the event effects in a set of data to which the model is fitted may be conditional on the validity of many of the assumptions embodied in the model. What is needed for this type of question is a model-free test—one that makes as few assumptions as possible. Applications of such tests are illustrated in Sec. 6.7.

If a model agrees with a number of the properties of a set of data, then the discrepancies that do appear may be instructive and may be useful guides in refining the model. One example is provided by the analysis of free-recall verbal learning by a model developed by Miller and McGill (1952). The model is intended to apply to an experiment in which, on each trial, a randomly arranged list of words is presented and the subject is asked to recall as many of the words as he can. The model assumes that the process that governs whether or not a particular word is recalled on a trial is independent of the process for any of the other words. The model is remarkably successful, but one discrepancy appears when the estimated recall probability after ν recalls, \hat{p}_ν , is examined as a function of ν and compared to the theoretical mean curve (Bush & Mosteller, 1955, p. 234). The observed proportions oscillate about the mean more than the model allows. This suggests the hypothesis that a subject learns words in clusters rather than independently and that either all or none of the words in a cluster tend to be recalled on a trial.

A second example of the baseline use of a model is provided by Sternberg's (1959b) analysis of the Goodnow data by means of the linear one-trial perseveration model (Eq. 29). Several properties of the data are described adequately by the model, but in at least one respect it is deficient. The model implies that for all trials $n \geq 2$, $\Pr \{x_n = 1 \mid x_{n-1} = 1\} - \Pr \{x_n = 1 \mid x_{n-1} = 0\} = \beta$, a constant. What is observed is that the difference between the estimates of these conditional probabilities decreases somewhat during the course of learning. It was inferred from this finding and from other properties of the data that the tendency to perseverate may change as a function of experience. This example may serve to caution us, however, and to indicate that the hypotheses suggested by a baseline analysis may be only tentative. The inference mentioned depends strongly on the use of a linear model. If the logistic perseveration model (Eq. 32) is used instead, the observed decrease in the difference between conditional probabilities is produced automatically, without requiring changes in parameter values during the course of learning.

A final use of models as aids to inference is in the study of methods of data analysis. In an effort to reveal some feature of data the investigator may define a statistic whose value is thought to reflect the feature. Because the relations between the behavior of the statistic and the feature of interest may not be known, errors of inference may occur. These errors are sometimes known as *artifacts*: the critical property of the statistic arises from its definition rather than from the feature of interest in the data.

A simple example of an artifact has already been mentioned in Sec. 6.3. If individuals differ in their p_1 -values and if the existence of response effects is assessed by means of the correlation over subjects between early and late errors, then a positive response effect may be inferred when there is none.

A second example is the evaluation of response-repetition tendencies. If subjects differ in their p -values on trial $n - 1$ and if the existence of a perseverative tendency is assessed by comparing the proportions that correspond to $\Pr \{x_n = 1 \mid x_{n-1} = 1\}$ and $\Pr \{x_n = 1 \mid x_{n-1} = 0\}$, then a perseverative tendency may be inferred where none is present. This occurs because the samples used to determine the two proportions are not randomly selected: they are chosen on the basis of the response on trial $n - 1$. The subjects used to determine $\Pr \{x_n = 1 \mid x_{n-1} = 1\}$ tend to have higher p -values on trial $n - 1$ (and consequently on trial n) than the subjects in the other sample. Selective sampling effects of this kind have been discussed by Anderson (1960, p. 85) and Anderson and Grant (1958).

The question in both of these examples is how extreme a value of the statistic must be observed in order to make the inference valid. In order to answer this question, the behavior of the statistic under the hypothesis

of no effect must be known. Such behavior can be studied by applying the method of analysis to data for which the underlying process is known, namely Monte Carlo sequences generated by a model.

A more complicated problem to which this type of study has been applied is the criterion-reference learning curve (Hayes & Pereboom, 1959). This is a method of data analysis in which an average learning curve is constructed by pooling scores for different subjects on trials that are specified by a performance criterion rather than by their ordinal position. Underwood (1957) developed a method of this kind in an effort to detect a cyclical component in serial verbal learning. The result of the analysis was a learning curve with a distinct cyclical component. Whether the inference of cyclical changes in p_n is warranted depends on the result of the analysis when it is applied to a process in which p_n increases monotonically. Hayes and Pereboom apply the method to Monte Carlo sequences generated by such a process and obtain a cyclical curve. We conclude that to infer cyclical changes in p_n the magnitude of the cycles in the criterion-reference curve must be carefully examined; the existence of cycles is not sufficient.

6.7 Testing Model Assumptions in Isolation

A particular learning model can be thought of as the embodiment of several assumptions about the learning process. Testing the model, then, is equivalent to testing all of these assumptions jointly. If the model fails, we are still left with the question of the assumptions that are at fault. Light may be shed on this question by the comparative method and the detailed analysis of discrepancies discussed in the last two sections. But a preferable technique is to test particular assumptions in as much isolation from the others as possible. Examples of several techniques are described in this section.

To illustrate the equivalence of a model to several assumptions, each of which might be tested separately, let us consider the analysis of the data from an escape-avoidance shuttlebox experiment by means of the linear-operator model (Eqs. 7 to 11 in Sec. 2.4). Suppose that an estimation procedure yields an avoidance operator that is more potent than the escape operator ($\alpha_1 < \alpha_2$). Some of the assumptions involved in this model are listed:

1. The effect of an event on $\mathbf{p}_n = \Pr \{\text{escape}\}$ is manifested completely on the next trial.
2. Conditional on the value of \mathbf{p}_n , the effect of an event is independent of the previous sequence of events.

3. Conditional on the value of \mathbf{p}_n , the effect of avoidance (reward) on \mathbf{p}_n is greater than the effect of escape (nonreward).

4. The reduction in \mathbf{p}_n caused by an event is proportional to the value of \mathbf{p}_n .

5. The proportional reduction is constant throughout learning; therefore the change in \mathbf{p}_n induced by avoidance, for example, decreases during the course of learning.

6. The value of \mathbf{p}_n for a subject depends only on the number of avoidances and escapes on the first $n - 1$ trials and not on the order in which they occurred.

7. All subjects have the same values of p_1 , α_1 , and α_2 .

Let us consider the third assumption listed: avoidance has more effect than escape. It has already been indicated that to test this assumption simply by examining the parameter estimates for a model may be misleading. For the Solomon-Wynne data, Bush and Mosteller (1955) have found the values $\hat{\alpha}_1 = 0.80$, $\hat{\alpha}_2 = 0.92$, and $p_1 = 1.00$ for the linear model, confirming the third assumption. In Table 5 it is shown that estimates for the beta model may or may not confirm the assumption, depending on the method of estimation used.

We wish to test this assumption without the encumbrance of all the others. One step in this direction is to apply a more general model, in which at least some of the constraints are discarded. Hanania's work (1959) provides an example of this approach, in which she uses a linear, commutative model, but with trial-dependent operators, so that it is quasi-independent of path:

$$\mathbf{p}_{n+1} = \begin{cases} w_{n+1}\mathbf{p}_n & \text{if } \mathbf{x}_n = 1 & \text{(nonreward)} \\ \theta w_{n+1}\mathbf{p}_n & \text{if } \mathbf{x}_n = 0 & \text{(reward).} \end{cases} \quad (96)$$

Although θ is assumed to be constant, w_n may take on a different value for each trial number. The explicit formula, after a redefinition of the parameters, is

$$\mathbf{p}_n = \theta^{s_n} u_n, \quad (97)$$

where

$$\mathbf{s}_n = \sum_{j=1}^{n-1} \mathbf{x}_j.$$

When the w_n (or the u_n) are all equal, this formulation reduces to the Bush-Mosteller model. The relative effect of reward and nonreward is reflected by the value of θ , for which Hanania develops statistical tests.

This method is an improvement, but a good many assumptions are still needed. A more direct, if less powerful, method that requires fewer

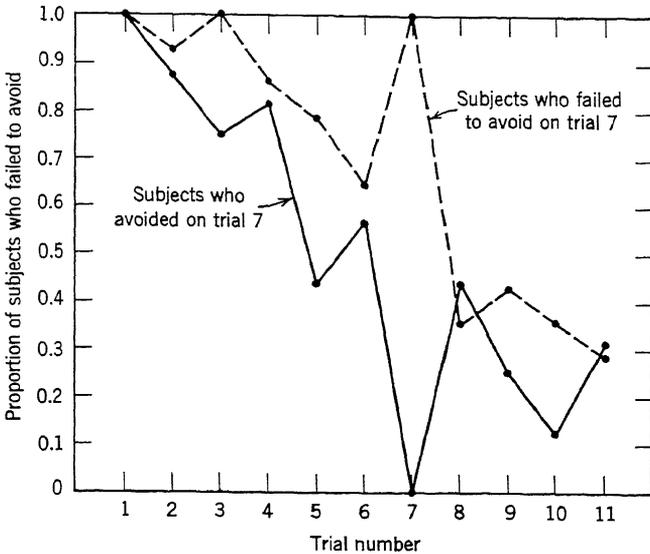


Fig. 15. Performance of two groups of dogs in the Solomon-Wynne avoidance-learning experiment, selected on the basis of their response on trial 7. The broken curve represents the performance of the 14 dogs who failed to avoid on trial 7. The solid curve represents the performance of the 16 dogs who avoided on trial 7.

assumptions is illustrated in Figs. 15 and 16. A trial is selected on which each response is performed by about half the subjects. The subjects are divided into two groups, according to the response they perform, and learning curves are plotted separately for each group. In Fig. 15 this analysis is performed on the Solomon-Wynne shuttlebox data. Subjects are selected on the basis of their seventh response (x_7). It can be seen that animals who escape on trial 7 have a relatively high escape probability on the preceding trials. There is a positive correlation between escape on trial 7 and the number of escapes on earlier trials.

One assumption is needed in order to make the desired inference: the absence of individual differences in parameter values. Individual differences alone, with no positive response effect, could produce a result of this kind; slower learners would tend to fall into the escape group on trial seven. On the other hand, if we assume no individual differences, the result strongly suggests that there is a positive response effect, confirming the third assumption. This result also casts doubt on the validity of the beta model for these data. As shown by Table 5, estimates for that model are either in conflict with the third assumption or require an absurdly low value for the initial escape probability.

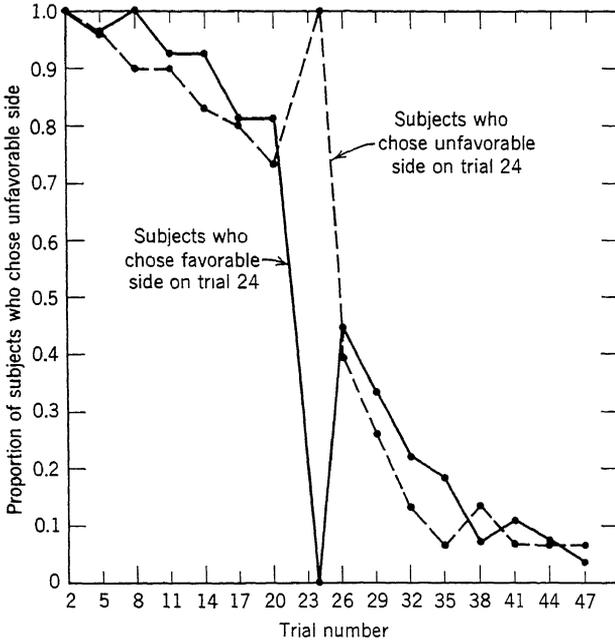


Fig. 16. Performance of two groups of rats in the Galanter-Bush experiment on reversal after overlearning, selected on the basis of their response on trial 24. The broken curve represents the performance of the 10 rats that chose the unfavorable side on trial 24. The solid curve represents the performance of the nine rats that chose the favorable side on trial 24. Except for the point at trial 24, points represent average proportions for blocks of three trials.

Figure 15 also illustrates the errors in sampling that can occur if a subject-controlled event is used as a criterion for selection. This type of error was touched on briefly in Sec. 6.6. It is exemplified by an alternative method that we might have used for assessing the response effect. In this method we would compare the performance of the two subgroups on trials *after* the seventh. Escape on trial 7 is associated with a high escape probability on future trials. It would be an error, however, to infer from this fact alone that the seventh response caused the observed differences; the subgroups are far from identical on trials before the seventh. The method of selecting subjects on the basis of an outcome and examining their future behavior to assess the effect of the outcome must be used with caution. It can be applied when either the occurrence of the outcome is controlled by the experimenter and is independent of the state of the

subject or when it can be demonstrated that the subgroups do not differ before the outcome. For an example of this type of demonstration in a model-free analysis of the effects of subject-controlled events, see Sheffield (1948).

Figure 16 gives the result of dividing subjects on the basis of the twenty-fourth trial of an experiment on reversal after overlearning (Galanter & Bush, 1959). These data are mentioned in Sec. 4.5 (Table 3), in which the coefficient of variation of u_1 is shown to be unusually small, and in Sec. 6.4, in which the inability of a linear-operator model to fit the learning curve is discussed. The results of Fig. 16 are the reverse of those in Fig. 15: animals that make errors on trial 24 tend to make *fewer* errors on preceding and following trials than those that give the correct response on trial 24. This negative relationship cannot be attributed to failure of the homogeneity assumption; individual differences in parameter values would tend to produce the opposite effect. Therefore we can conclude, without having to call on even the homogeneity assumption, that in this experiment there is a negative response effect.

This result gives additional information to help choose between the linear (Galanter & Bush, 1959) and beta (Bush, Galanter, & Luce, 1959) models for the data on reversal after overlearning. Estimation for the linear model suggested a positive response effect, which had to be increased if the learning curve was to be even roughly approximated. Because of its positive response effect, the model produced a value for $\text{Var}(u_1)$ that was far too large. For the beta model, on the other hand, estimation produces results in agreement with the analysis of Fig. 16 and the value for $\text{Var}(u_1)$ is slightly too small, a result consistent with the existence of small individual differences that are not incorporated in the model. The conclusion seems clear that, of the two, the beta model is to be preferred.

We are in the embarrassing position of having discredited both the linear and beta models, each in a different experiment. Unfortunately we cannot conclude that one applies to rats and the other to dogs; evidence similar to that presented clearly supports the linear model for a T-maze experiment on reversal without overlearning (Galanter & Bush, 1959, Experiment III, Period 2).

It has not been mentioned that when outcomes are independent of the state of the subject a model-free analysis of their effects can be performed. As an example, let us consider a T-maze experiment with a 75:25 reward schedule. The favorable (75%) side is baited with probability 0.75, independent of the rat's response. Suppose that we wish to examine the effect of reward on response probability. Rats that choose the favorable side on a selected trial are divided into those that are rewarded on that

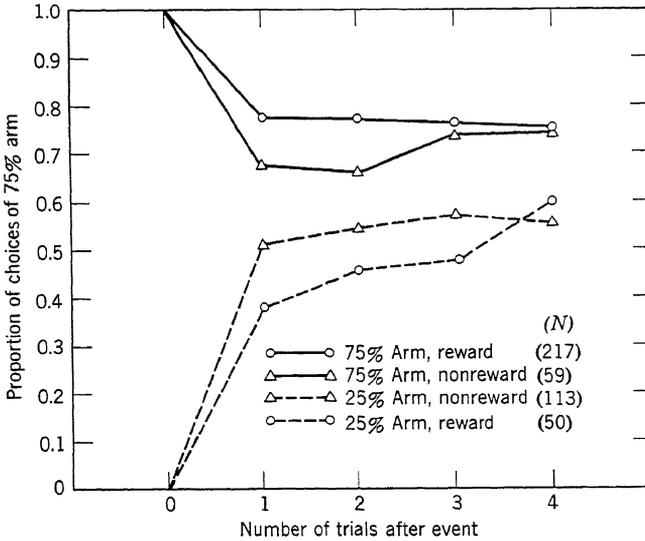


Fig. 17. Performance after four different events in the Weinstock 75:25 T-maze experiment, as a function of the number of trials after the event. The solid (broken) curves represent performance after choice of the 75% (25%) arm. Circles (triangles) represent performance after reward (nonreward). The number of observations used for each curve is indicated.

trial and those that are not. The behavior of the subgroups can be compared on future trials, and any differences can be attributed to the reward effect. To enlarge the sample size, averaging procedures can be employed. The same comparison can be made among the rats that choose the unfavorable side on the selected trial. Results of an analysis of this kind are shown in Fig. 17. The data are the first 20 trials of a 75:25 T-maze experiment conducted by Weinstock (1955). On each trial, n , the rats were divided into four subgroups on the basis of the response and outcome, and the number of choices during each of the next four trials was tabulated for each subgroup. These choice frequencies for all values of n , $1 \leq n \leq 20$, were added and the proportions given in the figure were obtained.

The results are in comforting agreement with the assumptions in several of our models. After reward, the performed response has a higher probability of occurrence than after nonreward, and this is true for both responses. In keeping with the first assumption mentioned in this section, there is no evidence of a delayed effect of the reinforcing event: the hypothesis that its full effect is manifested on the next trial cannot be rejected. On the contrary, there is a tendency for the effect to be reduced

as trials proceed; this last finding would be expected if the effect of reward were less than that of nonreward.

A method that has been used to study a "negative-recency effect" in the binary prediction experiment (e.g., Jarvik, 1951; Nicks, 1959; Edwards, 1961) provides us with a final example of the testing of model assumptions in isolation. The assumption in question is that the direction in which p_n is changed by an event is the same, regardless of the value of p_n and the sequence of prior events.²⁹

Consider the prediction experiment in terms of four experimenter-subject controlled events, as presented in Table 1. On a trial on which O_1 occurs, either A_1 or A_2 may occur, so that two events are possible. Moreover, which of these two events occurs on a trial depends on the state of the subject. Separate analysis of their effects is therefore difficult, as explained earlier in this section. Fortunately, we are willing to assume that both events have effects that are in the same direction: if either (A_1O_1) or (A_2O_1) occurs on trial n , then $p_{n+1} > p_n$. This assumption allows us to perform the test by averaging over all subjects that experienced O_1 on trial n , regardless of their response, and thus examining the average effect of O_1 on the average p -value. This is equivalent to examining an average of the effects of the two events (A_1O_1) and (A_2O_1) , and therefore the test lacks power: if only one of the events violates the assumption in question, the test can fail to detect the violation.³⁰

The variable of interest is the length of the immediately preceding tuple of O_1 's. Does the direction of the effect of O_1 on $\Pr\{A_1\}$ depend on this length? The method involves averaging the proportion of A_1 responses on trials after all j -tuples of O_1 's in the outcome sequence for various values of j and considering the average proportion as a function of j . The results of such an analysis, for O_2 's as well as O_1 's (Nicks, 1959) are given in Fig. 18. The data are from a 380-trial, 67:33 prediction experiment, and the analysis is performed separately for each quarter of the outcome sequence. Under the assumption in question, and in the light of the fact that a 1-tuple of O_1 's (an O_1 that is preceded by an O_2) markedly increases $\Pr\{A_1\}$, all curves in the figure should have slopes that are uniformly nonnegative. Such is not the case, and the results lead us to reject the assumption.³¹ If we assume in this experiment that the effects

²⁹ As mentioned in Sec. 2.3, simple models exist in which the direction of the effect of an event depends on the value of p_n . Such models have seldom been applied, however.

³⁰ If we assume that the experiment consists of two experimenter-controlled events, then this criticism does not apply; however, this is precisely the sort of additional assumption that we do not wish to make.

³¹ Using this type of analysis of performance in a prediction experiment after 520 trials of practice, Edwards (1961) obtained results favorable to the assumption.

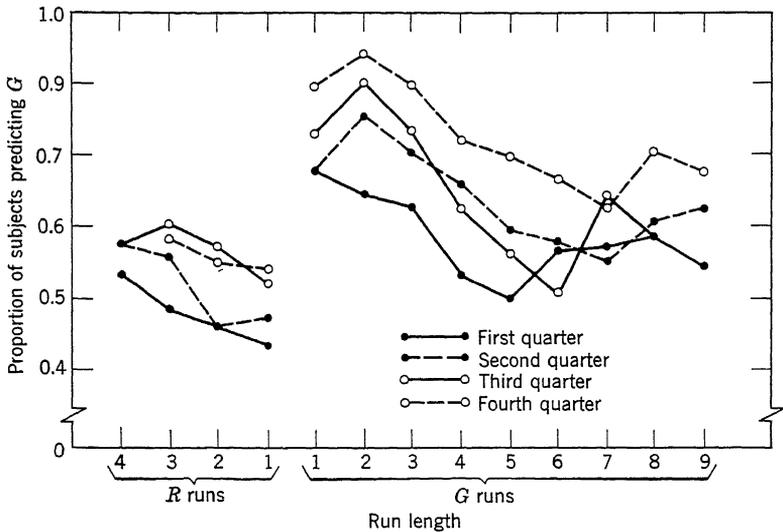


Fig. 18. Proportion of subjects predicting "green" immediately after outcome runs of various lengths of green (G) and red (R) lights in Nick's 67:33 binary prediction experiment. Separate curves are presented for the four quarters of the 380-trial sequence. After Nicks, 1959, Fig. 3.

of events sharing the same outcome are in the same direction, then the direction of the effect of an event appears to depend in a complicated way on the prior sequence of events.

7. CONCLUSION

Implicit in these pages are two alternative views of the place of stochastic models in the study of learning. The first view is that a model furnishes a sophisticated statistical method for drawing inferences about the effects of individual trial events in a learning experiment or for providing descriptive statistics of the data. The method has to be sophisticated because the problem is difficult: the time-series in question is usually nonstationary³² and involves only a small number of observations per subject; if the observations are discrete rather than continuous, each one gives us little information. The use of a model, then, can be thought of as a method of

³² In an experiment in which there is no over-all trend (i.e., $V_{1,n}$ and $V_{2,n}$ are constant), statistical methods for the analysis of stationary time series can be used (see, e.g., Hannan, 1960, and Cox, 1958). Interesting models for such experiments have been developed and applied by Cane (1959, 1961).

combining data—of averaging—in a process with trend. The model is not expected to fit exactly even the most refined experiment; it is simply a tool.

The second view is that a model, or a family of models, is a mathematical representation of a theory about the learning process. In this case our focus shifts from features of a particular set of data to the extent to which a model describes those data and to the variety of experiments the model can describe. We become more concerned with the assumptions that give rise to the model and with crucial experiments or discriminating statistics to use in its evaluation. We attempt to refine our experiments so that the process purportedly described by the model can be observed in its pure form.

Whether we are concerned more with describing particular aspects of the process or more with evaluating an over-all theory, many of the fundamental questions that arise about learning can be answered only by the use of explicit models. The use of models, however, does not automatically produce easy answers.

References

- Anderson, N. H. An analysis of sequential dependencies. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 248–264.
- Anderson, N. H. Effect of first-order conditional probability in a two-choice learning situation. *J. exp. Psychol.*, 1960, **59**, 73–93.
- Anderson, N. H., & Grant, D. A. Correction and reanalysis. *J. exp. Psychol.*, 1958, **56**, 453–454.
- Anscombe, F. J. On estimating binomial response relations. *Biometrika*, 1956, **43**, 461–464.
- Audley, R. J. A stochastic description of the learning behavior of an individual subject. *Quart. J. exp. Psychol.*, 1957, **9**, 12–20.
- Audley, R. J., & Jonckheere, A. R. The statistical analysis of the learning process. *Brit. J. Stat. Psychol.*, 1956, **9**, 87–94.
- Bailey, N. T. J. Some problems in the statistical analysis of epidemic data. *J. Roy. Stat. Soc. (B)*, 1955, **17**, 35–57.
- Barucha-Reid, A. T. *Elements of the theory of Markov processes and their applications*. New York: McGraw-Hill, 1960.
- Behrend, E. R., & Bitterman, M. E. Probability-matching in the fish. *Amer. J. Psychol.*, 1961, **74**, 542–551.
- Berkson, J. Maximum likelihood and minimum χ^2 estimates of the logistic function. *J. Amer. Stat. Assoc.*, 1955, **50**, 130–162.
- Berkson, J. Tables for the maximum-likelihood estimation of the logistic function. *Biometrics*, 1957, **13**, 28–34.
- Bush, R. R. *A block-moment method of estimating parameters in learning models*. Unpublished manuscript, 1955.

- Bush, R. R. Sequential properties of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 215-227.
- Bush, R. R. Some properties of Luce's beta model for learning. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences, 1959*. Stanford: Stanford Univer. Press, 1960. Pp. 254-264. (a)
- Bush, R. R. A survey of mathematical learning theory. In R. D. Luce (Ed.), *Developments in mathematical psychology*. Glencoe, Ill. Free Press, 1960. Pp. 123-170. (b)
- Bush, R. R., & Estes, W. K. (Eds.). *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959.
- Bush, R. R., Galanter, E., & Luce, R. D. Tests of the "beta model." In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 381-399.
- Bush, R. R., & Mosteller, F. A mathematical model for simple learning. *Psychol. Rev.*, 1951, **58**, 313-323.
- Bush, R. R., & Mosteller, F. *Stochastic models for learning*. New York: Wiley, 1955.
- Bush, R. R., & Mosteller, F. A comparison of eight models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 293-307.
- Bush, R. R., Mosteller, F., & Thompson, G. L. A formal structure for multiple-choice situations. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*. New York: Wiley, 1954. Pp. 99-126.
- Bush, R. R., & Sternberg, S. H. A single-operator model. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 204-214.
- Bush, R. R., & Wilson, T. R. Two-choice behavior of paradise fish. *J. exp. Psychol.*, 1956, **51**, 315-322.
- Cane, Violet R. Behaviour sequences as semi-Markov chains. *J. Roy. Stat. Soc. (B)*, 1959, **21**, 36-58.
- Cane, Violet R. Review of R. D. Luce, Individual Choice Behavior. *J. Roy. Stat. Soc. (A)*, 1960, **22**, 486-488.
- Cane, Violet R. Some ways of describing behavior. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behavior*. Cambridge: Cambridge Univer. Press, 1961. Pp. 361-388.
- Cox, D. R. The regression analysis of binary sequences. *J. Roy. Stat. Soc. (B)*, 1958, **20**, 215-232.
- Edwards, W. Reward probability, amount, and information as determiners of sequential two-alternative decisions. *J. exp. Psychol.*, 1956, **52**, 177-188.
- Edwards, W. Probability learning in 1000 trials. *J. exp. Psychol.*, 1961, **62**, 385-394.
- Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.
- Estes, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: a study of a science. Vol. II. General systematic formulations, learning, and special processes*. New York: McGraw-Hill, 1959. Pp. 380-491.
- Estes, W. K. Learning theory. *Ann. Rev. Psychol.*, 1962, **13**, 107-144.
- Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.*, 1954, **47**, 225-234.
- Estes, W. K., & Suppes, P. Foundations of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 137-179.

- Feldman, J., & Newell, A. A note on a class of probability matching models. *Psychometrika*, 1961, **26**, 333-337.
- Feller, W. *An introduction to probability theory and its applications*, second edition. New York: Wiley, 1957.
- Galanter, E., & Bush, R. R. Some T-maze experiments. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 265-289.
- Goldberg, S. *Introduction to difference equations*. New York: Wiley, 1958.
- Gulliksen, H. A rational equation of the learning curve based on Thorndike's law of effect. *J. gen. Psychol.*, 1934, **11**, 395-434.
- Hanania, Mary I. A generalization of the Bush-Mosteller model with some significance tests. *Psychometrika*, 1959, **24**, 53-68.
- Hannan, E. J. *Time series analysis*. London: Methuen, 1960.
- Hayes, K. J., & Pereboom, A. C. Artifacts in criterion-reference learning curves. *Psychol. Rev.*, 1959, **66**, 23-26.
- Hodges, J. L., Jr. Fitting the logistic by maximum likelihood. *Biometrics*, 1958, **14**, 453-461.
- Hull, C. L. *Principles of behavior*. New York: Appleton-Century Crofts, 1943.
- Hull, C. L. *A behavior system*. New Haven: Yale Univer. Press, 1952.
- Irwin, F. W. On desire, aversion, and the affective zero. *Psychol. Rev.*, 1961, **68**, 293-300.
- Jarvik, M. E. Probability learning and a negative recency effect in the serial anticipation of alternating symbols. *J. exp. Psychol.*, 1951, **41**, 291-297.
- Kanal, L. Analysis of some stochastic processes arising from a learning model. Unpublished doctoral thesis, Univer. Penn., 1960.
- Kanal, L. A functional equation analysis of two learning models. *Psychometrika*, 1962, **27**, 89-104. (a)
- Kanal, L. The asymptotic distribution for the two-absorbing-barrier beta model. *Psychometrika*, 1962, **27**, 105-109. (b)
- Karlin, S. Some random walks arising in learning models. *Pacific J. Math.*, 1953, **3**, 725-756.
- Kendall, D. G. Stochastic processes and population growth. *J. Roy. Stat. Soc. (B)*, 1949, **11**, 230-264.
- Lamperti, J., & Suppes, P. Some asymptotic properties of Luce's beta learning model. *Psychometrika*, 1960, **25**, 233-241.
- Logan, F. A. *Incentive*. New Haven: Yale Univer. Press, 1960.
- Luce, R. D. *Individual choice behavior*. New York: Wiley, 1959.
- Luce, R. D. Some one-parameter families of commutative learning operators. In R. C. Atkinson (Ed.), *Studies in mathematical psychology, 1963*. Stanford: Stanford Univer. Press, 1963, in press.
- Miller, G. A., & McGill, W. J. A statistical description of verbal learning. *Psychometrika*, 1952, **17**, 369-396.
- Mosteller, F. Stochastic learning models. In *Proc. Third Berkeley Symp. Math. Stat. Probability*, 1955, **5**, Pp. 151-167.
- Mosteller, F., & Tatsuoka, M. Ultimate choice between two attractive goals: predictions from a model. *Psychometrika*, 1960, **25**, 1-17.
- Nicks, D. C. Prediction of sequential two-choice decisions from event runs. *J. exp. Psychol.*, 1959, **57**, 105-114.
- Restle, F. A theory of discrimination learning. *Psychol. Rev.*, 1955, **62**, 11-19.
- Restle, F. *Psychology of judgment and choice*. New York: Wiley, 1961.

- Sheffield, F. D. Avoidance training and the contiguity principle. *J. comp. physiol. Psychol.*, 1948, **41**, 165-177.
- Sternberg, S. H. A path-dependent linear model. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 308-339. (a)
- Sternberg, S. H. Application of four models to sequential dependence in human learning. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press: 1959. Pp. 340-380. (b)
- Tatsuoka, M., & Mosteller, F. A commuting-operator model. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 228-247.
- Thurstone, L. L. The learning function. *J. gen. Psychol.*, 1930, **3**, 469-491.
- Underwood, B. J. A graphical description of rote learning. *Psychol. Rev.*, 1957, **64**, 119-122.
- Wald, A. Asymptotic properties of the maximum-likelihood estimate of an unknown parameter of a discrete stochastic process. *Ann. math. Stat.*, 1948, **19**, 40-46.
- Weinstock, S. Unpublished experiment, 1955.
- Wilks, S. S. *Mathematical statistics*. New York: Wiley, 1962.

IO

*Stimulus Sampling Theory*¹

Richard C. Atkinson and
William K. Estes

Stanford University

1. Preparation of this chapter was supported in part by Contract Nonr-908(16) between the Office of Naval Research and Indiana University, Grant M-5184 from the National Institute of Mental Health to Stanford University, and Contract Nonr-225(17) between the Office of Naval Research and Stanford University.

Contents

| | |
|--|-----|
| 1. One-Element Models | 125 |
| 1.1. Learning of a single stimulus response association, | 126 |
| 1.2. Paired-associate learning, | 128 |
| 1.3. Probabilistic reinforcement schedules, | 141 |
| 2. Multi-Element Pattern Models | 153 |
| 2.1. General formulation, | 153 |
| 2.2. Treatment of the simple noncontingent case, | 162 |
| 2.3. Analysis of a paired-comparison learning experiment, | 181 |
| 3. A Component Model for Stimulus Compounding and Generalization | 191 |
| 3.1. Basic concepts; conditioning and response axioms, | 191 |
| 3.2. Stimulus compounding, | 193 |
| 3.3. Sampling axioms and major response theorem of fixed sample size model, | 198 |
| 3.4. Interpretation of stimulus generalization, | 200 |
| 4. Component and Linear Models for Simple Learning | 206 |
| 4.1. Component models with fixed sample size, | 207 |
| 4.2. Component models with stimulus fluctuation, | 219 |
| 4.3. The linear model as a limiting case, | 226 |
| 4.4. Applications to multiperson interactions, | 234 |
| 5. Discrimination Learning | 238 |
| 5.1. The pattern model for discrimination learning, | 239 |
| 5.2. A mixed model, | 243 |
| 5.3. Component models, | 249 |
| 5.4. Analysis of a signal detection experiment, | 250 |
| 5.5. Multiple-process models, | 257 |
| References | 265 |

Stimulus Sampling Theory

Stimulus sampling theory is concerned with providing a mathematical language in which we can state assumptions about learning and performance in relation to stimulus variables. A special advantage of the formulations to be discussed is that their mathematical properties permit application of the simple and elegant theory of Markov chains (Feller, 1957; Kemeny, Snell, & Thompson, 1957; Kemeny & Snell, 1959) to the tasks of deriving theorems and generating statistical tests of the agreement between assumptions and data. This branch of learning theory has developed in close interaction with certain types of experimental analyses; consequently it is both natural and convenient to organize this presentation around the theoretical treatments of a few standard reference experiments.

At the level of experimental interpretation most contemporary learning theories utilize a common conceptualization of the learning situation in terms of *stimulus*, *response*, and *reinforcement*. The stimulus term of this triumvirate refers to the environmental situation with respect to which behavior is being observed, the response term to the class of observable behaviors whose measurable properties change in some orderly fashion during learning, and the reinforcement term to the experimental operations or events believed to be critical in producing learning. Thus, in a simple paired-associate experiment concerned with the learning of English equivalents to Russian words, the stimulus might consist in presentation of the printed Russian word alone, the response measure in the relative frequency with which the learner is able to supply the English equivalent from memory, and reinforcement in paired presentation of the stimulus and response words.

In other chapters of this *Handbook*, and in the general literature on learning theory, the reader will encounter the notions of sets of responses and sets of reinforcing events. In the present chapter mathematical sets are used to represent certain aspects of the stimulus situation. It should be emphasized from the outset, however, that the mathematical models to be considered are somewhat abstract and that the empirical interpretations of stimulus sets and their elements are not to be considered fixed and immutable. Two main types of interpretations are discussed: in one the empirical correspondent of a stimulus element is the full pattern of stimulation effective on a given trial; in the other the correspondent of an

element is a component, or aspect, of the full pattern of stimulation. In the first, we speak of "pattern models" and in the second, of "component models" (Estes, 1959b).

There are a number of ways in which characteristics of the stimulus situation are known to affect learning and transfer. Rates and limits of conditioning and learning generally depend on stimulus magnitude, or intensity, and on stimulus variability from trial to trial. Retention and transfer of learning depend on the similarity, or communality, between the stimulus situations obtaining during training and during the test for retention or transfer. These aspects of the stimulus situation can be given direct and natural representations in terms of mathematical sets and relations between sets.

The basic notion common to all stimulus sampling theories is the conceptualization of the totality of stimulus conditions that may be effective during the course of an experiment in terms of a mathematical set. Although it is not a necessary restriction, it is convenient for mathematical reasons to deal only with finite sets, and this limitation is assumed throughout our presentation. Stimulus variability is taken into account by assuming that of the total population of stimuli available in an experimental situation generally only a part actually affects the subject on any one trial. Translating this idea into the terms of a stimulus sampling model, we may represent the total population by a set of "stimulus elements" and the stimulation effective on any one trial by a sample from this set. Many of the simple mathematical properties of the models to be discussed arise from the assumption that these trial samples are drawn randomly from the population, with all samples of a given size having equal probabilities. Although it is sometimes convenient and suggestive to speak in such terms, we should not assume that the stimulus elements are to be identified with any simple neurophysiological unit, as, for example, receptor cells. At the present stage of theory construction we mean to assume only that certain properties of the set-theoretical model represent certain properties of the process of stimulation. If these assumptions prove to be sufficiently well substantiated when the model is tested against behavioral data, then it will be in order to look for neurophysiological variables that might underlie the correspondences. Just as the ratio of sample size to population size is a natural way of representing stimulus variability, sample size per se may be taken as a correspondent of stimulus intensity, and the amount of overlap (i.e., proportion of common elements) between two stimulus sets may be taken to represent the degree of communality between two stimulus situations.

Our concern in this chapter is not to survey the rapidly developing area of stimulus sampling theory but simply to present some of the fundamental

mathematical techniques and illustrate their applications. For general background the reader is referred to Bush (1960), Bush & Estes (1959), Estes (1959a, 1962), and Suppes & Atkinson (1960). We shall consider first, and in some detail, the simplest of all learning models—the pattern model for simple learning. In this model the population of available stimuli is assumed to comprise a set of distinct stimulus patterns, exactly one of which is sampled on each trial. In the important special case of the one-element model it is assumed that there is only one such pattern and that it recurs intact at the beginning of each experimental trial. Granting that the one-element model represents a radical idealization of even the most simplified conditioning situations, we shall find that it is worthy of study not only for expositional purposes but also for its value as an analytic device in relation to certain types of learning data. After a relatively thorough treatment of pattern models for simple acquisition and for learning under probabilistic reinforcement schedules, we shall take up more briefly the conceptualization of generalization and transfer; component models in which the patterns of stimulation effective on individual trials are treated not as distinct elements but as overlapping samples from a common population; and, finally, some examples of the more complex multiple-process models that are becoming increasingly important in the analysis of discrimination learning, concept formation, and related phenomena.

1. ONE-ELEMENT MODELS

We begin by considering some one-element models that are special cases of the more general theory. These examples are especially simple mathematically and provide us with the opportunity to develop some mathematical tools that will be necessary in later discussions. Application of these models is appropriate when the stimulus situation is sufficiently stable from trial to trial that it may be theoretically represented (to a good approximation) by a single stimulus element which is sampled with probability 1 on each trial. At the start of a trial the element is in one of several possible conditioning states; it may or may not remain in this conditioning state, depending on the reinforcing event for that trial. In the first part of this section we consider a model for paired-associate learning. In the second part we consider a model for a two-choice learning situation involving a probabilistic reinforcement schedule. The models generate some predictions that are undoubtedly incorrect, except possibly under ideal experimental conditions; nevertheless, they provide a useful introduction to more general cases which we pursue in Section 2.

1.1 Learning of a Single Stimulus-Response Association

Imagine the simplest possible learning situation. A single stimulus pattern, S , is to be presented on each of a series of trials and each trial is to terminate with reinforcement of some designated response, the "correct response" in this situation. According to stimulus sampling theory, learning occurs in an all-or-none fashion with respect to S .

1. If the correct response is not originally conditioned to ("connected to") S , then, until learning occurs, the probability of the correct response is zero.
2. There is a fixed probability c that the reinforced response will become conditioned to S on any trial.
3. Once conditioned to S , the correct response occurs with probability 1 on every subsequent trial.

These assumptions constitute the simplest case of the "one-element pattern model." Learning situations that completely meet the specifications laid down above are as unlikely to be realized in psychological experiments as perfect vacuums or frictionless planes in the physics laboratory. However, reasonable approximations to these conditions can be attained. The requirement that the same stimulus pattern be reproduced on each trial is probably fairly well met in the standard paired-associate experiment with human subjects. In one such experiment, conducted in the laboratory of one of the writers (W. K. E.), the stimulus member of each item was a trigram and the correct response an English word, for example,

| | |
|-----|-------|
| S | R . |
| xvk | house |

On a reinforced trial the stimulus and response members were exposed together, as shown. Then, after several such items had received a single reinforcement, each of the stimuli was presented alone, the subject being instructed to give the correct response from memory, if he could. Then each item was given a second reinforcement, followed by a second test, and so on.

According to the assumptions of the one-element pattern model, a subject should be expected to make an incorrect response on each test with a given stimulus until learning occurs, then a correct response on every subsequent trial; if we represent an error by a 1 and a correct response by a 0, the protocol for an individual item over a series of trials should, then, consist in a sequence of 0's preceded in most cases by a sequence of 1's. Actual protocols for several subjects are shown below:

| | | | | | | | | | | |
|----------|---|---|---|---|---|---|---|---|---|---|
| <i>a</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>b</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>c</i> | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>d</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>e</i> | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>f</i> | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>g</i> | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>h</i> | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| <i>i</i> | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

The first seven of these correspond perfectly to the idealized theoretical picture; the last two deviate slightly. The proportion of "fits" and "misfits" in this sample is about the same as in the full set of 80 cases from which the sample was taken. The occasional lapses, that is, errors following correct responses, may be symptomatic of a forgetting process that should be incorporated into the theory, or they may be simply the result of minor uncontrolled variables in the experimental situation which are best ignored for theoretical purposes. Without judging this issue, we may conclude that the simple one-element model at least merits further study.

Before we can make quantitative predictions we need to know the value of the conditioning parameter c . Statistical learning theory includes no formal axioms that specify precisely what variables determine the value of c , but on the basis of considerable experience we can safely assume that this parameter will vary with characteristics of the populations of subjects and items represented in a particular experiment. An estimate of the value of c for the experiment under consideration is easy to come by. In the full set of 80 cases (40 subjects, each tested on two items) the proportion of correct responses on the test given after a single reinforcement was 0.39. According to the model, the probability is c that a reinforced response will become conditioned to its paired stimulus; consequently c is the expected proportion of successful conditionings out of 80 cases, and therefore the expected proportion of correct responses on the subsequent test. Thus we may simply take the observed proportion 0.39 as an estimate of c .

In order to test the model, we need now to derive theoretical expressions for other aspects of the data. Suppose we consider the sequences of correct and incorrect responses, 000, 001, etc., on the first three trials. According to the model, a correct response should never be followed by an error, so the probability of the sequence 000 is simply c , and the probabilities of 001, 010, 011, and 101 are all zero. To obtain an error on the first trial followed by a correct response on the second, conditioning must fail on the first reinforcement but occur on the second, and this joint event has

probability $(1 - c)c$. Similarly, the probability that the first correct response will occur on the third trial is given by $(1 - c)^2c$ and the probability of no correct response in three trials by $(1 - c)^3$. Substituting the estimate 0.39 for c in each of these expressions, we obtain the predicted

Table 1 Observed and Predicted (One-Element Model) Values for Response Sequences Over First Three Trials of a Paired-Associate Experiment

| Sequence* | Observed Proportions | Theoretical Proportions |
|-----------|----------------------|-------------------------|
| 000 | 0.36 | 0.39 |
| 001 | 0.02 | 0 |
| 010 | 0.01 | 0 |
| 011 | 0 | 0 |
| 100 | 0.27 | 0.24 |
| 101 | 0 | 0 |
| 110 | 0.11 | 0.14 |
| 111 | 0.23 | 0.23 |

* 0 = correct response
1 = error

values which are compared with the corresponding empirical values for this experiment in Table 1. The correspondences are seen to be about as close as could be expected with proportions based on 80 response sequences.

1.2 Paired-Associate Learning

In order to apply the one-element model to paired-associate experiments involving fixed lists of items, it is necessary to adjust the "boundary conditions" appropriately. Consider, for example, an experiment reported by Estes, Hopkins, and Crothers (1960). The task assigned their subjects was to learn associations between the numbers 1 through 8, serving as responses, and eight consonant trigrams, serving as stimuli. Each subject was given two practice trials and two test trials. On the first practice trial the eight syllable-number pairs were exhibited singly in a random order. Then a test was given, the syllables alone being presented singly in a new random order and the subjects attempting to respond to each syllable with the correct number. Four of the syllable-number pairs were presented on a second practice trial, and all eight syllables were included in a final test trial.

In writing an expression for the probability of a correct response on the first test in this experiment, we must take account of the fact that, after the first practice trial, the subjects knew that the responses were the numbers 1 to 8 and were in a position to guess at the correct answers when shown syllables that they had not yet learned. The minimum probability of achieving a correct response to an unlearned item by guessing would be $\frac{1}{8}$. Thus we would have for p_0 , the probability of a correct response on the first test,

$$p_0 = c + \frac{1 - c}{8},$$

that is, the probability c that the correct association was formed plus the probability $(1 - c)/8$ that the association was not formed but the correct response was achieved by guessing. Setting this expression equal to the observed proportion of correct responses on the first trial for the twice reinforced items, we readily obtain an estimate of c for these experimental conditions,

$$0.404 = c + (1 - c)(0.125),$$

and so

$$\hat{c} = 0.32.$$

Now we can proceed to derive expressions for the joint probabilities of various combinations of correct and incorrect responses on the first and second tests for the twice reinforced items. For the probability of correct responses to a given item in both tests, we have

$$p_{00} = c + (1 - c)(0.125)c + (1 - c)^2(0.125)^2.$$

With probability c , conditioning occurs on the first reinforced trial, and then correct responses necessarily occur on both tests; with probability $(1 - c)c(0.125)$, conditioning does not occur on the first reinforced trial but does on the second, and a correct response is achieved by guessing on the first test; with probability $(1 - c)^2(0.125)^2$, conditioning occurs on neither reinforced trial but correct responses are achieved by guessing on both tests. Similarly, we obtain

$$p_{01} = (1 - c)^2(0.875)(0.125)$$

$$p_{10} = (1 - c)(0.875)[c + (1 - c)(0.125)]$$

and

$$p_{11} = (1 - c)^2(0.875)^2.$$

Substituting for c in these expressions the estimate computed above, we

arrive at the predicted values which we compare with the corresponding observed values below.

| | Observed | Predicted |
|----------|----------|-----------|
| p_{00} | 0.35 | 0.35 |
| p_{01} | 0.05 | 0.05 |
| p_{10} | 0.27 | 0.24 |
| p_{11} | 0.33 | 0.35 |

Although this comparison reveals some disparities, which we might hope to reduce with a more elaborate theory, it is surprising, to the writers at least, that the patterns of observed response proportions in both experiments considered can be predicted as well as they are by such an extremely simple model.

Ordinarily, experiments concerned with paired-associate learning are not limited to a couple of trials, like those just considered, but continue until the subjects meet some criterion of learning. Under these circumstances it is impractical to derive theoretical expressions for all possible sequences of correct and incorrect responses. A reasonable goal, instead, is to derive expressions for various statistics that can be conveniently computed for the data of the standard experiment; examples of such statistics are the mean and variance of errors per item, frequencies of runs of errors or correct responses, and serial correlation of errors over trials with any given lag. Bower (1961, 1962) carried out the first major analysis of this type for the one-element model. We shall use some of his results to illustrate application of the model to a full "learning-to-criterion" experiment. Essential details of his experiment are as follows: a list of 10 items was learned by 29 undergraduates to a criterion of two consecutive errorless trials. The stimuli were different pairs of consonant letters and the responses were the integers 1 and 2; each response was assigned as correct to a randomly selected five items for each subject. A response was obtained from the subject on each presentation of an item, and he was informed of the correct answer following his response.

As in the preceding application, we shall assume that each item in the list is to be represented theoretically by exactly one stimulus element, which is sampled with probability 1 when the item is presented, and that the correct response to that item is conditioned in an all-or-none fashion. On trial n of the experiment an element is in one of two "conditioning states": In state C the element is conditioned to the correct response; in state \bar{C} the element is not conditioned.

The response the subject makes depends on his conditioning state.

When the element is in state C , the correct response occurs with probability 1. The probability of the correct response when the element is in state \bar{C} depends on the experimental procedure. In Bower's experiment the subjects were told the r responses available to them and each occurred equally often as the to-be-learned response. Therefore we may assume that in the unconditioned state the probability of a correct response is $1/r$, where r is the number of alternative responses.

The conditioning assumptions can readily be restated in terms of the conditioning states:

1. On any reinforced trial, if the sampled element is in state \bar{C} , it has probability c of going into state C .
2. The parameter c is fixed in value in a given experiment.
3. Transitions from state C to state \bar{C} have probability zero.

We shall now derive some predictions from the model and compare these with observed data. The data of particular interest will be a subject's sequence of correct and incorrect responses to a specific stimulus item over trials. Similarly, in deriving results from the model we shall consider only an isolated stimulus item and its related sequence of responses. However, when we apply the model to data, we assume that all items in the list are comparable, that is, all items have the same conditioning parameter c and all items start out in the same conditioning state (\bar{C}). Consequently the response sequence associated with any given item is viewed as a sample of size 1 from a population of sequences all generated by the same underlying process.

A feature of this model which makes it especially tractable for purposes of deriving various statistics is the fact that the sequences of transitions between states C and \bar{C} constitute a Markov chain. This means that, given the state on any one trial, we can specify the probability of each state on the next trial without regard to the previous history. If we represent by C_n and \bar{C}_n the events that an item is in the conditioned or unconditioned state, respectively, on trial n , and by q_{11} and q_{21} the probabilities of transitions from state C to state C and from \bar{C} to C , respectively, the conditioning assumptions lead directly to the relations²

$$\begin{aligned} q_{11} &= \Pr(C_{n+1} | C_n) = 1, \\ q_{21} &= \Pr(C_{n+1} | \bar{C}_n) = c, \end{aligned}$$

² See Feller (1957) for a discussion of conditional probabilities. In brief, if H_1, \dots, H_n are a set of mutually exclusive events of which one necessarily occurs, then any event A can occur only in conjunction with some H_j . Since the AH_j are mutually exclusive, their probabilities add. Applying the well-known theorem on compound probabilities, we obtain $\Pr(A) = \sum_j \Pr(AH_j) = \sum_j \Pr(A | H_j) \Pr(H_j)$.

and

$$Q = \begin{bmatrix} 1 & 0 \\ c & 1 - c \end{bmatrix},$$

where Q is the matrix of one-step transition probabilities, the first row and column referring to C and the second row and column to \bar{C} . Now the matrix of probabilities for transitions between any two states in n trials is simply the n th power of Q , as may be verified by mathematical induction (see, e.g., Kemeny, Snell, & Thompson, 1957, p. 327),

$$Q^n = \begin{bmatrix} 1 & 0 \\ 1 - (1 - c)^n & (1 - c)^n \end{bmatrix}.$$

Henceforth we shall assume that all stimulus elements are in state \bar{C} at the onset of the first trial of our experiment. Given that the state is \bar{C} on trial 1, the probability of being in state \bar{C} at the start of trial n is $(1 - c)^{n-1}$, which goes to 0 as n becomes large, for $c > 0$. Thus with probability 1 the subject is eventually to be found in the conditioned state.

Next we prove some theorems about the observable sequence of correct and incorrect responses in terms of the underlying sequence of unobservable conditioning states. We define the response random variable

$$A_n = \begin{cases} 0 & \text{if a correct response occurred on trial } n, \\ 1 & \text{if an error occurred on trial } n. \end{cases}$$

By our assumed response rule the probabilities of an error, given that the subject is in the conditioned or unconditioned state, respectively, are

$$\Pr(A_n = 1 \mid C_n) = 0$$

and

$$\Pr(A_n = 1 \mid \bar{C}_n) = 1 - \frac{1}{r}.$$

To obtain the probability of an error on trial n , namely $\Pr(A_n = 1)$, we sum these conditional probabilities weighted by the probabilities of being in the respective states:

$$\begin{aligned} \Pr(A_n = 1) &= \Pr(A_n = 1 \mid C_n) \Pr(C_n) + \Pr(A_n = 1 \mid \bar{C}_n) \Pr(\bar{C}_n) \\ &= \left(1 - \frac{1}{r}\right) (1 - c)^{n-1}. \end{aligned} \quad (1)$$

Consider next the infinite sum of the random variables A_1, A_2, A_3, \dots which we denote \bar{A} ; specifically,

$$\bar{A} = \sum_{n=1}^{\infty} A_n.$$

But

$$\begin{aligned}
 E(\bar{A}) &= \sum E(A_n) \\
 &= \sum \Pr(A_n = 1) \\
 &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{r}\right) (1 - c)^{n-1} \\
 &= \frac{1 - (1/r)}{c}.
 \end{aligned} \tag{2}$$

Thus the number of errors expected during the learning of any given item is given by Eq. 2.

Equation 2 provides an easy method for estimating c . For any given subject we can obtain his average number of errors over stimulus items, equate this number to the right-hand side of Eq. 2 with $r = 2$, and solve for c . We thereby obtain an estimate of c for each subject, and intersubject differences in learning are reflected in the variability of these estimates. Bower, in analyzing his data, chose to assume that c was the same for all subjects; thus he set $E(\bar{A})$ equal to the observed number of errors averaged over both list items and subjects and obtained a single estimate of c . This group estimate of c simplifies the computations involved in generating predictions. However, it has the disadvantage that a discrepancy between observed and predicted values may arise as a consequence of assuming equal c 's when, in fact, the theory is correct but c varies from subject to subject. Fortunately, Bower has obtained excellent agreement between theory and observation using the group estimate of c and, for the particular conditions he investigated, any increase in precision that might be achieved by individual estimates of c does not seem crucial.

For the experiment described above, Bower reports 1.45 errors per stimulus item averaged over all subjects. Equating $E(\bar{A})$ in Eq. 2 to 1.45, with $r = 2$, we obtain the estimate $c = 0.344$. All predictions that we derive from the model for this experiment will be based on this single estimate of c . It should be remarked that the estimate of c in terms of Eq. 2 represents only one of many methods that could have been used. The method one selects depends on the properties of the particular estimator (e.g., whether the estimator is unbiased and efficient in relation to other estimators). Parameter estimation is a theory in its own right, and we shall not be able to discuss the many problems involved in the estimation of learning parameters. The reader is referred to Suppes & Atkinson (1960) for a discussion of various methods and their properties. Associated with this topic is the problem of assessing the statistical agreement between data and theory (i. e., the goodness-of-fit between predicted and observed values) once parameters have been estimated. In our analysis of data

in this chapter we offer no statistical evaluation of the predictions but simply display the results for the reader's inspection. Our reason is that we present the data only to illustrate features of the theory and its application; these results are not intended to provide a test of the model. However, in rigorous analyses of such models the problem of goodness-of-fit is extremely important and needs careful consideration. Here again the reader is referred to Suppes & Atkinson (1960) for a discussion of some of the problems and possible statistical tests.

By using Eq. 1 with the estimate of c obtained above we have generated the predicted learning curve presented in Fig. 1. The fit is sufficiently close that most of the predicted and observed points cannot be distinguished on the scale of the graph.

As a basis for the derivation of other statistics of total errors, we require an expression for the probability distribution of \bar{A} . To obtain this, we note first that the probability of no errors at all occurring during learning is given by

$$c\left(\frac{1}{r}\right) + (1-c)\left(\frac{1}{r}\right)^2 c + \dots = \frac{c}{r} \sum_{i=0}^{\infty} \left(\frac{1-c}{r}\right)^i = \frac{c}{r[1 - (1-c)/r]} = \frac{b}{r},$$

where $b = c/[1 - (1-c)/r]$. This event may arise if a correct response occurs by guessing on the first trial and conditioning occurs on the first reinforcement, if a correct response occurs by guessing on the first two

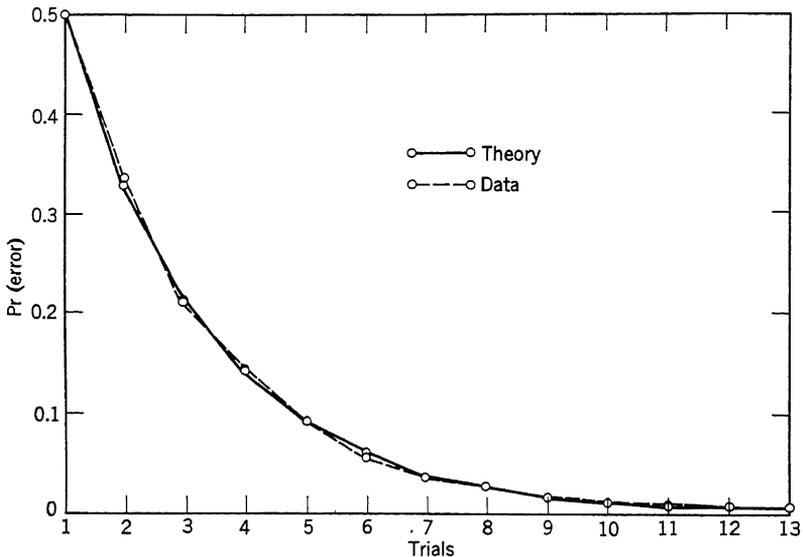


Fig. 1. The average probability of an error on trial n in Bower's paired-associate experiment.

trials and conditioning occurs on the second reinforcement, and so on. Similarly, the probability of no additional errors following an error on any given trial is given by

$$c + c \frac{1-c}{r} + \dots = c \sum_{i=0}^{\infty} \left(\frac{1-c}{r} \right)^i = \frac{c}{1-(1-c)/r} = b.$$

To have exactly k errors, we must have a first error (if $k > 0$), which has probability $1 - b/r$, $k - 1$ additional errors, each of which has probability $1 - b$, and then no more errors. Therefore the required probability distribution is

$$\Pr(\bar{A} = 0) = \frac{b}{r} \tag{3}$$

$$\Pr(\bar{A} = k) = b(1 - b/r)(1 - b)^{k-1}, \quad \text{for } k \geq 1.$$

Equation 3 can be applied to data directly to predict the form of the frequency distribution of total errors. It may also be utilized in deriving, for example, the variance of this distribution. Preliminary to computing the variance, we need the expectation of \bar{A}^2 ,

$$\begin{aligned} E(\bar{A}^2) &= \sum_{k=0}^{\infty} k^2 b \left(\frac{r-b}{r} \right) (1-b)^{k-1} \\ &= b \left(\frac{r-b}{r} \right) \sum_{k=0}^{\infty} [k(k-1) + k] (1-b)^{k-1} \\ &= (1-b)b \left(\frac{r-b}{r} \right) \sum_{k=0}^{\infty} [k(k-1) + k] (1-b)^{k-2}, \end{aligned}$$

where the second step is taken in order to facilitate the summation. Using the familiar expression

$$\sum_{k=0}^{\infty} (1-b)^k = \frac{1}{b}$$

for the sum of a geometric series, together with the relations

$$\frac{d}{db} (1-b)^k = -k(1-b)^{k-1},$$

$$\frac{d^2}{db^2} (1-b)^k = k(k-1)(1-b)^{k-2},$$

and

$$-\sum_{k=0}^{\infty} \frac{d}{db} (1-b)^k = -\frac{d}{db} \sum_{k=0}^{\infty} (1-b)^k = -\frac{d}{db} \left(\frac{1}{b} \right) = \frac{1}{b^2},$$

$$\sum_{k=0}^{\infty} \frac{d^2}{db^2} (1-b)^k = \frac{d^2}{db^2} \sum_{k=0}^{\infty} (1-b)^k = \frac{d^2}{db^2} \left(\frac{1}{b} \right) = \frac{2}{b^3},$$

we obtain

$$E(\bar{A}^2) = b \left(\frac{r-b}{r} \right) \left[\frac{2(1-b)}{b^3} + \frac{1}{b^2} \right]$$

and

$$\begin{aligned} \text{Var}(\bar{A}) &= E(\bar{A}^2) - [E(\bar{A})]^2 \\ &= b \left(\frac{r-b}{r} \right) \left[\frac{2(1-b)}{b^3} + \frac{1}{b^2} \right] - \frac{[1 - (1/r)]^2}{c^2} \\ &= \left(1 - \frac{1}{r} \right) \frac{(2c - cr + r - 1)}{c^2 r} \\ &= \frac{(r-1)(2c - cr + r - 1)}{rc} \\ &= \frac{(r-1)(cr + 2c - 2cr + r - 1)}{rc} \\ &= \frac{(r-1)}{rc} \left[1 + \frac{(2c-1)(1-r)}{rc} \right] \\ &= E(\bar{A})[1 + E(\bar{A})(1-2c)]. \end{aligned} \tag{4}$$

Inserting in Eq. 4 the estimates $E(\bar{A}) = 1.45$ and $c = 0.344$ from Bower's data, we obtain 1.44 for the predicted standard deviation of total errors, which may be compared with the observed value of 1.37.

Another useful statistic of the error sequence is $E(\mathbf{A}_n \mathbf{A}_{n+k})$; namely, the expectation of the product of error random variables on trials n and $n+k$. This quantity is related to the autocorrelation between errors on trials $n+k$ and trial n . By elementary probability theory,

$$\begin{aligned} E(\mathbf{A}_n \mathbf{A}_{n+k}) &= E(\mathbf{A}_{n+k} | \mathbf{A}_n) E(\mathbf{A}_n) \\ &= \Pr(\mathbf{A}_{n+k} = 1 | \mathbf{A}_n = 1) \Pr(\mathbf{A}_n = 1). \end{aligned}$$

But for an error to occur on trial $n+k$ conditioning must have failed to occur during the intervening k trials and the subject must have guessed incorrectly on trial $n+k$. Hence

$$\Pr(\mathbf{A}_{n+k} = 1 | \mathbf{A}_n = 1) = (1-c)^k \left(1 - \frac{1}{r} \right).$$

Substitution of this result into the preceding expression, along with the result presented in Eq. 1, yields the following expression:

$$\begin{aligned} E(\mathbf{A}_n \mathbf{A}_{n+k}) &= \left(1 - \frac{1}{r} \right) (1-c)^k (1-c)^{n-1} \left(1 - \frac{1}{r} \right) \\ &= \left(1 - \frac{1}{r} \right)^2 (1-c)^{n+k-1}. \end{aligned} \tag{5}$$

A convenient statistic for comparison with data (directly related to the average autocorrelation of errors with lag k , but easier to compute) is obtained by summing the cross product of A_n and A_{n+k} over all trials. We define c_k as the mean of this random variable, where

$$\begin{aligned} c_k &= \sum_{n=1}^{\infty} E(A_{n+k}A_n) \\ &= E(\bar{A}) \left(1 - \frac{1}{r}\right) (1 - c)^k. \end{aligned} \quad (6)$$

To be explicit, consider the following response protocol running in time from left to right: 1101010010000. The observed values for c_k are $c_1 = 1$, $c_2 = 2$, $c_3 = 2$, and so on.

The predictions for c_1 , c_2 , and c_3 computed from the c estimate given above were 0.479, 0.310, and 0.201. Bower's observed values were 0.486, 0.292, and 0.187.

Next we consider the distribution of the number of errors between the k th and $(k + 1)$ st success. The methods to be used in deriving this result are general and can be used to derive the distribution of errors between the k th and $(k + m)$ th success for any nonnegative integer m . The only limitation is that the expressions become unwieldy as m increases. We shall define J_k as the random variable for the number of errors between the k th and $(k + 1)$ st success; its values are 0, 1, 2, An error following the k th success can occur only if the k th success itself occurs as a result of guessing; that is, the subject necessarily is in state \bar{C} when the k th success occurs. Letting g_k denote the probability that the k th success occurs by guessing, we can write the probability distribution

$$\Pr(J_k = i) = \begin{cases} 1 - \alpha g_k & \text{for } i = 0 \\ (1 - \alpha)\alpha^i g_k & \text{for } i > 0, \end{cases} \quad (7)$$

where $\alpha = (1 - c)[1 - (1/r)]$. To obtain $\Pr(J_k = 0)$, we note that 0 errors can occur in one of three ways: (1) the k th success occurs because the subject is in state C (which has probability $1 - g_k$) and necessarily a correct response occurs on the next trial; (2) the k th success occurs by guessing, the subject remaining in state \bar{C} and again guessing correctly on the next trial [which has probability $g_k(1 - c)(1/r)$]; or (3) the k th success occurs by guessing but conditioning is effective on the trial (which has probability $g_k c$). Thus $\Pr(J_k = 0) = 1 - g_k + g_k(1 - c)(1/r) + g_k c = 1 - \alpha g_k$. The event of i errors ($i > 0$) between the k th and $(k + 1)$ st successes can occur in one of two ways: (1) the k th and $(k + 1)$ st successes occur by guessing {with probability $g_k(1 - c)^{i+1}[1 - (1/r)]^i(1/r)$ }, or (2)

the k th success occurs by guessing and conditioning does not take place until the trial immediately preceding the $(k + 1)$ st success {with probability $g_k(1 - c)^i[1 - (1/r)]^i c$ }. Hence

$$\begin{aligned} \Pr(\mathbf{J}_k = i) &= g_k(1 - c)^{i+1} \left(1 - \frac{1}{r}\right)^i \frac{1}{r} + g_k(1 - c)^i \left(1 - \frac{1}{r}\right)^i c \\ &= g_k \left(1 - \frac{1}{r}\right)^i (1 - c)^i \left[c + \frac{1}{r}(1 - c)\right] \\ &= g_k \alpha^i (1 - \alpha). \end{aligned}$$

From Eq. 7 we may obtain the mean and variance of \mathbf{J}_k , namely

$$E(\mathbf{J}_k) = \sum_{i=0}^{\infty} i \Pr(\mathbf{J}_k = i) = \frac{\alpha g_k}{1 - \alpha}, \quad (8)$$

and

$$\begin{aligned} \text{Var}(\mathbf{J}_k) &= \sum_{i=0}^{\infty} i^2 \Pr(\mathbf{J}_k = i) - E(\mathbf{J}_k)^2 \\ &= \frac{\alpha g_k (1 + \alpha)}{(1 - \alpha)^2} - \frac{\alpha^2 g_k^2}{(1 - \alpha)^2} \\ &= \frac{\alpha g_k}{(1 - \alpha)^2} [1 + \alpha(1 - g_k)]. \end{aligned} \quad (9)$$

In order to evaluate these quantities, we require an expression for g_k . Consider g_1 , the probability that the first success will occur by guessing. It could occur in one of the following ways: (1) the subject guesses correctly on trial 1 (with probability $1/r$); (2) the subject guesses incorrectly on trial 1, conditioning does not occur, and the subject guesses successfully on trial 2 {this joint event has probability $[1 - (1/r)](1 - c)(1/r)$ }; or (3) conditioning does not occur on trials 1 and 2, and the subject guesses incorrectly on both of these trials but guesses correctly on trial 3 {with probability $[1 - (1/r)]^2(1 - c)^2(1/r)$ }, and so forth. Thus

$$\begin{aligned} g_1 &= \frac{1}{r} + \left(1 - \frac{1}{r}\right)(1 - c) \frac{1}{r} + \left(1 - \frac{1}{r}\right)^2 (1 - c)^2 \frac{1}{r} + \dots \\ &= \frac{1}{r} \sum_{i=0}^{\infty} \left(1 - \frac{1}{r}\right)^i (1 - c)^i \\ &= \frac{1}{(1 - \alpha)r}. \end{aligned}$$

Now consider the probability that the k th success occurs by guessing for $k > 1$. In order for this event to occur it must be the case that (1) the $(k - 1)$ st success occurs by guessing, (2) conditioning fails to occur on the

trial of the $(k - 1)$ st success, and (3) since the subject is assumed to be in state \bar{C} on the trial following the $(k - 1)$ st success, the next correct response occurs by guessing, which has probability g_1 . Hence

$$g_k = g_{k-1}(1 - c)g_1.$$

Solving this difference equation³ we obtain

$$g_k = (1 - c)^{k-1}g_1^k.$$

Finally, substituting the expression obtained for g_1 yields

$$g_k = \frac{(1 - c)^{k-1}}{(r - \alpha r)^k}. \quad (10)$$

We may now combine Eqs. 7 and 10, inserting our original estimate of c , to obtain predictions about the number of errors between the k th and $(k + 1)$ st success in Bower's data. To illustrate, for $k = 1$, the predicted mean is 0.361 and the observed value is 0.350.

To conclude our analysis of this model, we consider the probability p_k that a response sequence to a stimulus item will exhibit the property of no errors following the k th success. This event can occur in one of two ways: (1) the k th success occurs when the subject is in state C (the probability of which is $1 - g_k$), or (2) the k th success occurs when the subject is in state \bar{C} and no errors occur on subsequent trials. Let b denote the probability of no more errors following a correct guess. Then

$$\begin{aligned} p_k &= (1 - g_k) + g_k b \\ &= 1 - g_k(1 - b). \end{aligned} \quad (11)$$

But the probability of no more errors following a successful guess is simply

$$\begin{aligned} b &= c + (1 - c)\frac{1}{r}c + (1 - c)^2\left(\frac{1}{r}\right)^2c + \dots \\ &= \frac{c}{\alpha + c}. \end{aligned}$$

Substituting this result for b into Eq. 11, along with our expression for g_k in Eq. 10, we obtain

$$p_k = 1 - \frac{\alpha(1 - c)^{k-1}}{(\alpha + c)(r - \alpha r)^k}. \quad (12)$$

Observed and predicted values of p_k for Bower's experiment are shown in Table 2.

³ The solution of this equation can quickly be obtained. Note that $g_2 = g_1(1 - c)g_1 = (1 - c)g_1^2$. Similarly, $g_3 = g_2(1 - c)g_1$; substituting the result for g_2 , we obtain $g_3 = (1 - c)g_1^2(1 - c)g_1 = (1 - c)^2g_1^3$. If we continue in this fashion, it will be obvious that $g_k = (1 - c)^{k-1}g_1^k$.

We shall not pursue more consequences of this model.⁴ The particular results we have examined were selected because they illustrated fundamental features of the model and also introduced mathematical techniques that will be needed later. In Bower's paper more than 30 predictions of the type presented here were tested, with results comparable to those exhibited above. The goodness-of-fit of theory to data in these instances is quite

Table 2 Observed and Predicted Values for p_k , the Probability of No Errors Following the k th Success

| k | Observed p_k | Predicted p_k |
|-----|----------------|-----------------|
| 0 | 0.255 | 0.256 |
| 1 | 0.628 | 0.636 |
| 2 | 0.812 | 0.822 |
| 3 | 0.869 | 0.912 |
| 4 | 0.928 | 0.957 |
| 5 | 0.963 | 0.979 |
| 6 | 0.973 | 0.990 |
| 7 | 0.990 | 0.995 |
| 8 | 0.990 | 0.997 |
| 9 | 0.993 | 0.998 |
| 10 | 0.996 | 0.999 |
| 11 | 1.000 | 1.000 |

(Interpret p_0 as the probability of no errors at all during the course of learning).

representative of what we may now expect to obtain routinely in simple learning experiments when experimental conditions have been appropriately arranged to approximate the simplifying assumptions of the mathematical model.

Concepts of the sort developed in this section can be extended to more traditional types of verbal learning situations involving stimulus similarity, meaningfulness, and the like. For example, Atkinson (1957) has presented a model for rote serial learning which is based on similar ideas and deals

⁴ Bower also has compared the one-element model with a comparable single-operator linear model presented by Bush and Sternberg (1959). The linear model assumes that the probability of an incorrect response on trial n is a fixed number p_n , where $p_{n+1} = (1 - c)p_n$ and $p_1 = [1 - (1/r)]$. The one-element model and the linear model generate many identical predictions (e.g., mean learning curve), and it is necessary to look at the finer structure of the data to differentiate models. Among the 20 possible comparisons Bower makes between the two models, he finds that the one-element model comes closer to the data on 18.

with such variables as intertrial interval, list length, and types of errors (perseverative, anticipatory, or response-failure). Unfortunately, theoretical analyses of this sort for traditional experimental routines often lead to extremely complicated mathematical models with the result that only a few consequences of the axioms can be derived. Stated differently, a set of concepts may be general in terms of the range of situations to which it is applicable; nevertheless, in order to provide rigorous and detailed tests of these concepts, it is frequently necessary to contrive special experimental routines in which the theoretical analyses generate tractable mathematical systems.

1.3 Probabilistic Reinforcement Schedules

We shall now examine a one-element model for some simple two-choice learning problems. The one-element model for this situation, as contrasted with the paired-associate model, generates some predictions of behavior that are quite unrealistic, and for this reason we defer an analysis of experimental data until we consider comparable multi-element processes. The reason for presenting the one-element model is that it represents a convenient introduction to multi-element models and permits us to develop some mathematical tools in a simple fashion. Further, when we do discuss multi-element models, we shall employ a rather restrictive set of conditioning axioms. However, for the one-element model we may present an extremely general set of conditioning assumptions without getting into too much mathematical complexity. Therefore the analysis of the one-element case will suggest lines along which the multi-element models can be generalized.

The reference experiment (see, e.g., Estes & Straughan, 1954; Suppes & Atkinson, 1960) involves a long series of discrete trials. Each trial is initiated by the onset of a signal. To the signal the subject is required to make one of two responses which we denote A_1 and A_2 . The trial is terminated with an E_1 or E_2 reinforcing event; the occurrence of E_i indicates that response A_i was the correct response for that trial. Thus in a human learning situation the subject is required on each trial to predict the reinforcing event he expects will occur by making the appropriate response—an A_1 if he expects E_1 and an A_2 if he expects E_2 ; at the end of the trial he is permitted to observe which event actually occurred. Initially the subject may have no preference between responses, but as information accrues to him over trials his pattern of choices undergoes systematic changes. The role of a model is to predict the detailed features of these changes.

The experimenter may devise various schedules for determining the sequence of reinforcing events over trials. For example, the probability of an E_1 may be (1) some function of the trial number, (2) dependent on previous responses of the subject, (3) dependent on the previous sequence of reinforcing events, or (4) some combination of the foregoing. For simplicity we consider a *noncontingent* reinforcement schedule. The case is defined by the condition that the probability of E_1 is constant over trials and independent of previous responses and reinforcements. It is customary in the literature to call this probability π ; thus $\Pr(E_{1,n}) = \pi$ for all n . Here we are denoting by $E_{i,n}$ the event that reinforcement E_i occurs on trial n . Similarly, we shall represent by $A_{i,n}$ the event that response A_i occurs on trial n .

We assume that the stimulus situation comprising the signal light and the context in which it occurs can be represented theoretically by a single stimulus element that is sampled with probability 1 when the signal occurs. At the start of a trial the element is in one of three conditioning states: in state C_1 the element is conditioned to the A_1 -response and in state C_2 to the A_2 -response; in state C_0 the element is not conditioned to A_1 or to A_2 . The response rules are similar to those presented earlier. When the subject is in C_1 or C_2 , the A_1 - or A_2 -response occurs with probability 1. In state C_0 we assume that either response will be elicited equiprobably; that is, $\Pr(A_{1,n} | C_{0,n}) = \frac{1}{2}$. For some subjects a response bias may exist that would require the assumption $\Pr(A_{1,n} | C_{0,n}) = \beta$, where $\beta \neq \frac{1}{2}$. For these subjects it would be necessary to estimate β when applying the model. However, for simplicity we shall pursue only the case in which responses are equiprobable when the subject is in C_0 .

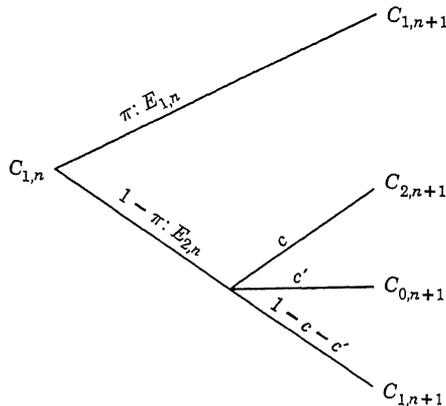


Fig. 2. Branching process, starting from state C_1 on trial n , for a one-element model in a two-choice, noncontingent case.

We now present a general set of rules governing changes in conditioning states. As the model is developed it will become obvious that for some experimental problems restrictions that greatly simplify the process can be imposed.

If the subject is in state C_1 and an E_1 occurs (i.e., the subject makes an A_1 -response, which is correct), then he will remain in C_1 . However, if the subject is in C_1 and an E_2 occurs, then with probability c the subject goes to C_2 and with probability c' to C_0 . Comparable rules apply when the subject is in C_2 . Thus, if the subject is in C_1 or C_2 and his response is correct, he will remain in C_1 or C_2 . If, however, he is in C_1 or C_2 and his response is not correct, then he may shift to one of the other conditioning states, which reduces the probability of repeating the same response on the next trial.

Finally, if the subject is in C_0 and an E_1 or E_2 occurs, then with probability c'' the subject moves to C_1 or C_2 , respectively.⁵ Thus, to summarize, for $i, j = 1, 2$ and $i \neq j$,

$$\begin{aligned} \Pr(C_{i,n+1} | E_{i,n}C_{i,n}) &= 1 \\ \Pr(C_{0,n+1} | E_{j,n}C_{i,n}) &= c' \\ \Pr(C_{j,n+1} | E_{j,n}C_{i,n}) &= c \\ \Pr(C_{i,n+1} | E_{i,n}C_{0,n}) &= c'' \end{aligned} \quad (13)$$

where $0 < c'' \leq 1$ and $0 < c + c' \leq 1$.

We now use the assumptions of the preceding paragraphs and the particular assumptions for the noncontingent case to derive the transition matrix in the conditioning states. In making such a derivation it is convenient to represent the various *possible* occurrences on a trial by a tree. Each set of branches emanating from a point represents a mutually exclusive and exhaustive set of possibilities. For example, suppose that at the start of trial n the subject is in state C_1 ; the tree in Fig. 2 represents the possible changes that can occur in the conditioning state.

⁵ Here we assume that the subject's response does not affect the change; that is, if the subject is in C_0 and an E_1 occurs, then he will move to C_1 with probability c'' , no matter whether A_1 or A_2 has occurred. This assumption is not necessary and we could readily have the actual response affect change. For example, we might postulate c_1'' for an A_1E_1 or A_2E_2 combination, and c_2'' for the A_1E_2 or A_2E_1 combination; that is,

$$\Pr(C_{1,n+1} | E_{1,n}A_{1,n}C_{0,n}) = \Pr(C_{2,n+1} | E_{2,n}A_{2,n}C_{0,n}) = c_1''$$

and

$$\Pr(C_{1,n+1} | E_{1,n}A_{2,n}C_{0,n}) = \Pr(C_{2,n+1} | E_{2,n}A_{1,n}C_{0,n}) = c_2''$$

where

$$c_1'' \neq c_2''.$$

However, such additions make the mathematical process more complicated and should be introduced only when the data clearly require them.

The first set of branches is associated with the reinforcing event on trial n . If the subject is in C_1 and an E_1 occurs, then he will stay in state C_1 on the next trial. However, if an E_2 occurs, then with probability c he will go to C_2 , with probability c' he will go to C_0 , and with probability $1 - c - c'$ he will remain in C_1 .

Each path of a tree, from a beginning point to a terminal point, represents a possible outcome on a given trial. The probability of each path is obtained by multiplying the appropriate conditional probabilities. Thus for the tree in Fig. 2 the probability of the bottom path may be represented by $\Pr(E_{2,n} | C_{1,n}) \Pr(C_{1,n+1} | E_{2,n} C_{1,n}) = (1 - \pi)(1 - c - c')$. Two of the four paths lead from C_1 to C_1 ; hence

$$p_{11} = \Pr(C_{1,n+1} | C_{1,n}) = \pi + (1 - \pi)(1 - c - c').$$

Similarly, $p_{10} = (1 - \pi)c'$ and $p_{12} = (1 - \pi)c$, where p_{ij} denotes the probability of a one-step transition from C_i to C_j .

For the C_0 state we have the tree given in Fig. 3. On the top branch an E_1 event is indicated; by Eq. 13 the probability of going to C_1 is c'' and of staying in C_0 is $1 - c''$. A similar analysis holds for the bottom branches. Thus we have

$$p_{01} = \pi c''$$

$$p_{02} = (1 - \pi)c''$$

$$p_{00} = 1 - c''.$$

A combination of these results and the comparable results for C_2 yields the following transition matrix:

$$P = \begin{matrix} & \begin{matrix} C_1 & C_0 & C_2 \end{matrix} \\ \begin{matrix} C_1 \\ C_0 \\ C_2 \end{matrix} & \begin{bmatrix} 1 - (1 - \pi)(c' + c) & c'(1 - \pi) & c(1 - \pi) \\ c''\pi & 1 - c'' & c''(1 - \pi) \\ c\pi & c'\pi & 1 - \pi(c' + c) \end{bmatrix} \end{matrix} \quad (14)$$

As in the case of the paired-associate model, a large number of predictions can be derived easily for this process. However, we shall select only a few that will help to clarify the fundamental properties of the model. We begin by considering the asymptotic probability of a particular conditioning state and, in turn, the asymptotic probability of an A_1 -response. The following notation will prove useful: let $[p_{ij}]$ be the transition matrix and define $p_{ij}^{(n)}$ as the probability of being in state j on trial $r + n$, given that at trial r the subject was in state i . The quantity is defined recursively:

$$p_{ij}^{(1)} = p_{ij}, \quad p_{ij}^{(n+1)} = \sum_v p_{iv} p_{vj}^{(n)}.$$

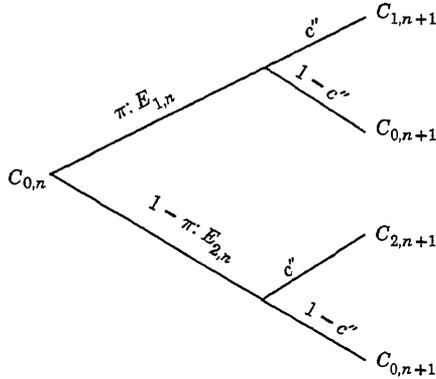


Fig. 3. Branching process, starting from state C_0 on trial n , for a one-element model in a two-choice, noncontingent case.

Moreover, if the appropriate limit exists and is independent of i , we set

$$u_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)}.$$

The limiting quantities u_j exist for any finite-state Markov chain that is irreducible and aperiodic. A Markov chain is irreducible if there is no closed proper subset of states; that is, no proper subset of states such that once within this set the probability of leaving it is 0. For example, the chain whose transition matrix is

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{matrix}$$

is reducible because the set $\{1, 2\}$ of states is a proper closed subset. A Markov chain is aperiodic if there is no fixed period for return to any state and periodic if a return to some initial state j is impossible except at $t, 2t, 3t, \dots$ trials for $t > 1$. Thus the chain whose matrix is

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{matrix}$$

has period $t = 3$ for return to each state.

If there are r states, we call the vector $\mathbf{u} = (u_1, u_2, \dots, u_r)$ the *stationary probability* vector of the chain. It may be shown (Feller, 1957; Kemeny & Snell, 1959) that the components of this vector are the solutions of the r linear equations

$$\begin{aligned} u_1 &= \sum_{v=1}^r u_v p_{v1} \\ u_2 &= \sum_{v=1}^r u_v p_{v2} \end{aligned} \quad (15)$$

$$u_r = \sum_{v=1}^r u_v p_{vr}$$

such that $\sum_{v=1}^r u_v = 1$. Thus, to find the asymptotic probabilities u_j of the states, we need find only the solution of the r equations. The intuitive basis of this system of equations seems clear. Consider a two-state chain. Then the probability p_{n+1} of being in state 1 on trial $n + 1$ is the probability of being in state 1 on trial n and going to 1 plus the probability of being in state 2 on trial n and going to 1; that is

$$p_{n+1} = p_{11}p_n + p_{21}(1 - p_n).$$

But at asymptote $p_{n+1} = p_n = u_1$ and $1 - p_n = u_2$, whence

$$u_1 = p_{11}u_1 + p_{21}u_2,$$

which is the first of the two equations of the system when $r = 2$.

It is clear that the chain represented by the matrix P of Eq. 14 is irreducible and aperiodic; thus the asymptotes exist and are independent of the initial probability distribution on the states. Let $[p_{ij}]$ ($i, j = 1, 2, 3$) be any 3×3 transition matrix. Then we seek the numbers u_j such that $u_j = \sum_v u_v p_{vj}$ and $\sum u_j = 1$. The general solution is given by $u_j = D_j/D$, where

$$\begin{aligned} D_1 &= p_{31}(1 - p_{22}) + p_{21}p_{32} \\ D_2 &= p_{31}p_{12} + p_{32}(1 - p_{11}) \\ D_3 &= (1 - p_{11})(1 - p_{22}) - p_{21}p_{12} \\ D &= D_1 + D_2 + D_3. \end{aligned} \quad (16)$$

Inserting in these equations the equivalents of the p_{ij} from the transition matrix and renumbering the states appropriately, we obtain

$$\begin{aligned} D_1 &= \pi c''(c + c'\pi) \\ D_0 &= \pi(1 - \pi)c'(c' + 2c) \\ D_2 &= (1 - \pi)c''[c + c'(1 - \pi)]. \end{aligned}$$

Since D is the sum of the D_j 's and since $u_j = D_j/D$, we may divide the numerator and denominator by $(c'')^2$ and obtain

$$\begin{aligned}
 u_1 &= \frac{\pi(\rho + \epsilon\pi)}{\pi(\rho + \epsilon\pi) + \pi(1 - \pi)\epsilon(\epsilon + 2\rho) + (1 - \pi)[\rho + \epsilon(1 - \pi)]} \\
 u_0 &= \frac{\pi(1 - \pi)\epsilon(\epsilon + 2\rho)}{\pi(\rho + \epsilon\pi) + \pi(1 - \pi)\epsilon(\epsilon + 2\rho) + (1 - \pi)[\rho + \epsilon(1 - \pi)]} \quad (17) \\
 u_2 &= 1 - u_1 - u_0,
 \end{aligned}$$

where $\rho = c/c''$ and $\epsilon = c'/c''$.

By our response axioms we have

$$\Pr(A_{1,n}) = \Pr(C_{1,n}) + \frac{1}{2} \Pr(C_{0,n})$$

for all n . Hence

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \Pr(A_{1,n}) &= u_1 + \frac{1}{2}u_0 \\
 &= \frac{\pi(\rho + \epsilon\rho + \frac{1}{2}\epsilon^2) + \pi^2(\epsilon - \epsilon\rho - \frac{1}{2}\epsilon^2)}{\pi(\epsilon^2 + 2\epsilon\rho - 2\epsilon) + \pi^2(2\epsilon - \epsilon^2 - 2\epsilon\rho) + \rho + \epsilon}. \quad (18)
 \end{aligned}$$

An inspection of Eq. 18 indicates that the asymptotic probability of an A_1 -response is a function of π , ρ , and ϵ . As will become clear later, the value of $\Pr(A_{1,\infty})$ is bounded in the open interval from $\frac{1}{2}$ to $\pi^2/[\pi^2 + (1 - \pi)^2]$; whether $\Pr(A_{1,\infty})$ is above or below π depends on the values of ρ and ϵ .

We now consider two special cases of our one-element model. The first case is comparable to the multi-element models to be discussed later, whereas the second case is, in some respects, the complement of the first case.

Case of $c' = 0$. Let us rewrite Eq. 14 with $c' = 0$. Then the transition matrix has the following canonical form:

$$P = \begin{matrix} & \begin{matrix} C_1 & C_2 & C_0 \end{matrix} \\ \begin{matrix} C_1 \\ C_2 \\ C_0 \end{matrix} & \begin{bmatrix} 1 - c(1 - \pi) & c(1 - \pi) & 0 \\ c\pi & 1 - c\pi & 0 \\ c''\pi & c''(1 - \pi) & 1 - c'' \end{bmatrix} \end{matrix} \quad (19)$$

We note that once the subject has left state C_0 he can never return. In fact, it is obvious that $\Pr(C_{0,n}) = \Pr(C_{0,1})(1 - c'')^{n-1}$ where $\Pr(C_{0,1})$ is the initial probability of being in C_0 . Thus, except on early trials, C_0 is not part of the process, and the subject in the long run fluctuates between C_1 and C_2 , being in C_1 on a proportion π of the trials.

From Eq. 19 we have also

$$\Pr(C_{1,n+1}) = \Pr(C_{1,n})[1 - c(1 - \pi)] + \Pr(C_{2,n})c\pi + \Pr(C_{0,n})c''\pi;$$

that is, the probability of being in C_1 on trial $n + 1$ is equal to the probability of being in C_1 on trial n times the probability p_{11} of going from C_1 to C_1 plus the probability of being in C_2 times p_{21} plus the probability of being in C_0 times p_{01} . For simplicity let $x_n = \Pr(C_{1,n})$, $y_n = \Pr(C_{2,n})$, and $z_n = \Pr(C_{0,n})$. Now we know that $z_n = z_1(1 - c'')^{n-1}$ and also that $x_n + y_n + z_n = 1$, or $y_n = 1 - x_n - z_1(1 - c'')^{n-1}$. Making these substitutions in the foregoing recursion yields

$$\begin{aligned} x_{n+1} &= x_n[1 - c(1 - \pi)] + z_1 c'' \pi (1 - c'')^{n-1} + c\pi[1 - x_n - z_1(1 - c'')^{n-1}] \\ &= x_n(1 - c) + z_1(1 - c'')^{n-1} \pi (c'' - c) + c\pi. \end{aligned}$$

This difference equation has the following solution⁶:

$$x_n = \pi - (\pi - x_1)(1 - c)^{n-1} - \pi z_1[(1 - c'')^{n-1} - (1 - c)^{n-1}].$$

But $\Pr(A_{1,n}) = x_n + \frac{1}{2}z_n$; hence

$$\begin{aligned} \Pr(A_{1,n}) &= \pi - [\pi - \pi \Pr(C_{0,1}) - \Pr(C_{1,1})](1 - c)^{n-1} \\ &\quad - \Pr(C_{0,1})(\pi - \frac{1}{2})(1 - c'')^{n-1}. \quad (20) \end{aligned}$$

If $\Pr(C_{0,1}) = 0$, then we have a simple exponential learning function starting at $\Pr(C_{1,1})$ and approaching π at a rate determined by c . If $\Pr(C_{0,1}) \neq 0$, then the rate of approach is a function of both c and c'' .

We now consider one simple sequential prediction to illustrate another feature of the one-element model for $c' = 0$. Specifically, consider the probability of an A_1 -response on trial $n + 1$ given a reinforced A_1 -response on trial n ; namely $\Pr(A_{1,n+1} | E_{1,n}A_{1,n})$. Note first of all that

$$\Pr(A_{1,n+1} | E_{1,n}A_{1,n}) \Pr(E_{1,n}A_{1,n}) = \Pr(A_{1,n+1}E_{1,n}A_{1,n}).$$

⁶ The solution of such a difference equation can readily be obtained. Consider $x_{n+1} = ax_n + bc^{n-1} + d$ where a , b , c , and d are constants. Then

$$(1) \quad x_2 = ax_1 + b + d.$$

Similarly, $x_3 = ax_2 + bc + d$ and substituting (1) for x_2 we obtain

$$(2) \quad x_3 = a^2x_1 + ab + ad + bc + d.$$

Similarly, $x_4 = ax_3 + bc^2 + d$ and substituting (2) for x_3 we obtain

$$(3) \quad x_4 = a^3x_1 + a^2b + a^2d + abc + ad + bc^2 + d.$$

If we continue in this fashion, it will be obvious that for $n \geq 2$

$$x_n = a^{n-1}x_1 + d \sum_{i=0}^{n-2} a^i + a^{n-2}b \sum_{i=0}^{n-2} \left(\frac{c}{a}\right)^i.$$

Carrying out the summations yields the desired results. See Jordan (1950, pp. 583-584) for a detailed treatment.

Further, we may write

$$\begin{aligned} & \Pr(A_{1,n+1}E_{1,n}A_{1,n}) \\ &= \sum_{i,j} \Pr(A_{1,n+1}C_{i,n+1}E_{1,n}A_{1,n}C_{j,n}) \\ &= \sum_{i,j} \Pr(A_{1,n+1} \mid C_{i,n+1}E_{1,n}A_{1,n}C_{j,n}) \Pr(C_{i,n+1} \mid E_{1,n}A_{1,n}C_{j,n}) \\ & \quad \cdot \Pr(E_{1,n} \mid A_{1,n}C_{j,n}) \Pr(A_{1,n} \mid C_{j,n}) \Pr(C_{j,n}). \end{aligned}$$

By assumption the probability of a response is determined solely by the conditioning state, hence

$$\Pr(A_{1,n+1} \mid C_{i,n+1}E_{1,n}A_{1,n}C_{j,n}) = \Pr(A_{1,n+1} \mid C_{i,n+1}).$$

Further, by assumption, the probability of an E_1 -event is independent of other events, and $\Pr(E_{1,n} \mid A_{1,n}C_{j,n}) = \pi$. Substituting these results in the foregoing expression, we obtain

$$\begin{aligned} \Pr(A_{1,n+1}E_{1,n}A_{1,n}) &= \pi \sum_{i,j} \Pr(A_{1,n+1} \mid C_{i,n+1}) \Pr(C_{i,n+1} \mid E_{1,n}A_{1,n}C_{j,n}) \\ & \quad \cdot \Pr(A_{1,n} \mid C_{j,n}) \Pr(C_{j,n}). \end{aligned}$$

Both i and j run over 0, 1, and 2, and therefore there are nine terms in the sum; but note that when $i = 2$, the term $\Pr(A_{1,n+1} \mid C_{i,n+1})$ is zero and when $j = 2$ the term $\Pr(A_{1,n} \mid C_{j,n})$ is zero. Consequently it suffices to limit i and j to 0 and 1, and we have

$$\begin{aligned} & \Pr(A_{1,n+1}E_{1,n}A_{1,n}) \\ &= \pi \sum_{i=0}^1 \Pr(A_{1,n+1} \mid C_{i,n+1}) \Pr(C_{i,n+1} \mid E_{1,n}A_{1,n}C_{1,n}) \Pr(A_{1,n} \mid C_{1,n}) \Pr(C_{1,n}) \\ &+ \pi \sum_{i=0}^1 \Pr(A_{1,n+1} \mid C_{i,n+1}) \Pr(C_{i,n+1} \mid E_{1,n}A_{1,n}C_{0,n}) \Pr(A_{1,n} \mid C_{0,n}) \Pr(C_{0,n}). \end{aligned}$$

Since the subject cannot leave state C_1 on a trial when A_1 is reinforced, we know that

$$\Pr(C_{1,n+1} \mid E_{1,n}A_{1,n}C_{1,n}) = 1 \quad \text{and} \quad \Pr(C_{0,n+1} \mid E_{1,n}A_{1,n}C_{1,n}) = 0;$$

further, $\Pr(A_{1,n+1} \mid C_{1,n+1}) = 1$. Therefore the first sum is simply $\pi \Pr(C_{1,n})$. For the second sum, $\Pr(C_{1,n+1} \mid E_{1,n}A_{1,n}C_{0,n}) = c''$ and $\Pr(C_{0,n+1} \mid E_{1,n}A_{1,n}C_{0,n}) = 1 - c''$. Further, $\Pr(A_{1,n} \mid C_{0,n}) = \frac{1}{2}$; hence for the second sum we obtain

$$\pi[c''\frac{1}{2} + \frac{1}{2}(1 - c'')\frac{1}{2}] \Pr(C_{0,n}).$$

Combining these results,

$$\Pr(A_{1,n+1}E_{1,n}A_{1,n}) = \pi\{\Pr(C_{1,n}) + \frac{1}{2}\Pr(C_{0,n})[c'' + (1 - c'')\frac{1}{2}]\}.$$

But

$$\Pr(E_{1,n}A_{1,n}) = \Pr(E_{1,n} | A_{1,n}) \Pr(A_{1,n}) = \pi \Pr(A_{1,n}),$$

whence

$$\Pr(A_{1,n+1} | E_{1,n}A_{1,n}) = \frac{\Pr(C_{1,n}) + \frac{1}{2} \Pr(C_{0,n})[c'' + (1 - c'')^{\frac{1}{2}}]}{\Pr(A_{1,n})}.$$

We know that $\Pr(C_{1,n})$ and $\Pr(A_{1,n})$ both approach π in the limit and that $\Pr(C_{0,n})$ approaches 0. Therefore we predict that

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n+1} | E_{1,n}A_{1,n}) = 1.$$

This prediction provides a sharp test for this particular case of the model and one that is certain to fail in almost any experimental situation; that is, even after a large number of trials it is hard to conceive of an experimental procedure such that a response will be repeated with probability 1 if it occurred and was reinforced on the preceding trial. Later we shall consider a multi-element model that provides an excellent description of many sets of data but is based on essentially the same conditioning rules specified by this case of $c' = 0$. It should be emphasized that deterministic predictions of the sort given in the foregoing equation are peculiar to one-element models; for the multi-element case such difficulties do not arise. This point is amplified later.

Case of $c = 0$. We now consider the case in which direct counter-conditioning does not occur, that is, $c = 0$, and thus $\rho = 0$ and $0 < \epsilon < \infty$. With this restriction the chain is still ergodic, since it is possible to go from every state to every other state, but transitions between C_1 and C_2 must go by way of C_0 . Letting $\rho = 0$ in Eq. 18, we obtain

$$\Pr(A_{1,\infty}) = \frac{\pi^2 + \frac{1}{2}\pi(1 - \pi)\epsilon}{\pi^2 + \pi(1 - \pi)\epsilon + (1 - \pi)^2}. \quad (21)$$

From Eq. 21 we can draw some interesting conclusions about the relationship of the asymptotic response probabilities to the ratio $\epsilon = c'/c''$. Differentiating with respect to ϵ , we obtain

$$\frac{\partial}{\partial \epsilon} \Pr(A_{1,\infty}) = \frac{\pi(1 - \pi)(\frac{1}{2} - \pi)}{[\pi^2 + (1 - \pi)^2 + \pi(1 - \pi)\epsilon]^2}.$$

If $\pi(1 - \pi)(\frac{1}{2} - \pi) \neq 0$, then $\Pr(A_{1,\infty})$ has no maximum for ϵ in the open interval $(0, \infty)$, which is the permissible range on ϵ . In fact, since the sign of the derivative is independent of ϵ , we know that $\Pr(A_{1,\infty})$ is either monotone increasing or monotone decreasing in ϵ : strictly increasing if $\pi(1 - \pi)(\frac{1}{2} - \pi) > 0$ (i.e., $\pi < \frac{1}{2}$) and decreasing if $\pi(1 - \pi)(\frac{1}{2} - \pi) < 0$ (i.e., $\pi > \frac{1}{2}$). Moreover, because of the monotonicity of $\Pr(A_{1,\infty})$ in ϵ ,

it is easy to compute bounds from Eq. 21. First, we see immediately that the lower bound (assuming $\pi > \frac{1}{2}$) is $\lim_{\epsilon \rightarrow \infty} \Pr(A_{1,\infty}) = \frac{1}{2}$. Second, when ϵ is very small, $\Pr(A_{1,\infty})$ approaches $\frac{\pi^2}{\pi^2 + (1 - \pi)^2}$. Note, however, that Eq. 21 is inapplicable when $\epsilon = 0$; for if both $c = 0$ and $c' = 0$ the transition matrix (Eq. 14) reduces to

$$P = \begin{bmatrix} 1 & 0 & 0 \\ c''\pi & 1 - c'' & c''(1 - \pi) \\ 0 & 0 & 1 \end{bmatrix},$$

and, if the process starts in C_0 , $\Pr(A_{1,\infty}) = \pi$. But for $\epsilon > 0$, if $\pi > \frac{1}{2}$, $\Pr(A_{1,\infty})$ is a decreasing function of ϵ and its values lie in the half-open interval

$$\frac{1}{2} \leq \Pr(A_{1,\infty}) < \frac{\pi^2}{\pi^2 + (1 - \pi)^2}.$$

It is readily determined that probability matching would not generally be predicted in this case. When c'/c'' is greater than 2, the predicted value of $\Pr(A_{1,\infty})$ is less than π , and when this ratio is less than 2 the predicted value of $\Pr(A_{1,\infty})$ is greater than π .

Finally, we derive $\Pr(A_{1,n+1} | E_{1,n}A_{1,n})$ for this case. The derivation is identical to that given for $c' = 0$. Hence

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n+1} | E_{1,n}A_{1,n}) = \frac{u_1 + \frac{1}{2}u_0[c'' + (1 - c'')\frac{1}{2}]}{u_1 + \frac{1}{2}u_0}.$$

Note, however, that for $c = 0$ the quantity u_0 is never 0 (except for $\pi = 0, 1$), and consequently $\Pr(A_{1,n+1} | E_{1,n}A_{1,n})$ is always less than 1.

Contingent Reinforcement. As a final example we shall apply the one-element model to a situation in which the reinforcing event on trial n is contingent on the response on that trial. Simple contingent reinforcement is defined by two probabilities π_{11} and π_{21} such that

$$\Pr(E_{1,n} | A_{1,n}) = \pi_{11} \quad \text{and} \quad \Pr(E_{1,n} | A_{2,n}) = \pi_{21}.$$

We consider the case of the model in which $c' = 0$ and $\Pr(C_{0,1}) = 0$; that is, the subject is not in state C_0 on trial 1 and (since $c' = 0$) he can never reach C_0 from C_1 or C_2 . Hence on all trials he is in C_1 or C_2 , and transitions between these states are governed by the single parameter c . The trees for the C_1 and C_2 states are given in Fig. 4.

The transition matrix is

$$P = \begin{matrix} & \begin{matrix} C_1 & C_2 \end{matrix} \\ \begin{matrix} C_1 \\ C_2 \end{matrix} & \begin{bmatrix} 1 - (1 - \pi_{11})c & (1 - \pi_{11})c \\ c\pi_{21} & 1 - c\pi_{21} \end{bmatrix}, \end{matrix}$$

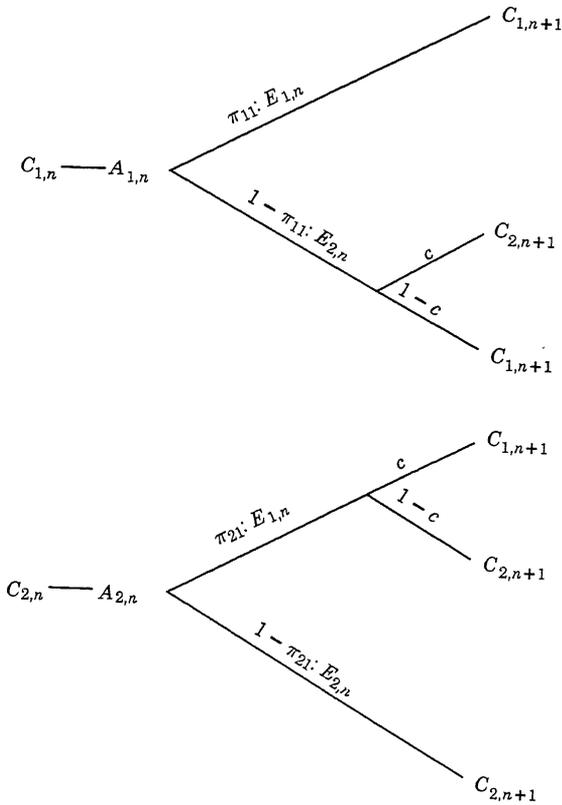


Fig. 4. Branching process for one-element model in two-choice, contingent case.

and, in terms of this matrix, we may write

$$\Pr(C_{1,n+1}) = \Pr(C_{1,n})[1 - (1 - \pi_{11})c] + \Pr(C_{2,n})c\pi_{21}.$$

But $\Pr(C_{2,n}) = 1 - \Pr(C_{1,n})$ and $\Pr(C_{1,n}) = \Pr(A_{1,n})$; hence

$$\Pr(A_{1,n+1}) = \Pr(A_{1,n})[1 - (1 - \pi_{11})c - c\pi_{21}] + c\pi_{21}.$$

This difference equation has the solution

$$\Pr(A_{1,n}) = \Pr(A_{1,\infty}) - [\Pr(A_{1,\infty}) - \Pr(A_{1,1})][1 - c(1 - \pi_{11} + \pi_{21})]^{n-1},$$

where

$$\Pr(A_{1,\infty}) = \frac{\pi_{21}}{1 - \pi_{11} + \pi_{21}}.$$

The asymptote is independent of c , and the rate of approach is determined by the quantity $c(1 - \pi_{11} + \pi_{21})$. It is interesting to note that the learning function for $\text{Pr}(A_{1,n})$ in this case of the one-element model is identical to that of the linear model (cf. Estes & Suppes, 1959a).

2. MULTI-ELEMENT PATTERN MODELS

2.1 General Formulation

In the literature of stimulus sampling theory a variety of proposals has been made for conceptually representing the stimulus situation. Fundamental to all of these suggestions has been the distinction between pattern elements and component elements. For the one-element case this distinction does not play a serious role, but for multi-element formulations these alternative representations of the stimulus situation specify different mathematical processes.

In component models the stimulating situation is represented as a population of elements which the learner is viewed as sampling from trial to trial. It is assumed that the conditioning of individual elements to responses occurs independently as the elements are sampled in conjunction with reinforcing events and that the response probability in the presence of a sample containing a number of elements is determined by an averaging rule. The principal consideration has been to account for response variability to an apparently constant stimulus situation by postulating random fluctuations from trial to trial in the particular sample of stimulus elements affecting the learner. These component models have provided a mechanism for effecting a reconciliation between the picture of gradual change usually exhibited by the learning curve and the all-or-none law of association.

For many experimental situations a detailed account of the quantitative properties of learning can be given by component models that assume discrete associations between responses and the independently variable elements of a stimulating situation. However, in some cases predictions from component models fail, and it appears that a simple account of the learning process requires the assumption that responses become associated, not with separate components or aspects of a stimulus situation, but with total patterns of stimulation considered as units. The model presented in this section is intended to represent such a case. In it we assume that an experimentally specified stimulating situation can be conceived as an assemblage of distinct, mutually exclusive patterns of stimulation, each of which becomes conditioned to responses on an all-or-none basis. By

“mutually exclusive” we mean that exactly one of the patterns occurs (is sampled by the subject) on each trial. By “distinct” we mean that no generalization occurs from one pattern to another. Thus the clearest experimental interpretation would involve a set of patterns having no common elements (i.e., common properties or components). In practice the pattern model has also been applied with considerable success to experiments in which the alternative stimuli have some common elements but nevertheless are sufficiently discriminable so that generalization effects (e.g., “confusion errors”) are small and can be neglected without serious error.

In this presentation we shall limit consideration to cases in which patterns are sampled randomly with equal likelihood so that if there are N patterns each has probability $1/N$ of being sampled on a trial. This sampling assumption represents only one way of formulating the model and is presented here because it generates a fairly simple mathematical process and provides a good account of a variety of experimental results. However, this particular scheme for sampling patterns has restricted applicability. For example, in certain experiments it can be demonstrated that the stimulus array to which the subject responds is in large part determined by events on previous trials; that is, trace stimulation associated with previous responses and rewards determines the stimulus pattern to which the subject responds. When this is the case, it is necessary to postulate a more general rule for sampling patterns than the random scheme proposed (e.g., see the discussion of “hypothesis models” in Suppes & Atkinson, 1960).

Before stating the axioms for the pattern model to be considered in this section, we define the following notions. As before, the behaviors available to the subject are categorized into mutually exclusive and exhaustive response classes (A_1, A_2, \dots, A_r). The possible experimenter-defined outcomes of a trial (e.g., giving or withholding reward, unconditioned stimulus, knowledge of results) are classified by their effect on response probability and are represented by a mutually exclusive and exhaustive set of reinforcing events (E_0, E_1, \dots, E_r). The event E_i ($i \neq 0$) indicates that response A_i is reinforced and E_0 represents any trial outcome whose effect is neutral (i.e., reinforces none of the A_i 's). The subject's response and the experimenter-defined outcomes are observable, but the occurrence of E_i is a purely hypothetical event that represents the reinforcing effect of the trial outcome. Event E_i is said to have occurred when the outcome of a trial increases the probability of response A_i in the presence of the given stimulus—provided, of course, that this probability is not already at its maximum value.

We now present the axioms. The first group of axioms deals with the

conditioning of sampled patterns, the second group with the sampling of patterns, and the third group with responses.

Conditioning Axioms

- C1. *On every trial each pattern is conditioned to exactly one response.*
- C2. *If a pattern is sampled on a trial, it becomes conditioned with probability c to the response (if any) that is reinforced on the trial; if it is already conditioned to that response, it remains so.*
- C3. *If no reinforcement occurs on a trial (i.e., E_0 occurs), there is no change in conditioning on that trial.*
- C4. *Patterns that are not sampled on a trial do not change their conditioning on that trial.*
- C5. *The probability c that a sampled pattern will be conditioned to a reinforced response is independent of the trial number and the preceding events.*

Sampling Axioms

- S1. *Exactly one pattern is sampled on each trial.*
- S2. *Given the set of N patterns available for sampling on a trial, the probability of sampling a given pattern is $1/N$, independent of the trial number and the preceding events.*

Response Axiom

- R1. *On any trial that response is made to which the sampled pattern is conditioned.*

Later in this section we apply these axioms to a two-choice learning experiment and to a paired-comparison study. First, however, we shall prove several general theorems. Before we can begin our analysis it is necessary to define the notion of a conditioning state. For the axioms given, all patterns are sampled with equal probability, and it suffices to let the state of conditioning indicate the number of patterns conditioned to each response. Hence for r responses the conditioning states are the ordered r -tuples $\langle k_1, k_2, \dots, k_r \rangle$ where $k_i = 0, 1, 2, \dots, N$ and $k_1 + k_2 + \dots + k_r = N$; the integer k_i denotes the number of patterns conditioned to the A_i response. The number of possible conditioning states is $\binom{N+r-1}{N}$. (In a generalized model, which permitted different patterns to have different likelihoods of being sampled, it would be necessary to specify not only the number of patterns conditioned to a response but also the sampling probabilities associated with the patterns.) For simplicity we limit consideration in this section to the case of two alternatives, except for one example in which $r = 3$. Given only two alternatives, we denote the conditioning state on trial n of an experiment

as $C_{i,n}$, where $i = 0, 1, 2, \dots, N$; the subscript i indicates the number of patterns conditioned to A_1 and $N - i$ the number conditioned to A_2 . TRANSITION PROBABILITIES. Only one pattern is sampled per trial; therefore the subject can go from state C_i only to one of the three states C_{i-1} , C_i , or C_{i+1} on any given trial. The probabilities of these transitions depend on the value of the conditioning parameter c , the reinforcement schedule, and the value of i . We now proceed to compute these probabilities.

If the subject is in state C_i on trial n and an E_1 occurs, then the possible outcomes are indicated by the tree in Fig. 5. On the upper main branch, which has probability i/N , a pattern that is conditioned to A_1 is sampled and, since an E_1 -reinforcement occurs, the pattern remains conditioned to A_1 . Hence the conditioning state on trial $n + 1$ is the same as on trial n (see Axiom C2). On the lower main branch, which has probability $(N - i)/N$, a pattern conditioned to A_2 is sampled; then with probability c the pattern is conditioned to A_1 and the subject moves to conditioning state C_{i+1} , whereas with probability $1 - c$ conditioning is not effective and the subject remains in state C_i . Putting these results together, we obtain

$$\Pr(C_{i+1,n+1} | E_{1,n}C_{i,n}) = c \frac{N - i}{N}$$

$$\Pr(C_{i,n+1} | E_{1,n}C_{i,n}) = 1 - c + c \frac{i}{N}. \tag{22a}$$

Similarly, if an E_2 occurs on trial n ,

$$\Pr(C_{i-1,n+1} | E_{2,n}C_{i,n}) = c \frac{i}{N}$$

$$\Pr(C_{i,n+1} | E_{2,n}C_{i,n}) = 1 - c + c \frac{N - i}{N}. \tag{22b}$$

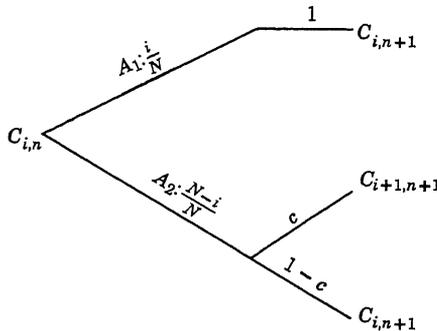


Fig. 5. Branching process for N -element model on a trial when the subject starts in state C_i and an E_1 -event occurs.

By Axiom C3, if an E_0 occurs, then

$$\Pr(C_{i,n+1} \mid E_{0,n}C_{i,n}) = 1. \tag{22c}$$

Noting that a transition upward can occur only when a pattern conditioned to A_2 is sampled on an E_1 -trial and a transition downward can occur only when a pattern conditioned to A_1 is sampled on an E_2 -trial, we can combine the results from Eq. 22a-c to obtain

$$\Pr(C_{i+1,n+1} \mid C_{i,n}) = c \frac{N-i}{N} \Pr(E_{1,n} \mid A_{2,n}C_{i,n}) \tag{23a}$$

$$\Pr(C_{i-1,n+1} \mid C_{i,n}) = c \frac{i}{N} \Pr(E_{2,n} \mid A_{1,n}C_{i,n}) \tag{23b}$$

$$\begin{aligned} \Pr(C_{i,n+1} \mid C_{i,n}) = & 1 - c + c \left[\frac{i}{N} \Pr(E_{1,n} \mid A_{1,n}C_{i,n}) \right. \\ & + \frac{N-i}{N} \Pr(E_{2,n} \mid A_{2,n}C_{i,n}) \\ & \left. + \Pr(E_{0,n} \mid C_{i,n}) \right] \tag{23c} \end{aligned}$$

for the probabilities of one-step transitions between states. Equation 23a, for example, states that the probability of moving from the state with i elements conditioned to A_1 to the state with $i + 1$ elements conditioned to A_1 is the product of the probability $(N - i)/N$ that an element not already conditioned to A_1 is sampled and the probability $c\Pr(E_{1,n} \mid A_{2,n}C_{i,n})$ that, under the given circumstances, conditioning occurs.

As defined earlier, we have a Markov process in the conditioning states if the probability of a transition from any state to any other state depends at most on the state existing on the trial preceding the transition. By inspection of Eq. 23 we see that the Markov condition may be satisfied by limiting ourselves to reinforcement schedules in which the probability of a reinforcing event E_i depends at most on the response of the given trial; that is, in learning-theory terminology, to noncontingent and simple contingent schedules. This restriction will be assumed throughout the present section except for a few remarks in which we explicitly consider various lines of generalization.

With these restrictions in mind, we define

$$\pi_{ij} = \Pr(E_{j,n} \mid A_{i,n}),$$

where $j = 0$ to r , $i = 1$ to r , and $\sum_j \pi_{ij} = 1$; that is, the reinforcement on a trial depends at most on the response of the given trial. Further, the

reinforcement probabilities do not depend on the trial number. We may then rewrite Eq. 23 as follows:

$$q_{i,i+1} = c \frac{N-i}{N} \pi_{21} \quad (24a)$$

$$q_{i,i} = 1 - c \frac{N-i}{N} \pi_{21} - c \frac{i}{N} \pi_{12} \quad (24b)$$

$$q_{i,i-1} = c \frac{i}{N} \pi_{12}. \quad (24c)$$

Note that we use the notation q_{ij} in place of $\Pr(C_{j,n+1} | C_{i,n})$. The reason is that the transition probabilities do not depend on n , given the restrictions on the reinforcement schedule stated above, and the simpler notation expresses this fact.

RESPONSE PROBABILITIES AND MOMENTS. By Axioms S1, S2, and R1 we know that the relation between response probability and the conditioning state is simply

$$\Pr(A_{1,n} | C_{i,n}) = \frac{i}{N}.$$

Hence

$$\begin{aligned} \Pr(A_{1,n}) &= \sum_{i=0}^N \Pr(A_{1,n} | C_{i,n}) \Pr(C_{i,n}) \\ &= \sum_{i=0}^N \frac{i}{N} \Pr(C_{i,n}). \end{aligned} \quad (25)$$

But note that by definition of the transition probabilities q_{ij}

$$\begin{aligned} \Pr(C_{i,n}) &= \Pr(C_{0,n-1})q_{0i} + \Pr(C_{1,n-1})q_{1i} + \dots + \Pr(C_{N,n-1})q_{Ni} \\ &= \sum_{j=0}^N \Pr(C_{j,n-1})q_{ji}. \end{aligned} \quad (26)$$

The latter expression, together with Eq. 25, serves as the basis for a general recursion in $\Pr(A_{1,n})$:

$$\Pr(A_{1,n}) = \sum_{i=0}^N \frac{i}{N} \sum_{j=0}^N \Pr(C_{j,n-1})q_{ji}.$$

Now substituting for q_{ji} in terms of Eq. 24 and rearranging the sum we have

$$\begin{aligned} \Pr(A_{1,n}) &= \sum_{i=0}^N \frac{i}{N} \Pr(C_{i,n-1}) - c\pi_{12} \sum_{i=1}^N \frac{i^2}{N^2} \Pr(C_{i,n-1}) \\ &\quad - c\pi_{21} \sum_{i=0}^{N-1} \frac{i(N-i)}{N^2} \Pr(C_{i,n-1}) \\ &\quad + c\pi_{21} \sum_{i=0}^{N-1} \frac{(i+1)(N-i)}{N^2} \Pr(C_{i,n-1}) \\ &\quad + c\pi_{12} \sum_{i=1}^N \frac{i(i-1)}{N^2} \Pr(C_{i,n-1}). \end{aligned}$$

The first sum is, by Eq. 25, $\Pr(A_{1,n-1})$. Let us define

$$\alpha_{2,n} = \sum_{i=0}^N (i^2/N^2) \Pr(C_{i,n});$$

then the second sum is simply $-c\pi_{12}\alpha_{2,n-1}$. Similarly, the third sum is

$$\begin{aligned} -c\pi_{21}[\Pr(A_{1,n-1}) - \Pr(C_{N,n-1}) - \alpha_{2,n-1} + \Pr(C_{N,n-1})] \\ = -c\pi_{21}[\Pr(A_{1,n-1}) - \alpha_{2,n-1}], \end{aligned}$$

and so forth. Carrying out the summation and simplifying, we obtain the following recursion in $\Pr(A_{1,n})$:

$$\Pr(A_{1,n}) = \left[1 - \frac{c}{N}(\pi_{12} + \pi_{21}) \right] \Pr(A_{1,n-1}) + \frac{c}{N}\pi_{21}. \tag{27}$$

This difference equation has the well-known solution (cf. Bush & Mos-teller, 1955; Estes, 1959b; Estes & Suppes, 1959)

$$\Pr(A_{1,n}) = \Pr(A_{1,\infty}) - [\Pr(A_{1,\infty}) - \Pr(A_{1,1})] \left[1 - \frac{c}{N}(\pi_{12} + \pi_{21}) \right]^{n-1}, \tag{28}$$

where

$$\Pr(A_{1,\infty}) = \frac{\pi_{21}}{\pi_{21} + \pi_{12}}.$$

At this point it will also be instructive to calculate the variance of the distribution of response probabilities $\Pr(A_{1,n} | C_{i,n})$. The second raw moment, as defined above, is

$$\alpha_{2,n} = \sum_{i=0}^N \frac{i^2}{N^2} \Pr(C_{i,n}) = \sum_{i=0}^N \frac{i^2}{N^2} \sum_{j=0}^N \Pr(C_{j,n-1}) q_{ji}. \tag{29}$$

Carrying out the summation, as in the case of Eq. 27, we obtain

$$\begin{aligned} \alpha_{2,n} = \alpha_{2,n-1} \left[1 - \frac{2c}{N}(\pi_{12} + \pi_{21}) \right] \\ + \Pr(A_{1,n-1}) \left[c\pi_{12} \frac{1}{N^2} + c\pi_{21} \left(\frac{2}{N} - \frac{1}{N^2} \right) \right] + \frac{c}{N^2}\pi_{21}. \end{aligned}$$

Subtracting the square of $\Pr(A_{1,n})$, as given in Eq. 28, from $\alpha_{2,n}$ yields the variance of the response probabilities. The second and higher moments of the response probabilities are of experimental interest primarily because they enter into predictions concerning various sequential statistics. We shall return to this point later.

ASYMPTOTIC DISTRIBUTIONS. The pattern model has one particularly advantageous feature not shared by many other learning models that have appeared in the literature. This feature is a simple calculational procedure

for generating the complete asymptotic distribution of conditioning states and therefore the asymptotic distribution of responses. The derivation to be given assumes that all elements $q_{i,i-1}$, $q_{i,i}$, $q_{i,i+1}$ of the transition matrix are nonzero; the same technique can be applied if there are zero entries, except, of course, that in forming ratios one must keep the zeros out of the denominators.

As in Sec. 1.3, we let $\lim_{n \rightarrow \infty} \Pr(C_{i,n}) = u_i$. The theorem to be proved is that all of the asymptotic conditioning state probabilities u_i can be expressed recursively in terms of u_0 ; since the u_i 's must sum to unity, this recursion suffices to determine the entire distribution.

By Eq. 26 we note that

$$u_0 = u_0 q_{00} + u_1 q_{10},$$

hence

$$\frac{u_0}{u_1} = \frac{q_{10}}{1 - q_{00}} = \frac{q_{10}}{q_{01}}.$$

We now prove by induction that a similar relation holds for any adjacent pair of states; that is,

$$\frac{u_i}{u_{i+1}} = \frac{q_{i+1,i}}{q_{i,i+1}}.$$

For any state i we have by Eq. 26

$$u_i = u_{i-1} q_{i-1,i} + u_i q_{i,i} + u_{i+1} q_{i+1,i}.$$

Rearranging,

$$u_i(1 - q_{i,i}) = u_{i-1} q_{i-1,i} + u_{i+1} q_{i+1,i}.$$

However, under the inductive hypothesis we may replace u_{i-1} by its equivalent $u_i q_{i,i-1} / q_{i-1,i}$. Hence

$$\begin{aligned} u_i(1 - q_{i,i}) &= \frac{u_i q_{i,i-1} q_{i-1,i}}{q_{i-1,i}} + u_{i+1} q_{i+1,i} \\ &= u_i q_{i,i-1} + u_{i+1} q_{i+1,i} \end{aligned}$$

or

$$u_i(1 - q_{i,i} - q_{i,i-1}) = u_{i+1} q_{i+1,i}.$$

However, $1 - q_{i,i} - q_{i,i-1} = q_{i,i+1}$, since $q_{i,i-1} + q_{i,i} + q_{i,i+1} = 1$, and therefore

$$\frac{u_i}{u_{i+1}} = \frac{q_{i+1,i}}{q_{i,i+1}},$$

which concludes the proof.

Thus we may write

$$u_1 = \frac{q_{01}}{q_{10}} u_0, \quad u_2 = \frac{q_{12}}{q_{21}} u_1 = \frac{q_{12} q_{01}}{q_{21} q_{10}} u_0,$$

and so forth. Since the u_i 's must sum to unity, u_0 also is determined. To illustrate the application of this technique, we consider some simple cases. For the noncontingent case discussed in Sec. 1.3.

$$\begin{aligned} \pi &= \pi_{21} = \pi_{11} \\ 1 - \pi &= \pi_{12} = \pi_{22}. \end{aligned}$$

By Eq. 24 we have

$$\begin{aligned} q_{i,i+1} &= c \frac{N-i}{N} \pi \\ q_{i,i-1} &= c \frac{i}{N} (1-\pi). \end{aligned}$$

Applying the technique of the previous paragraph,

$$\begin{aligned} \frac{u_1}{u_0} &= \frac{c\pi}{c(1/N)(1-\pi)} = \frac{N\pi}{(1-\pi)} \\ \frac{u_2}{u_1} &= \frac{\pi c[(N-1)/N]}{(1-\pi)c(2/N)} = \frac{(N-1)\pi}{2(1-\pi)} \end{aligned}$$

and in general

$$\frac{u_k}{u_{k-1}} = \frac{(N-k+1)\pi}{k(1-\pi)}.$$

This result has two interesting features. First, we note that the asymptotic probabilities are independent of the conditioning parameter c . Second, the ratio of u_k to u_{k-1} is the same as that of neighboring terms

$$\binom{N}{k} \pi^k (1-\pi)^{N-k} \quad \text{and} \quad \binom{N}{k-1} \pi^{k-1} (1-\pi)^{N-k+1}$$

in the expansion of $[\pi + (1-\pi)]^N$. Therefore the asymptotic probabilities in this case are binomially distributed. For a population of subjects whose learning is described by the model, the limiting proportion of subjects having all N patterns conditioned to A_1 is π^N ; the proportion having all but one of the N patterns conditioned to A_1 is $N\pi^{N-1}(1-\pi)$; and so on.

For the case of simple contingent reinforcement,

$$\frac{u_k}{u_{k-1}} = \frac{(N-k+1)\pi_{21}c}{N} \bigg/ \frac{k\pi_{12}c}{N} = \frac{(N-k+1)\pi_{21}}{k\pi_{12}}.$$

Again we note that the u_i are independent of c . Further the ratio u_k to u_{k-1} is the same as that of

$$\binom{N}{k} \pi_{21}^k \pi_{12}^{N-k} \quad \text{to} \quad \binom{N}{k-1} \pi_{21}^{k-1} \pi_{12}^{N-k+1}.$$

Therefore the asymptotic state probabilities are the terms in the expansion of

$$\left(\frac{\pi_{21}}{\pi_{21} + \pi_{12}} + \frac{\pi_{12}}{\pi_{21} + \pi_{12}} \right)^N.$$

Explicit formulas for state probabilities are useful primarily as intermediary expressions in the derivation of other quantities. In the special case of the pattern model (unlike other types of stimulus sampling models) the strict determination of the response on any trial by the conditioning state of the trial sample permits a relatively direct empirical interpretation, for the moments of the distribution of state probabilities are identical with the moments of the response random variable. Thus in the simple contingent case we have immediately for the mean and variance of the response random variable A_∞

$$E(A_\infty) = \sum_{k=1}^N \frac{k}{N} \binom{N}{k} \left(\frac{\pi_{21}}{\pi_{21} + \pi_{12}} \right)^k \left(\frac{\pi_{12}}{\pi_{21} + \pi_{12}} \right)^{N-k} = \frac{\pi_{21}}{\pi_{21} + \pi_{12}}$$

and

$$\begin{aligned} \text{Var}(A_\infty) &= \sum_{k=1}^N \frac{k^2}{N^2} \binom{N}{k} \left(\frac{\pi_{21}}{\pi_{21} + \pi_{12}} \right)^k \left(\frac{\pi_{12}}{\pi_{21} + \pi_{12}} \right)^{N-k} - [E(A_\infty)]^2 \\ &= \frac{\pi_{21}\pi_{12}}{(\pi_{21} + \pi_{12})^2}. \end{aligned}$$

A bit of caution is needed in applying this last expression to data. If we select some fixed trial n (large enough so that the learning process may be assumed asymptotic), then the theoretical variance for the A_1 -response totals of a number of independent samples of K subjects on trial n is simply $K[\pi_{21}\pi_{12}/(\pi_{21} + \pi_{12})^2]$ by the familiar theorem for the variance of a sum of independent random variables. However, this expression does not hold for the variance of A_1 -response totals over a block of K successive trials. The additional considerations involved in the latter case are discussed in the next section.

2.2 Treatment of the Simple Noncontingent Case

In this section we shall consider various predictions that may be derived from the pattern model for simple predictive behavior in a two-choice situation with noncontingent reinforcement. Each trial in the reference experiment begins with the presentation of a ready signal; the subject's task is to respond to the signal by operating one of a pair of response keys, A_1 or A_2 , indicating his prediction as to which of two reinforcing lights will appear. The reinforcing lights are programmed by the experimenter to

occur in random sequence, exactly one on each trial, with probabilities that are constant throughout the series and independent of the subject's behavior.

For illustrative purposes, we shall use data from two experiments of this sort. In one of these, henceforth designated the 0.6 series, 30 subjects were run, each for a series of 240 trials, with probabilities of 0.6 and 0.4 for the two reinforcing lights. Details of the experimental procedure, and a more complete analysis of the data than we shall undertake here, are given in Suppes & Atkinson (1960, Chapter 10). In the other experiment, henceforth designated the 0.8 series, 80 subjects were run, each for a series of 288 trials, with probabilities of 0.8 and 0.2 for the two reinforcing lights. Details of the procedure and results have been reported by Friedman et al. (1960). A possibly important difference between the conditions of the two experiments is that in the 0.6 series the subjects were new to this type of experiment, whereas in the 0.8 series the subjects were highly practiced, having had experience with a variety of noncontingent schedules in two previous experimental sessions.

For our present purposes it will suffice to consider only the simplest possible interpretation of the experimental situation in terms of the pattern model. Let O_1 denote the more frequently occurring reinforcing light and O_2 the less frequent light. We then postulate a one-to-one correspondence between the appearance of light O_i and the reinforcing event E_i which is associated with A_i (the response of predicting O_i). Also we assume that the experimental conditions determine a set of N distinct stimulus patterns, exactly one of which is present at the onset of any given trial. Since, in experiments of the sort under consideration, the experimenter usually presents the same ready signal at the beginning of every trial, we might assume that N would necessarily equal unity. However, we shall not impose this restriction on the model. Rather, we shall let N appear as a free parameter in theoretical expressions; then we shall seek to determine from the data the value of N required to minimize the disparities between theoretical and observed values.

If the data of a particular experiment yield an estimate of N greater than unity and if, with this estimate, the model provides a satisfactory account of the empirical relationships in question, we shall conclude that the learning process proceeds as described by the model but that, regardless of the experimenter's intention, the subjects are sampling a population of stimulus patterns. The pattern effective at the onset of a given trial might comprise the experimenter's ready signal together with stimulus traces (perhaps verbally mediated) of the reinforcing events and responses of one or more preceding trials.

It will be apparent that the pattern model could scarcely be expected to

provide a completely adequate account of the data of two-choice experiments run under the conditions sketched above. First, if the stimulus patterns to which the subject responds include cues from preceding events, then it is extremely unlikely that all of the available patterns would have equal sampling probabilities as assumed in the model. Second, the different patterns must have component cues in common, and these would be expected to yield transfer effects (at least on early trials) so that the response to a pattern first sampled on trial n would be influenced by conditioning that occurred when components of that pattern were present on earlier trials. However, the pattern model assumes that all of the patterns available for sampling are distinct in the sense that reinforcement of a response to one pattern has no effect on response probabilities associated with other patterns.

Despite these complications, many investigators (e.g., Suppes & Atkinson, 1960; Estes, 1961b; Suppes & Ginsberg, 1962; Bower, 1961) have found it a useful strategy to apply the pattern model in the simple form presented in the preceding section. The goal in these applications is not the perhaps impossible one of accounting for every detail of the experimental results but rather the more modest, yet realizable, one of obtaining valuable information about various theoretical assumptions by comparing manageably simple models that embody different combinations of assumptions. This procedure is illustrated in the remainder of the section.

SEQUENTIAL PREDICTIONS. We begin our application of the pattern model with a discussion of sequential statistics. It should be emphasized that one of the major contributions of mathematical learning theory has been to provide a framework within which the sequential aspects of learning can be scrutinized. Before the development of mathematical models little attention was paid to trial-by-trial phenomena; at the present time, for many experimental problems, such phenomena are viewed as the most interesting aspect of the data.

Although we consider only the noncontingent case, the same methods may be used to obtain results for more general reinforcement schedules. We shall develop the proofs in terms of two responses, but the results hold for any number of alternatives. If there are r responses in a given experimental application, any one response can be denoted A_1 and the rest regarded as members of a single class, A_2 .

We consider first the probability of an A_1 response, given that it occurred and was reinforced on the preceding trial; that is, $\Pr(A_{1,n+1} | E_{1,n}A_{1,n})$. It is convenient to deal first with the joint probability $\Pr(A_{1,n+1}E_{1,n}A_{1,n})$ and to conditionalize later. First we note that

$$\Pr(A_{1,n+1}E_{1,n}A_{1,n}) = \sum_{i,j} \Pr(A_{1,n+1}C_{j,n+1}E_{1,n}A_{1,n}C_{i,n}), \quad (30)$$

and that $\Pr(A_{1,n+1}C_{j,n+1}E_{1,n}A_{1,n}C_{i,n})$ may be expressed in terms of conditional probabilities as

$$\Pr(A_{1,n+1} \mid C_{j,n+1}E_{1,n}A_{1,n}C_{i,n}) \Pr(C_{j,n+1} \mid E_{1,n}A_{1,n}C_{i,n}) \cdot \Pr(E_{1,n} \mid A_{1,n}C_{i,n})\Pr(A_{1,n} \mid C_{i,n}) \Pr(C_{i,n}).$$

But from the sampling and response axioms the probability of a response on trial n is determined solely by the conditioning state on trial n ; that is, the first factor in the expansion can be rewritten simply as $\Pr(A_{1,n+1} \mid C_{j,n+1})$. Further, by Axiom R1, we have

$$\Pr(A_{1,n+1} \mid C_{j,n+1}) = j/N.$$

For the noncontingent case the probability of an E_1 on any trial is independent of previous events and consequently we may write

$$\Pr(E_{1,n} \mid A_{1,n}C_{i,n}) = \pi.$$

Next, we note that

$$\Pr(C_{j,n+1} \mid E_{1,n}A_{1,n}C_{i,n}) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j; \end{cases}$$

that is, an element conditioned to A_1 is sampled on trial n (since an A_1 -response occurs on n) and thus by Axiom C2 no change in the conditioning state can occur.

Putting these results together and substituting in Eq. 30, we obtain

$$\begin{aligned} \Pr(A_{1,n+1}E_{1,n}A_{1,n}) &= \pi \sum_i \frac{i^2}{N^2} \Pr(C_{i,n+1} \mid E_{1,n}A_{1,n}C_{i,n}) \Pr(C_{i,n}) \\ &= \pi \sum_i \frac{i^2}{N^2} \Pr(C_{i,n}) \\ &= \pi\alpha_{2,n}, \end{aligned} \tag{31a}$$

and

$$\begin{aligned} \Pr(A_{1,n+1} \mid E_{1,n}A_{1,n}) &= \frac{\pi\alpha_{2,n}}{\Pr(E_{1,n}A_{1,n})} \\ &= \frac{\alpha_{2,n}}{\Pr(A_{1,n})}. \end{aligned} \tag{31b}$$

In order to express this conditional probability in terms of the parameters π , c , N , and $\Pr(A_{1,i})$, we simply substitute into Eq. 31b the expression given for $\Pr(A_{1,n})$ in Eq. 28 and the corresponding expression for $\alpha_{2,n}$ that would be given by the solution of the difference equation (Eq. 29). Unfortunately, the expression so obtained is extremely cumbersome to work with. Consequently it is usually preferable in working with data to proceed in a different way.

Suppose the data to be treated consist of proportions of occurrences of the various trigrams $A_{k,n+1}E_{j,n}A_{i,n}$ over blocks of M trials. If, for example, $M = 5$, then in the protocol

| | | | | | |
|-------|----------|----------|----------|----------|----------|
| Trial | 1 | 2 | 3 | 4 | 5 |
| Event | A_1E_1 | A_1E_1 | A_2E_1 | A_1E_1 | A_1E_2 |

There are four opportunities for such trigrams. The combination $A_{1,n+1} \cdot E_{1,n}A_{1,n}$ occurs on two of these, $A_{2,n+1}E_{1,n}A_{1,n}$ on one and $A_{1,n+1}E_{1,n}A_{2,n}$ on the other; hence the proportions of occurrence of these trigrams are 0.50, 0.25, and 0.25, respectively. To deal theoretically with quantities such as these, we need only average both sides of Eq. 31a (and the corresponding expressions for other trigrams) over the appropriate block of trials, obtaining, for example, for the block running from trial n through trial $n + M - 1$

$$p_{111} = \frac{1}{M} \sum_{n'=n}^{n+M-1} \Pr(A_{1,n'+1}E_{1,n'}A_{1,n'}) = \frac{\pi}{M} \sum_{n'=n}^{n+M-1} \alpha_{2,n'} = \pi \bar{\alpha}_2(n, M), \quad (32a)$$

where $\bar{\alpha}_2(n, M)$ is the average value of the second moment of the response probabilities over the given trial block. By strictly analogous methods we can derive theoretical expressions for other trigram proportions:

$$p_{112} = \frac{1}{M} \sum_{n'=n}^{n+M-1} \Pr(A_{1,n'+1}E_{1,n'}A_{2,n'}) = \pi \left[\left(1 - \frac{c}{N} \right) \bar{\alpha}_1(n, M) + \frac{c}{N} - \bar{\alpha}_2(n, M) \right], \quad (32b)$$

$$p_{121} = \frac{1}{M} \sum_{n'=n}^{n+M-1} \Pr(A_{1,n'+1}E_{2,n'}A_{1,n'}) = (1 - \pi) \left[\bar{\alpha}_2(n, M) - \frac{c}{N} \bar{\alpha}_1(n, M) \right], \quad (32c)$$

$$p_{122} = \frac{1}{N} \sum_{n'=n}^{n+M-1} \Pr(A_{1,n'+1}E_{2,n'}A_{2,n'}) = (1 - \pi) [\bar{\alpha}_1(n, M) - \bar{\alpha}_2(n, M)], \quad (32d)$$

and so on; the quantity $\bar{\alpha}_1(n, M)$ denoting the average A_1 -probability (or, equivalently, the proportion of A_1 -responses) over the given trial block.

Now the average moments $\bar{\alpha}_i$ can be treated as parameters to be estimated from the data in order to mediate theoretical predictions. To illustrate, let us consider a sample of data from the 0.8 series. Over the first 12 trials of the $\pi = 0.8$ series, the observed proportion of A_1 -responses

for the group of 80 subjects was 0.63 and the observed values for the trigrams of Eq. 32a-d were $p_{111} = 0.379$, $p_{112} = 0.168$, $p_{121} = 0.061$, and $p_{122} = 0.035$. Using p_{111} to estimate $\bar{\alpha}_2(1, 12)$, we have from Eq. 32a

$$0.379 = 0.8[\bar{\alpha}_2(1, 12)],$$

which yields as our estimate

$$\hat{\bar{\alpha}}_2(1, 12) = 0.47.$$

Now we are in a position to predict the value of p_{122} . Substituting the appropriate parameter values into Eq. 32d, we have

$$p_{122} = 0.2(0.63 - 0.47) = 0.032,$$

which is not far from the observed value of 0.035. Proceeding similarly, we can use Eq. 32b to estimate c/N , namely,

$$p_{112} = 0.168 = 0.8 \left[\left(1 - \frac{c}{N} \right) (0.63) + \frac{c}{N} - 0.47 \right],$$

from which

$$\frac{\hat{c}}{N} = 0.135.$$

With this estimate in hand, together with those already obtained for the first and second moments, we can substitute into Eq. 32c and predict the value of p_{121} :

$$\begin{aligned} p_{121} &= 0.2[0.47 - 0.135(0.63)] \\ &= 0.077, \end{aligned}$$

which is somewhat high in relation to the observed value of 0.061.

It should be mentioned that the simple estimation method used above for illustrative purposes would be replaced, in a serious application of the model, by a more systematic procedure. For example, one might simultaneously estimate $\bar{\alpha}_2$ and c/N by least squares, employing all eight of the p_{ijk} ; this procedure would yield a better over-all fit of the theoretical and observed values.

A limitation of the method just described is that it permits estimation of the ratio c/N but not estimation of c and N separately. Fortunately, in the asymptotic case, the expressions for the moments α_i are simple enough so that expressions for the trigrams in terms of the parameters are manageable; and it turns out to be easy to evaluate the conditioning parameter and the number of elements from these expressions. The limit of $\alpha_{1,n}$ for large n is, of course, π in the simple noncontingent case. The limit, α_2 , of $\alpha_{2,n}$ may be obtained from the solution of Eq. 29; however, a simpler method of obtaining the same result is to note that, by definition,

$$\alpha_2 = \sum_i \frac{i^2}{N^2} u_i,$$

where u_i again represents the asymptotic probability of the state in which i elements are conditioned to A_1 . Recalling that the u_i are terms of the binomial distribution, we may then write

$$\begin{aligned}\alpha_2 &= \sum_i \frac{i^2}{N^2} \binom{N}{i} \pi^i (1 - \pi)^{N-i} \\ &= \frac{1}{N^2} \sum_i i^2 \binom{N}{i} \pi^i (1 - \pi)^{N-i}.\end{aligned}$$

The summation is the second raw moment of the binomial distribution with parameter π and sample size N . Therefore

$$\begin{aligned}\alpha_2 &= \frac{N\pi(1 - \pi) + N^2\pi^2}{N^2} \\ &= \frac{\pi(1 - \pi)}{N} + \pi^2.\end{aligned}\tag{33}$$

Using Eq. 33 and the fact that $\lim \Pr(A_{1,n}) = \pi$, we have

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n+1} | E_{1,n}A_{1,n}) = \pi \left(1 - \frac{1}{N}\right) + \frac{1}{N}.\tag{34a}$$

By identical methods we can establish that

$$\lim \Pr(A_{1,n+1} | E_{1,n}A_{2,n}) = \pi \left(1 - \frac{1}{N}\right) + \frac{c}{N},\tag{34b}$$

$$\lim \Pr(A_{1,n+1} | E_{2,n}A_{1,n}) = \pi \left(1 - \frac{1}{N}\right) + \frac{1 - c}{N},\tag{34c}$$

and

$$\lim \Pr(A_{1,n+1} | E_{2,n}A_{2,n}) = \pi \left(1 - \frac{1}{N}\right).\tag{34d}$$

With these formulas in hand, we need only apply elementary probability theory to obtain expressions for dependencies of responses on responses or responses on reinforcements, namely,

$$\lim \Pr(A_{1,n+1} | A_{1,n}) = \pi + \frac{(1 - c)(1 - \pi)}{N}\tag{35a}$$

$$\lim \Pr(A_{1,n+1} | A_{2,n}) = \pi - \frac{(1 - c)\pi}{N}\tag{35b}$$

$$\lim \Pr(A_{1,n+1} | E_{1,n}) = \left(1 - \frac{c}{N}\right)\pi + \frac{c}{N}\tag{35c}$$

$$\lim \Pr(A_{1,n+1} | E_{2,n}) = \left(1 - \frac{c}{N}\right)\pi.\tag{35d}$$

Given a set of trigram proportions from the asymptotic data of a two-choice experiment, we are now in a position to achieve a test of the model by using part of the data to estimate the parameters c and N , and then substituting these estimates into Eq. 34a-d and 35a-d to predict the values of all eight of these sequential statistics. We shall illustrate this procedure with the data of the 0.6 series. The observed transition frequencies $F(A_{i,n+1} | E_{j,n}A_{k,n})$ for the last 100 trials, aggregated over subjects, are as follows:

| | A_1 | A_2 |
|----------|-------|-------|
| A_1E_1 | 748 | 298 |
| A_1E_2 | 394 | 342 |
| A_2E_1 | 462 | 306 |
| A_2E_2 | 186 | 264 |

An estimate of the asymptotic probability of an A_1 -response given an A_1E_1 -event on the preceding trial can be obtained by dividing the first entry in row one by the sum of the row; that is, $\Pr(A_1 | E_1A_1) = 748/(748 + 298) = 0.715$. But, if we turn to Eq. 34a, we note that $\lim \Pr(A_{1,n+1} | E_{1,n}A_{1,n}) = \pi(1 - 1/N) + 1/N$. Hence, letting $0.715 = 0.6(1 - 1/N) + 1/N$, we obtain an estimate⁷ of $N = 3.48$. Similarly $\Pr(A_1 | E_1A_2) = 462/(462 + 306) = 0.602$, which by Eq. 34b is an estimate of $\pi(1 - 1/N) + c/N$; using our values of π and N we find that $c/N = 0.174$ and $c = 0.605$.

Having estimated c and N , we may now generate predictions for any of our asymptotic quantities. Table 3 presents predicted and observed values for the quantities given in Eq. 34a to Eq. 35d. Considering that only two degrees of freedom have been utilized in estimating parameters, the close correspondence between theoretical and observed quantities in Table 3 may be interpreted as giving considerable support to the assumptions of the model. A similar analysis of the asymptotic data from the 0.8 series, which has been reported elsewhere (Estes, 1961b), yields comparable agreement between theoretical and observed trigram proportions. The estimate of c/N for the 0.8 data is very close to that for the 0.6 data (0.172 versus 0.174), but the estimates of c and N (0.31 and 1.84, respectively) are both smaller for the 0.8 data. It appears that the more highly practiced subjects of the 0.8 series are, on the average, sampling from a smaller population of stimulus patterns and at the same time are less responsive to the reinforcing lights than the more naive subjects of the 0.6 series.

⁷ For any one subject, N must, of course, be an integer. The fact that our estimation procedures generally yield nonintegral values for N may signify that N varies somewhat between subjects, or it may simply reflect some contamination of the data by sources of experimental error not represented in the model.

Since no model can be expected to give a perfect account of fallible data arising from real experiments (as distinguished from the idealized experiments to which the model should apply strictly), it is difficult to know how to evaluate the goodness-of-fit of theoretical to observed values. In practice, investigators usually proceed on a largely intuitive basis, evaluating the fit in a given instance against that which it appears reasonable to hope for in the light of what is known about the precision of experimental control and measurement. Statistical tests of goodness-of-fit are sometimes

Table 3 Predicted (Pattern Model) and Observed Values of Sequential Statistics for Final 100 Trials of the 0.6 Series

| Asymptotic Quantity | Predicted | Observed |
|------------------------|-----------|----------|
| $\Pr(A_1 E_1 A_1)$ | 0.715 | 0.715 |
| $\Pr(A_1 E_2 A_1)$ | 0.541 | 0.535 |
| $\Pr(A_1 E_1 A_2)$ | 0.601 | 0.601 |
| $\Pr(A_1 E_2 A_2)$ | 0.428 | 0.413 |
| $\Pr(A_1 A_1)$ | 0.645 | 0.641 |
| $\Pr(A_1 A_2)$ | 0.532 | 0.532 |
| $\Pr(A_1 E_1)$ | 0.669 | 0.667 |
| $\Pr(A_1 E_2)$ | 0.496 | 0.489 |

possible (discussions of some tests which may be used in conjunction with stimulus sampling models are given in Suppes & Atkinson, 1960); however, statistical tests are not entirely satisfactory, taken by themselves, for a sufficiently precise test will often indicate significant differences between theoretical and observed values even in cases in which the agreement is as close as could reasonably be hoped for. Generally, once a degree of descriptive accuracy that appears satisfactory to investigators familiar with the given area has been attained, further progress must come largely via differential tests of alternative models.

In the case of the two-choice noncontingent situation the ingredients for one such test are immediately at hand; for we developed in Sec. 1.3 a one-element, guessing-state model that is comparable to the N -element model with respect to the number of free parameters and that to many might seem equally plausible on psychological grounds. These models embody the all-or-none assumption concerning the formation of learned associations, but they differ in the means by which they escape the deterministic features of the simple one-element model. It will be recalled that the one-element model cannot handle the sequential statistics considered

in this section because it requires, for example, a probability of unity for response A_i on any trial following a trial on which A_i occurred and was reinforced. In the N -element model (with $N \geq 2$), there is no such constraint, for the stimulus pattern present on the preceding reinforced trial may be replaced by another pattern, possibly conditioned to a different response, on the following trial. In the guessing-state model there is no strict determinacy, since the A_i -response may occur on the reinforced trial by guessing if the subject is in state C_0 ; and, if the reinforcement were not effective, a different response might occur, again through guessing, on the following trial.

The case of the guessing-state model with $c = 0$ (c , it will be recalled, being the counterconditioning parameter) provides a two-parameter model which may be compared with the two-parameter, N -element model. We will require an expression for at least one of the trigram proportions studied in connection with the N -element model. Let us take $\Pr(A_{1,n+1}E_{1,n}A_{1,n})$ for this purpose. In Sec. 1.3 we obtained an expression for $\Pr(A_{1,n+1} | E_{1,n}A_{1,n})$ for the case in which $c = 0$, and thus we can write at once

$$\Pr(A_{1,n+1}E_{1,n}A_{1,n}) = \pi\{u_{1,n} + \frac{1}{2}u_{0,n}[c'' + (1 - c'')^{\frac{1}{2}}]\}. \quad (36a)$$

Since we are interested only in the asymptotic case, we drop the n -subscript from the right-hand side of Eq. 36a and have for the desired theoretical asymptotic expression

$$p_{111} = \pi[u_1 + u_0(1 + c'')^{\frac{1}{2}}]. \quad (36b)$$

Substituting now into Eq. 36b the expressions for u_1 and u_0 derived in Sec. 1.3, we obtain finally

$$p_{111} = \pi^2 \frac{[4\pi + (1 - \pi)\epsilon(1 - c'')]}{4[\pi^2 + (1 - \pi)^2 + \pi(1 - \pi)\epsilon]}. \quad (36c)$$

To apply this model to the asymptotic data of the 0.6 series, we may first evaluate the parameter ϵ by setting the observed proportion of A_1 -responses over the terminal 100 trials, 0.593, equal to the right-hand side of Eq. 21 and solving for ϵ , namely,

$$\begin{aligned} 0.593 &= \frac{\pi[\pi + (1 - \pi)(\epsilon/2)]}{\pi^2 + (1 - \pi)^2 + \pi(1 - \pi)\epsilon} \\ &= \frac{0.6(0.6 + 0.2\epsilon)}{0.52 + 0.24\epsilon}, \end{aligned}$$

and

$$\epsilon = 2.315.$$

Now, by introducing this value for ϵ into Eq. 36c and simplifying, we obtain the prediction

$$p_{111} = 0.2782 + 0.0775c''.$$

Since the observed value of p_{111} for the 0.6 data is 0.249, it is apparent that no matter what value (in the admissible range $0 < c'' \leq 1$) is chosen for the parameter c'' the value predicted from the guessing state model will be too large. Further analysis, using the methods illustrated, makes it clear that for no combination of parameter estimates can the guessing-state model achieve predictive accuracy comparable to that demonstrated for the N -element model in Table 3. Although this one comparison cannot be considered decisive, we might be inclined to suspect that for interpretation of two-choice, probability learning the notion of a reaccessible guessing state is on the wrong track, whereas the N -element sampling model merits further investigation.

MEAN AND VARIANCE OF A_1 RESPONSE PROPORTION. By letting $\pi_{11} = \pi_{21} = \pi$ in Eq. 28, we have immediately an expression for the probability of an A_1 -response on trial n in the noncontingent case, namely,

$$\Pr(A_{1,n}) = \pi - [\pi - \Pr(A_{1,1})] \left(1 - \frac{c}{N}\right)^{n-1}. \quad (37)$$

If we define a response random variable A_n which equals 1 or 0 as A_1 or A_2 , respectively, occurs on trial n , then the right side of Eq. 37 also represents the expectation of this random variable on trial n . The expected number of A_1 -responses in a series of K trials is then given by the summation of Eq. 37 over trials,

$$E(\bar{A}_K) = \sum_{n=1}^K E(A_n) = K\pi - \frac{N}{c} [\pi - \Pr(A_{1,1})] \left[1 - \left(1 - \frac{c}{N}\right)^K\right]. \quad (38)$$

In experimental applications we are frequently interested in the learning curve obtained by plotting the proportion of A_1 -responses per K -trial block. A theoretical expression for this learning function is readily obtained by an extension of the method used to derive Eq. 38. Let x be the ordinal number of a K -trial block running from trial $K(x-1) + 1$ to Kx , where $x = 1, 2, \dots$, and define $P(x)$ as the proportion of A_1 -responses in block x . Then

$$\begin{aligned} P(x) &= \frac{1}{K} \left[\sum_{n=1}^{Kx} \Pr(A_{1,n}) - \sum_{n=1}^{K(x-1)} \Pr(A_{1,n}) \right] \\ &= \pi - \frac{N}{Kc} [\pi - \Pr(A_{1,1})] \left[1 - \left(1 - \frac{c}{N}\right)^{Kx} \right] \left(1 - \frac{c}{N}\right)^{K(x-1)}. \quad (39a) \end{aligned}$$

The value of $\Pr(A_{1,1})$ should be in the neighborhood of 0.5 if response bias

does not exist. However, to allow for sampling deviations we may eliminate $\Pr(A_{1,1})$ in favor of the observed value of $P(1)$. This can be done in the following way. Note that

$$P(1) = \pi - \frac{N}{Kc} [\pi - \Pr(A_{1,1})] \left[1 - \left(1 - \frac{c}{N} \right)^K \right].$$

Solving for $[\pi - \Pr(A_{1,1})]$ and substituting the result in Eq. 39a, we obtain

$$P(x) = \pi - [\pi - P(1)] \left(1 - \frac{c}{N} \right)^{K(x-1)}. \quad (39b)$$

Applications of Eq. 39b to data have led to results that are satisfying in some respects but perplexing in others (see, e.g., Estes, 1959a). In most instances the implication that the learning curve should have π as an asymptote has been borne out (Estes, 1961b, 1962), and further, with a suitable choice of values for c/N , the curve represented by Eq. 39b has served to describe the course of learning. However, in experiments run with naïve subjects, as has been nearly always the case, the value of c/N required to fit the mean learning curve has been substantially smaller than the value required to handle the sequential statistics discussed in Sec. 2.1. Consider, for example, the learning curve for the 0.6 series plotted by 20 trial blocks. The observed value of $P(1)$ is 0.48 and the value of c/N estimated from the sequential statistics of the second 20-trial block is 0.12. With these parameter values, Eq. 39b yields a prediction of 0.59 for $P(3)$ and the theoretical curve is essentially at asymptote from block 4 on. The empirical learning curve, however, does not approach 0.59 until block 6 and is still short of asymptote at the end of 12 blocks, the mean proportion of A_1 -responses over the last five blocks being 0.593 (Suppes & Atkinson, 1960, p. 197).

In the case of the 0.8 series there is a similar disparity between the value of c/N estimated from the sequential statistics and the value estimated from the mean learning curve. As we have already noted, an optimal account of the trigram proportions $\Pr(A_{k,n+1}E_{j,n}A_{i,n})$ requires a c/N -value of approximately 0.17. But, if this estimate is substituted into Eq. 39a, the predicted A_1 -frequency in the first block of 12 trials is 0.67, compared to an observed value of 0.63, and the theoretical curve runs appreciably above the empirical curve for another five blocks. A c/N -value of 0.06 yields a satisfactory graduation of the observed mean curve in terms of Eq. 39a, and a fit to the trigrams that does not look bad by usual standards for prediction in learning experiments. However, comparing predictions based on the two c/N -estimates for the trigrams that contain this parameter, we see that the estimate of 0.17 is distinctly superior. For the trigrams averaged over the first 12 trials, the result is as follows:

| | Observed | Theoretical: $c/N = 0.17$ | Theoretical: $c/N = 0.06$ |
|-----------|----------|---------------------------|---------------------------|
| p_{112} | 0.168 | 0.177 | 0.144 |
| p_{121} | 0.061 | 0.073 | 0.087 |
| p_{212} | 0.121 | 0.119 | 0.152 |
| p_{221} | 0.062 | 0.053 | 0.039 |

The reason for this discrepancy in the value of c/N required to give optimal descriptions of two different aspects of the data is not clear even after much investigation. One contributing factor might be individual differences in learning rates (c/N -values) among subjects; these would be expected to affect the two types of statistics differently. However, in the case of the 0.8 series, when a more homogeneous subgroup of subjects (the middle 50% on total A_1 frequency) is analyzed, the disparity, although somewhat reduced, is not eliminated; optimal c/N -values for the mean curve and the trigram statistics are now 0.08 and 0.15, respectively. The principal source of the remaining discrepancy in this homogeneous subgroup is a much smaller increment in A_1 -frequency from the first to the second 12-trial block than is predicted. Over the first three blocks the observed proportions are 0.633, 0.665, and 0.790; the proportions predicted from Eq. 39a with $c/N = 0.15$ run 0.657, 0.779, and 0.800. A possible explanation is that in the early part of the series the subjects are responding to cues, perhaps verbal in character, which are discarded (i.e., are not resampled) when they fail to elicit consistently correct responding. An interpretation of this sort could be incorporated into the model and subjected to formal testing, but this has not yet been done. In any event, we can see that analyses of data in terms of a model enables us to determine precisely which aspects of the subjects' behavior are and which are not accounted for in terms of a particular set of assumptions.

Next to the mean learning curve, the most frequently used behavioral measure in learning experiments is perhaps the variance of response occurrences in a block of trials. Predicting this variance from a theoretical model is an exceedingly taxing assignment; for the effects of individual differences in learning rate, together with those of all sources of experimental error not represented in the model, must be expected to increase the observed response variance. However, this statistic is relatively easy to compute for the pattern model, and the derivation may serve as a prototype for derivations of similar expressions in other learning models. For simplicity, we shall limit consideration here to the case of the variance of A_1 -response frequency in a trial block after the mean curve has reached asymptote.

As a preliminary to computation of the variance, we require a statistic

that is also of interest in its own right, the covariance of A_1 -responses on any two trials; that is,

$$\begin{aligned} \text{Cov}(A_{n+k}A_n) &= E(A_{n+k}A_n) - E(A_{n+k})E(A_n) \\ &= \Pr(A_{1,n+k}A_{1,n}) - \Pr(A_{1,n+k})\Pr(A_{1,n}). \end{aligned} \tag{40}$$

First, we can establish by induction that

$$\Pr(A_{1,n+k}A_{1,n}) = \pi \Pr(A_{1,n}) - [\pi \Pr(A_{1,n}) - \Pr(A_{1,n+1}A_{1,n})] \left(1 - \frac{c}{N}\right)^{k-1}.$$

This formula is obviously an identity for $k = 1$. Thus, assuming that the formula holds for trials n and $n + k$, we may proceed to establish it for trials n and $n + k + 1$. First we use our standard procedure to expand the desired quantity in terms of reinforcing events and states of conditioning. Letting $C_{j,n}$ denote the state in which exactly j of the N elements are conditioned to response A_1 , we may write

$$\begin{aligned} \Pr(A_{1,n+k+1}A_{1,n}) &= \sum_{i,j} \Pr(A_{1,n+k+1}E_{i,n+k}C_{j,n+k}A_{1,n}) \\ &= \sum_{i,j} \Pr(A_{1,n+k+1} \mid E_{i,n+k}C_{j,n+k}A_{1,n}) \Pr(E_{i,n+k}C_{j,n+k}A_{1,n}). \end{aligned}$$

Now we can make use of the assumptions that specify the noncontingent case to simplify the second factor to

$$\pi \Pr(C_{j,n+k}A_{1,n}) \quad \text{and} \quad (1 - \pi) \Pr(C_{j,n+k}A_{1,n})$$

for $i = 1, 2$, respectively. Also, we may apply the learning axioms to the first factor to obtain

$$\begin{aligned} \Pr(A_{1,n+k+1} \mid E_{1,n+k}C_{j,n+k}A_{1,n}) &= \frac{j^2}{N^2} + \left(1 - \frac{j}{N}\right) \left[\frac{(1-c)j}{N} + \frac{c(j+1)}{N} \right] \\ &= \left(1 - \frac{c}{N}\right) \frac{j}{N} + \frac{c}{N} \end{aligned}$$

and

$$\Pr(A_{1,n+k+1} \mid E_{2,n+k}C_{j,n+k}A_{1,n}) = \left(1 - \frac{c}{N}\right) \frac{j}{N}.$$

Combining these results, we have

$$\begin{aligned} \Pr(A_{1,n+k+1}A_{1,n}) &= \sum_j \left\{ \pi \left[\left(1 - \frac{c}{N}\right) \frac{j}{N} + \frac{c}{N} \right] + (1 - \pi) \left(1 - \frac{c}{N}\right) \frac{j}{N} \right\} \Pr(C_{j,n+k}A_{1,n}) \\ &= \sum_j \left\{ \left(1 - \frac{c}{N}\right) \frac{j}{N} + \pi \frac{c}{N} \right\} \Pr(C_{j,n+k}A_{1,n}) \\ &= \left(1 - \frac{c}{N}\right) \Pr(A_{1,n+k}A_{1,n}) + \pi \frac{c}{N} \Pr(A_{1,n}). \end{aligned}$$

Substitution into this expression in terms of our inductive hypothesis yields

$$\begin{aligned} \Pr(A_{1,n+k+1}A_{1,n}) &= \left(1 - \frac{c}{N}\right) \left\{ \pi \Pr(A_{1,n}) - [\pi \Pr(A_{1,n}) - \Pr(A_{1,n+1}A_{1,n})] \right. \\ &\quad \cdot \left. \left(1 - \frac{c}{N}\right)^{k-1} \right\} + \pi \frac{c}{N} \Pr(A_{1,n}) \\ &= \pi \Pr(A_{1,n}) - [\pi \Pr(A_{1,n}) - \Pr(A_{1,n+1}A_{1,n})] \left(1 - \frac{c}{N}\right)^k, \end{aligned}$$

as required.

We wish to take the limit of the right side of Eq. 40 as $n \rightarrow \infty$ in order to obtain the covariance of the response random variable on any two trials at asymptote. The limits of $\Pr(A_{1,n})$ and $\Pr(A_{1,n+k})$ we know to be equal to π , and from Eq. 35 we have the expression

$$\pi^2 + \pi(1 - \pi)[(1 - c)/N].$$

for the limit of $\Pr(A_{1,n+1}A_{1,n})$. Making the appropriate substitutions in Eq. 40, yields the simple result

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Cov}(\mathbf{A}_{n+k}\mathbf{A}_n) &= \pi^2 - \left[\pi^2 - \pi^2 - \pi(1 - \pi) \frac{(1 - c)}{N} \right] \left(1 - \frac{c}{N}\right)^{k-1} - \pi^2 \\ &= \frac{\pi(1 - \pi)(1 - c)}{N} \left(1 - \frac{c}{N}\right)^{k-1}. \end{aligned} \quad (41)$$

Now we are ready to compute $\text{Var}(\bar{\mathbf{A}}_K)$, the variance of A_1 -response frequencies in a block of K trials at asymptote, by applying the standard theorem for the variance of a sum of random variables (Feller, 1957):

$$\text{Var}(\bar{\mathbf{A}}_K) = \lim \left[K \text{Var}(\mathbf{A}_n) + 2 \sum_{i=1}^{j-1} \sum_{j=2}^K \text{Cov}(\mathbf{A}_{n+j}\mathbf{A}_{n+i}) \right].$$

Since

$$\lim_{n \rightarrow \infty} E(\mathbf{A}_n^2) = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi,$$

the limiting variance of \mathbf{A}_n is simply

$$\lim_{n \rightarrow \infty} \text{Var}(\mathbf{A}_n) = \lim_{n \rightarrow \infty} E(\mathbf{A}_n^2) - \lim_{n \rightarrow \infty} E(\mathbf{A}_n)^2 = \pi - \pi^2.$$

Substituting this result and that for $\lim \text{Cov}(\mathbf{A}_{n+k}\mathbf{A}_n)$ into the general expression for $\text{Var}(\bar{\mathbf{A}}_K)$, we obtain

$$\begin{aligned} \text{Var}(\bar{\mathbf{A}}_K) &= K\pi(1 - \pi) + 2 \sum_{i=1}^{j-1} \sum_{j=2}^K \frac{\pi(1 - \pi)(1 - c)}{N} \left(1 - \frac{c}{N}\right)^{j-i-1} \\ &= K\pi(1 - \pi) + \frac{2\pi(1 - \pi)(1 - c)}{N} \sum_{j=1}^K \frac{N}{c} \left[1 - \left(1 - \frac{c}{N}\right)^{j-1} \right] \\ &= K\pi(1 - \pi) + \frac{2\pi(1 - \pi)(1 - c)}{c} \left\{ K - \frac{N}{c} \left[1 - \left(1 - \frac{c}{N}\right)^K \right] \right\}. \end{aligned} \quad (42)$$

Application of this formula can be conveniently illustrated in terms of the asymptotic data for the 0.8 series. Least-squares determinations

of c/N and N from the trigram proportions (using Eq. 34*a-d*) yielded estimates of 0.17 and 1.84, respectively. Inserting these values into Eq. 42, we obtain for a 48-trial block at asymptote $\text{Var}(\bar{A}_K) = 37.50$; this variance corresponds to a standard deviation of 6.12. The observed standard deviation for the final 48-trial block was 6.94. Thus the theory predicts a variance of the right order of magnitude but, as anticipated, underestimates the observed value.

From the many other statistics that can be derived from the N -element model for two-choice learning data, we take one final example, selected primarily for the purpose of reviewing the technique for deriving sequential statistics. This technique is so generally useful that the major steps should be emphasized: first, expand the desired expression in terms of the conditioning states (as done, for example, in the case of Eq. 30); second, conditionalize responses and reinforcing events on the preceding sequence of events, introducing whatever simplifications are permitted by the boundary conditions of the case under consideration; third, apply the axioms and simplify to obtain the appropriate result. These steps are now followed in deriving an expression of considerable interest in its own right—the probability of an A_1 -response following a sequence of exactly v E_1 reinforcing events:

$$\begin{aligned}
 & \Pr(A_{1,n+v} \mid E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1}) \\
 &= \frac{1}{\pi^v(1-\pi)} \Pr(A_{1,n+v} E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1}) \\
 &= \frac{1}{\pi^v(1-\pi)} \sum_{i,j} \Pr(A_{1,n+v} C_{i,n+v} E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1} C_{j,n-1}) \\
 &= \frac{1}{\pi^v(1-\pi)} \sum_{i,j} \Pr(A_{1,n+v} \mid C_{i,n+v} E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1} C_{j,n-1}) \\
 &\quad \cdot \Pr(C_{i,n+v} \mid E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1} C_{j,n-1}) \\
 &\quad \cdot \Pr(E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1} \mid C_{j,n-1}) \Pr(C_{j,n-1}) \\
 &= \sum_{i,j} \frac{i}{N} \Pr(C_{i,n+v} \mid E_{1,n+v-1} \cdots E_{1,n} E_{2,n-1} C_{j,n-1}) \Pr(C_{j,n-1}) \\
 &= \sum_{j=0}^N \left[\left(1 - c \frac{j}{N}\right) \left\{ 1 - \left(1 - \frac{j}{N}\right) \left(1 - \frac{c}{N}\right)^v \right\} \right. \\
 &\quad \left. + c \frac{j}{N} \left\{ 1 - \left(1 - \frac{j-1}{N}\right) \left(1 - \frac{c}{N}\right)^v \right\} \right] \Pr(C_{j,n-1}) \\
 &= \sum_{j=0}^N \left[1 - \left(1 - \frac{j}{N}\right) \left(1 - \frac{c}{N}\right)^v - c \frac{j}{N} \cdot \frac{1}{N} \left(1 - \frac{c}{N}\right)^v \right] \Pr(C_{j,n-1}) \\
 &= 1 - (1 - p_{n-1}) \left(1 - \frac{c}{N}\right)^v - \frac{c}{N} p_{n-1} \left(1 - \frac{c}{N}\right)^v \\
 &= 1 - \left[1 - \left(1 - \frac{c}{N}\right) p_{n-1} \right] \left(1 - \frac{c}{N}\right)^v. \tag{43}
 \end{aligned}$$

The derivation has a formidable appearance, mainly because we have spelled out the steps in more than customary detail, but each step can readily be justified. The first involves simply using the definition of a conditional probability, $\Pr(A|B) = \Pr(AB)/\Pr(B)$, together with the fact that in the simple noncontingent case $\Pr(E_{1,n}) = \pi$ and $\Pr(E_{2,n}) = 1 - \pi$ for all n and $\Pr(E_{1,n+\nu-1} \dots E_{1,n} E_{2,n-1}) = \pi^\nu(1 - \pi)$. The second step introduces the conditioning states $C_{i,n+\nu}$ and $C_{j,n-1}$, denoting the states in which i elements are conditioned to A_1 on trial $n + \nu$ and j elements on trial $n - 1$, respectively. Their insertion into the right-hand expression of line 1 is permissible, since the summation of $\Pr(C_i)$ over all values of i is unity and similarly for the summation of $\Pr(C_j)$. The third step is based solely on repeated application of the defining equation for a conditional probability, which permits the expansion

$$\Pr(ABC \dots J) = \Pr(A|BC \dots J) \Pr(B|C \dots J) \dots \Pr(J).$$

The fourth step involves assumptions of the model: the conditionalization of $A_{1,n+\nu}$ on the preceding sequence can be reduced to $\Pr(A_{1,n+\nu}|C_{i,n+\nu}) = i/N$, since, according to the theory, the preceding history affects response probability on a given trial only insofar as it determines the state of conditioning, that is, the proportion of elements conditioned to the given response. The decomposition of

$$\Pr(E_{1,n+\nu-1} \dots E_{1,n} E_{2,n-1} C_{j,n-1}) \text{ into } \pi^\nu(1 - \pi) \Pr(C_{j,n-1})$$

is justified by the special assumptions of the simple noncontingent case. The fifth step involves calculating, for each value of j on trial $n - 1$, the expected proportion of elements conditioned to A_1 on trial $n + \nu$. There are two main branches to the process, starting with state C_j on trial $n - 1$. In one, which by the axioms has probability $1 - c(j/N)$, the state of conditioning is unchanged by the E_2 -event on trial $n - 1$; then, applying Eq. 37 with $\pi = 1$ (since from trial n onward we are dealing with a sequence of E_1 's) and $\Pr(A_{1,1}) = j/N$, we obtain the expression

$$\{1 - [1 - (j/N)][1 - (c/N)]^\nu\}$$

for the expected proportion of elements connected to A_1 on trial $n + \nu$ in this branch. In the other branch, which has probability $c(j/N)$, application of Eq. 37 with $\pi = 1$ and $\Pr(A_{1,1}) = (j - 1)/N$ yields the expression $\{1 - [1 - (j - 1)/N][1 - c/N]^\nu\}$ for the expected proportion of elements connected to A_1 on trial $n + \nu$. Carrying out the summation over j and using the by-now familiar property of the model that

$$\sum_{j=0}^N \frac{j}{N} \Pr(C_{j,n-1}) = \Pr(A_{1,n-1}) = p_{n-1},$$

we finally arrive at the desired expression for probability of A_1 following exactly νE_1 's.

Application of Eq. 43 can conveniently be illustrated in terms of the 0.8 series. Using the estimate of 0.17 for c/N (obtained previously from the trigram statistics) and taking $p_{n-1} = 0.83$ (the mean proportion of A_1 -responses over the last 96 trials of the 0.8 series), we can compute the following values for the conditional response proportions:

| r | 0 | 1 | 2 | 3 | 4 |
|-------------|-------|-------|-------|-------|-------|
| Theoretical | 0.689 | 0.742 | 0.786 | 0.822 | 0.852 |
| Observed | 0.695 | 0.787 | 0.838 | 0.859 | 0.897 |

It can be seen that the trend of the theoretical values represents quite well the trend of the observed proportions over the last 96 trials. Somewhat surprisingly, the observed proportions run slightly *above* the predicted values. There is no indication here of the "negative recency effect" (decrease in A_1 -proportion with increasing length of the E_1 -sequence) reported in a number of published two-choice studies (e.g., Jarvik, 1951; Nicks, 1959). It may be significant that no negative recency effect is observed in the 0.8 series, which, it will be recalled, involved well-practiced subjects who had had experience with a wide range of π -values in preceding series. However, the effect *is* observed in the 0.6 series, conducted with subjects new to this type of experiment (cf. Suppes & Atkinson, 1960, pp. 212-213). This differential result appears to support the idea (Estes, 1962) that the negative recency phenomenon is attributable to guessing habits carried over from everyday life to the experimental situation and extinguished during a long training series conducted with noncontingent reinforcement.

We shall conclude our analysis of the N -element pattern model by proving a very general "matching theorem." The substance of this theorem is that, so long as either an E_1 or an E_2 reinforcing event occurs on each trial, the proportion of A_1 -responses for any individual subject should tend to match the proportion of E_1 -events over a sufficiently long series of trials regardless of the reinforcement schedule.

For purposes of this derivation, we shall identify by a subscript x the probabilities and events associated with the individual x in a population of subjects; thus $p_{x1,n}$ will denote probability of an A_1 -response by subject x on trial n , and $E_{x1,n}$ and $A_{x1,n}$ will denote random variables which take on the values 1 or 0 according as an E_1 -event and an A_1 -response do or do not occur in this subject's protocol on trial n . With this notation, the probability of an A_1 -response by subject x on trial $n + 1$ can be expressed by the recursion

$$p_{x1,n+1} = p_{x1,n} + \frac{c}{N}(E_{x1,n} - A_{x1,n}). \quad (44)$$

The genesis of Eq. 44 should be reasonably obvious if we recall that $P_{x1,n}$ is equal to the proportion of elements currently conditioned to the A_1 -response. This proportion can change only if an E_1 -event occurs on a trial when a stimulus pattern conditioned to A_2 is sampled, in which case $E_{x1,n} - A_{x1,n} = 1 - 0 = 1$, or if an E_2 -event occurs on a trial when a pattern conditioned to A_1 is sampled, in which case

$$E_{x1,n} - A_{x1,n} = 0 - 1 = -1.$$

In the first case the proportion of patterns conditioned to A_1 increases by $1/N$ if conditioning is effective (which has probability c) and in the second case this proportion decreases by $1/N$ (again with probability c).

Consider now a series of, say, n^* trials: we can convert Eq. 44 into an analogous recursion for response proportions over the series simply by summing both sides over n and dividing by n^* , namely,

$$\frac{1}{n^*} \sum_{n=1}^{n^*} P_{x1,n+1} = \frac{1}{n^*} \sum_{n=1}^{n^*} P_{x1,n} + \frac{1}{n^*} \frac{c}{N} \sum_{n=1}^{n^*} (E_{x1,n} - A_{x1,n}).$$

Now we subtract the first sum on the right from both sides of the equation and distribute the second sum on the right to obtain

$$\frac{P_{x1,n+1} - P_{x1,1}}{n^*} = \frac{1}{n^*} \frac{c}{N} \sum_{n=1}^{n^*} E_{x1,n} - \frac{1}{n^*} \frac{c}{N} \sum_{n=1}^{n^*} A_{x1,n}.$$

The limit of the left side of this last equation is obviously zero as $n^* \rightarrow \infty$; thus taking the limit and rearranging we have⁸

$$\lim_{n^* \rightarrow \infty} \frac{1}{n^*} \sum_{n=1}^{n^*} A_{x1,n} = \lim_{n^* \rightarrow \infty} \frac{1}{n^*} \sum_{n=1}^{n^*} E_{x1,n}. \quad (45)$$

⁸ Equation 45 holds only if the two limits exist, which will be the case if the reinforcing event on trial n depends at most on the outcomes of some finite number of preceding trials. When this restriction is not satisfied, a substantially equivalent theorem can be derived simply by dividing both sides of the equation immediately preceding by

$\frac{1}{n^*} \sum_{n=1}^{n^*} E_{x1,n}$ before passing to the limit; that is

$$\frac{P_{x1,n+1} - P_{x1,1}}{\sum_{n=1}^{n^*} E_{x1,n}} = \frac{c}{N} - \frac{c \sum_{n=1}^{n^*} A_{x1,n}}{\sum_{n=1}^{n^*} E_{x1,n}}.$$

Except for special cases in which the sum in the denominators converges, the limit of the left-hand side is zero and

$$\lim_{n^* \rightarrow \infty} \frac{\sum_{n=1}^{n^*} A_{x1,n}}{\sum_{n=1}^{n^*} E_{x1,n}} = 1.$$

To appreciate the strength of this prediction, one should note that it holds for the data of an individual subject starting at any arbitrarily selected point in a learning series, provided only that a sufficiently long block of trials following that point is available for analysis. Further, it holds regardless of the values of the parameters N and c (provided that c is not zero) and regardless of the way in which the schedule of reinforcement may depend on preceding events, the trial number, the subject's behavior, or even events outside the system (e.g., the behavior of another individual in a competitive or cooperative social situation). Examples of empirical applications of this theorem under a variety of reinforcement schedules are to be found in studies reported by Estes (1957a) and Friedman et al. (1960).

2.3 Analysis of a Paired-Comparison Learning Experiment

In order to exhibit a somewhat different interpretation of the axioms of Sec. 2.1, we shall now analyze an experiment involving a paired-comparison procedure. The experimental situation consists of a sequence of discrete trials. There are r objects, denoted A_i ($i = 1$ to r). On each trial two (or more) of these objects are presented to the subject and he is required to choose between them. Once his response has been made the trial terminates with the subject winning or losing a fixed amount of money. The subject's task is to win as frequently as possible. There are many aspects of the situation that can be manipulated by the experimenter; for example, the strategy by which the experimenter makes available certain subsets of objects from which the subjects must choose, the schedule by which the experimenter determines whether the selection of a given object leads to a win or loss, and the amount of money won or lost on each trial.

The particular experiment for which we shall essay a theoretical analysis was reported by Suppes and Atkinson (1960, Chapter 11). The problem for the subjects involved repeated choices from subsets of a set of three objects, which may be denoted A_1 , A_2 , and A_3 . On each trial one of the following subsets of objects was presented: (A_1A_2) , (A_1A_3) , (A_2A_3) , or $(A_1A_2A_3)$. The subject selected one of the objects in the presentation set; then the trial terminated with a win or a loss of a small sum of money. The four presentation sets (A_1A_2) , (A_1A_3) , (A_2A_3) , and $(A_1A_2A_3)$ occurred with equal probabilities over the series of trials. Further, if object A_i was selected on a trial, then with probability λ_i the subject lost and with probability $1 - \lambda_i$ he won the predesignated amount. More complex schedules of reinforcement could be used; of particular interest is a schedule in which the likelihood of a win following the selection of a given object depends on the other available objects in the presentation group. For

example, the probability of a win following an A_1 choice could differ, depending on whether the (A_1A_2) , (A_1A_3) , or $(A_1A_2A_3)$ presentation group occurred. The analysis of these more complex schedules does not introduce new mathematical problems and may be pursued by the same methods we shall use for the simpler case.

Before the axioms of Sec. 2.1 can be applied to the present experiment, we need to provide an interpretation of the stimulus situation confronting the subject from trial to trial. The one we select is somewhat arbitrary and in Sec. 3 alternative interpretations are examined. Of course, discrepancies between predicted and observed quantities will indicate ways in which our particular analysis of the stimulus needs to be modified.

We represent the stimulus display associated with the presentation of the pair of objects (A_iA_j) by a set S_{ij} of stimulus patterns of size N ; the triple of objects $(A_1A_2A_3)$ is represented by a set of stimulus patterns S_{123} of size N^* . Thus there are four sets of stimulus patterns, and we assume that the sets are pairwise disjoint (i.e., have no patterns in common). Since, in the model under consideration, the stimulus element sampled on any trial represents the full pattern of stimulation effective on the trial, one might wonder why a given combination of objects, say (A_1A_2) , should have more than one element associated with it. It might be remarked in this connection that in introducing a parameter N to represent set size we do not necessarily assume $N > 1$. We simply allow for the possibility that such variations in the situation or different orders of presentation of the same set of objects on different trials might give rise to different stimulus patterns. The assumption that the stimulus patterns associated with a given presentation set are pairwise disjoint does not seem appealing on common-sense grounds; nevertheless, it is of interest to see how far we can go in predicting the data of a paired-comparison learning experiment with the simplified model incorporating this highly restrictive assumption. Even though we cannot attempt to handle the positive and negative transfer effects that must occur between different members of the set of patterns associated with a given combination of objects during learning, we may hope to account for statistics of asymptotic data.

When the pair of objects (A_iA_j) is presented, the subject must select A_i or A_j (i.e., make response A_i or A_j); hence all pattern elements in S_{ij} become conditioned to A_i or A_j . Similarly, all elements in S_{123} become conditioned to A_1 , A_2 , or A_3 . When (A_iA_j) is presented, the subject samples a single pattern from S_{ij} and makes the response to which the pattern is conditioned.

The final step, before applying the axioms of Sec. 2.1, is to provide an interpretation of reinforcing events. Our analysis is as follows: if (A_iA_j) is presented and the A_i -object is selected, then (a) the E_i reinforcing event

occurs if the A_i -response is followed by a win and (b) the E_j -event occurs if the A_i -response is followed by a loss. If $(A_i A_j A_k)$ is presented and the A_i -object is selected, then (a) the E_i -event occurs if the A_i -response is followed by a win and (b) E_j or E_k occurs, the two events having equal probabilities, if the A_i -response is followed by a loss. This collection of rules represents only one way of relating the observable trial outcomes to the hypothetical reinforcing events. For example, when A_i is selected from $(A_i A_j A_k)$ and followed by a loss, rather than having E_j or E_k occur with equal likelihoods one might postulate that they occur with probabilities dependent on the ratio of wins following A_j -responses to wins following A_k -responses over previous trials. Many such variations in the rules of correspondence between trial outcomes and reinforcing events have been explored; these variations become particularly important when the experimenter manipulates the amount of money won or lost, the magnitude of reward in animal studies, and related variables (see Estes, 1960b; Atkinson, 1962; and Suppes & Atkinson, 1960, Chapter 11, for discussions of this point).

In analyzing the model we shall use the following notation:

$A_{i,n}^{(ij)}$ = occurrence of an A_i -response on the n th presentation of $(A_i A_j)$ [note that the reference is not to the n th trial of the experiment but to the n th presentation of $(A_i A_j)$].

$W_n^{(ij)}$ = a win on the n th presentation of $(A_i A_j)$.

$L_n^{(ij)}$ = a loss on the n th presentation of $(A_i A_j)$.

We now proceed to derive the probability of an A_i -response on the n th presentation of $(A_i A_j)$; namely $\Pr(A_{i,n}^{(ij)})$. First we note that the state of conditioning of a stimulus pattern can change only when it is sampled. Since all of the sets of stimulus patterns are pairwise disjoint, the sequence of trials on which $(A_i A_j)$ is presented forms a learning process that may be studied independently of what happens on other trials (see Axiom C4); that is, the interspersing of other types of trials between the n th and $(n + 1)$ st presentation of $(A_i A_j)$ has no effect on the conditioning of patterns in set S_{ij} .

We now want to obtain a recursive expression for $\Pr(A_{i,n}^{(ij)})$. This can be done by using the same methods employed in Sec. 2.2. But, to illustrate another approach, we proceed differently in this case.

Let $\Pr(A_{i,n}^{(ij)}) = y_n$ and $\Pr(A_{j,n}^{(ij)}) = 1 - y_n$. The possible changes in y_n are given in Fig. 6. With probability $1 - c$ no change occurs in conditioning, regardless of trial events, hence $y_{n+1} = y_n$; with probability c change can occur. If A_i occurs and is followed by a win, then the sampled element remains conditioned to A_i ; however, if a loss occurs, the sampled element (which was conditioned to A_i) becomes conditioned to A_j and thus $y_{n+1} = y_n - 1/N$. If A_j occurs and is followed by a win, then

$y_{n+1} = y_n$; however, if it is followed by a loss, the sampled element (which was conditioned to A_j) becomes conditioned to A_i , hence $y_{n+1} = y_n + 1/N$. Putting these results together, we have

$$y_{n+1} = y_n(1 - c) + y_n[cy_n(1 - \lambda_i)] + \left(y_n - \frac{1}{N}\right)(cy_n\lambda_i) + y_n[c(1 - y_n)(1 - \lambda_j)] + \left(y_n + \frac{1}{N}\right)[c(1 - y_n)\lambda_j],$$

which simplifies to the expression

$$y_{n+1} = y_n \left[1 - \frac{c}{N} (\lambda_i + \lambda_j) \right] + \frac{c}{N} \lambda_j. \tag{46}$$

Solving this difference equation, we obtain

$$\Pr(A_{i,n}^{(ij)}) = \frac{\lambda_j}{\lambda_i + \lambda_j} - \left[\frac{\lambda_j}{\lambda_i + \lambda_j} - \Pr(A_{i,1}^{(ij)}) \right] \left[1 - \frac{c}{N} (\lambda_i + \lambda_j) \right]^{n-1}. \tag{47}$$

We now consider $\Pr(A_{i,n}^{(123)})$; for simplicity let $\alpha_n = \Pr(A_{1,n}^{(123)})$, $\beta_n = \Pr(A_{2,n}^{(123)})$, and $1 - \alpha_n - \beta_n = \Pr(A_{3,n}^{(123)})$. The possible changes in α_n are given in Fig. 7. For example, on the bottom branch conditioning is effective and an A_3 -response occurs which leads to a loss; hence E_1 or E_2 occur with equal probabilities. But an A_3 followed by E_1 makes

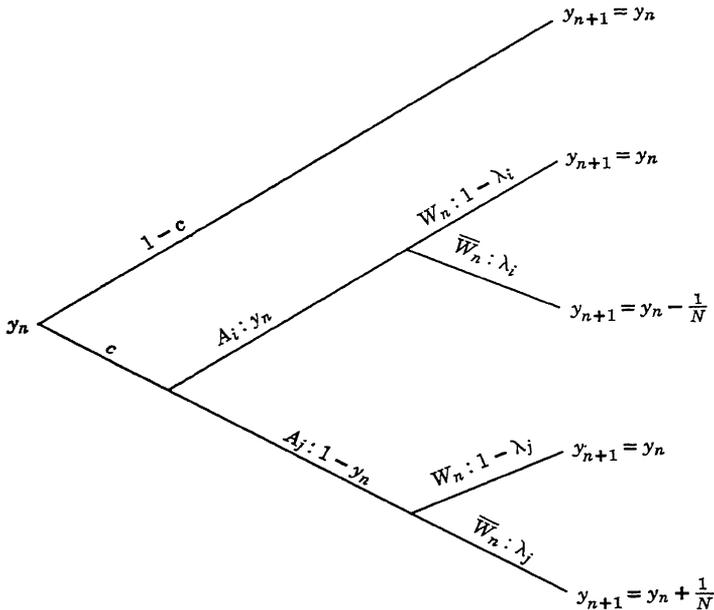


Fig. 6. Branching process for a diad probability on a paired comparison learning trial.

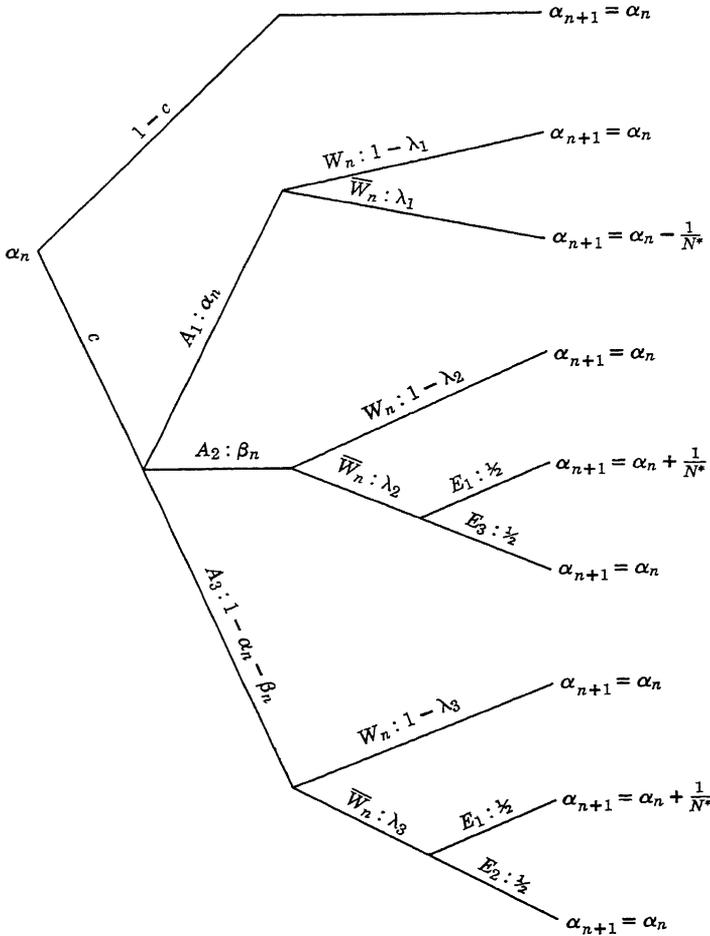


Fig. 7. Branching process for a triad probability on a paired comparison learning trial.

$\alpha_{n+1} = \alpha_n + 1/N$, and A_3 followed by E_2 makes $\alpha_{n+1} = \alpha_n$. Combining the results in this figure yields the following difference equation:

$$\begin{aligned} \alpha_{n+1} = & (1 - c)\alpha_n + \alpha_n[c\alpha_n(1 - \lambda_1)] + \left(\alpha_n - \frac{1}{N^*}\right)(c\alpha_n\lambda_1) \\ & + \alpha_n[c\beta_n(1 - \lambda_2)] + \left(\alpha_n + \frac{1}{N^*}\right)(c\beta_n\lambda_2\frac{1}{2}) + \alpha_n(c\beta_n\lambda_2\frac{1}{2}) \\ & + \alpha_n[c(1 - \alpha_n - \beta_n)(1 - \lambda_3)] + \left(\alpha_n + \frac{1}{N^*}\right)[c(1 - \alpha_n - \beta_n)\lambda_3\frac{1}{2}] \\ & + \alpha_n[c(1 - \alpha_n - \beta_n)\lambda_3\frac{1}{2}]. \end{aligned}$$

Simplifying this result, we obtain

$$x_{n-1} = x_n \left[1 - \frac{c}{2N^*} (2\lambda_1 + \lambda_3) \right] + \beta_n \frac{c}{2N^*} (\lambda_2 - \lambda_3) + \frac{c}{2N^*} \lambda_3. \quad (48a)$$

By a similar argument we obtain

$$\beta_{n-1} = \beta_n \left[1 - \frac{c}{2N^*} (2\lambda_2 + \lambda_3) \right] + \alpha_n \frac{c}{2N^*} (\lambda_1 - \lambda_3) + \frac{c}{2N^*} \lambda_3. \quad (48b)$$

Solutions for the pair of difference equations given by Eqs. 48a and 48b are well known and can be obtained by a number of different techniques (see Goldberg, 1958, pp. 130-133, or Jordan, 1950). Any solution presented can be verified by substituting into the appropriate difference equations. However for now we shall limit consideration to asymptotic results. In terms of the Markov chain property of our process it can be shown that the limits $\alpha = \lim_{n \rightarrow \infty} \alpha_n$ and $\beta = \lim_{n \rightarrow \infty} \beta_n$ exist. Letting $\alpha_{n+1} = \alpha_n = \alpha$ and $\beta_{n+1} = \beta_n = \beta$ in Eqs. 48a and 48b, we obtain

$$\begin{aligned} \alpha(2\lambda_1 + \lambda_3) &= \beta(\lambda_2 - \lambda_3) + \lambda_3 \\ \beta(2\lambda_2 + \lambda_3) &= \alpha(\lambda_1 - \lambda_3) + \lambda_3. \end{aligned}$$

Solving for α and β and rewriting, we have

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n}^{(123)}) = \frac{\lambda_2 \lambda_3}{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3}, \quad (49a)$$

$$\lim_{n \rightarrow \infty} \Pr(A_{2,n}^{(123)}) = \frac{\lambda_1 \lambda_3}{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3}, \quad (49b)$$

and

$$\lim_{n \rightarrow \infty} \Pr(A_{3,n}^{(123)}) = \frac{\lambda_1 \lambda_2}{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3}. \quad (49c)$$

The other moments of the distribution of response probabilities can be obtained by following the methods employed in Sec. 2.1; and at asymptote we can generate the entire distribution. In particular, for set S_{ij} the asymptotic probability that k patterns are conditioned to A_i and $N - k$ to A_j is simply

$$\binom{N}{k} \left(\frac{\lambda_j}{\lambda_i + \lambda_j} \right)^k \left(\frac{\lambda_i}{\lambda_i + \lambda_j} \right)^{N-k}.$$

For the set S_{123} the asymptotic probability of k_1 patterns conditioned to A_1 , k_2 to A_2 , and k_3 to A_3 (where $k_1 + k_2 + k_3 = N^*$) is

$$\frac{N^*!}{k_1! k_2! k_3!} \left(\frac{1}{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3} \right)^{N^*} (\lambda_2 \lambda_3)^{k_1} (\lambda_1 \lambda_3)^{k_2} (\lambda_1 \lambda_2)^{k_3}.$$

In analyzing data it is helpful also to examine the marginal limiting probability of an A_i -response, $\Pr(A_i)$, in addition to the other quantities already mentioned. We define $\Pr(A_i)$ as the probability of an A_i -response on any trial (regardless of the stimulus display) once the process has reached asymptote. Theoretically

$$\Pr(A_1) = \Pr(A_{1,\infty}^{(12)}) \Pr(D^{(12)}) + \Pr(A_{1,\infty}^{(13)}) \Pr(D^{(13)}) + \Pr(A_{1,\infty}^{(123)}) \Pr(D^{(123)}),$$

$$\Pr(A_2) = \Pr(A_{2,\infty}^{(12)}) \Pr(D^{(12)}) + \Pr(A_{2,\infty}^{(23)}) \Pr(D^{(23)}) + \Pr(A_{2,\infty}^{(123)}) \Pr(D^{(123)}),$$

and

$$\Pr(A_3) = 1 - \Pr(A_1) - \Pr(A_2),$$

where $\Pr(D^{(ij)})$ is the probability of presenting the pair of objects ($A_i A_j$).

The experimental results we consider were reported in preliminary form in Suppes & Atkinson (1960). Two groups, each involving 48 subjects, were run; subjects in one group won or lost one cent on each trial, and those in the other group won or lost five cents on each trial. We shall consider only the one-cent group, for an analysis of the differential effects of the two reward values requires a more elaborate interpretation of reinforcing events. Subjects were run for 400 trials with the following reinforcement schedule:

$$\lambda_1 = \frac{1}{3}, \quad \lambda_2 = \frac{6}{10}, \quad \lambda_3 = \frac{8}{10}.$$

Figure 8 presents the observed proportions of A_1 -, A_2 -, and A_3 -responses in successive 20-trial blocks. The three curves appear to be stable over the last 10 or so blocks; consequently we treat the data over trials 301 to 400 as asymptotic.

By Eq. 47 and Eq. 49a-c we may generate predictions for $\Pr(A_i^{(ij)})$ and $\Pr(A_{i,\infty}^{(123)})$. Given these values and the fact that the four presentation sets occur with equal probabilities, we may, as previously shown, generate

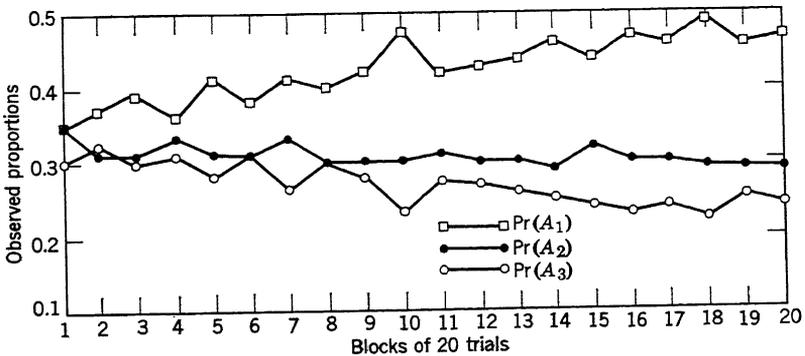


Fig. 8. Observed proportion of A_i -responses in successive 20-trial blocks for paired comparison experiment.

predictions for $\Pr(A_{i,\infty})$. The predicted values for these quantities and the observed proportions over the last 100 trials are presented in Table 4. The correspondence between predicted and observed values is very good, particularly for $\Pr(A_{i,\infty})$ and $\Pr(A_{i,\infty}^{(ij)})$. The largest discrepancy is for the triple presentation set, in which we note that the observed value of $\Pr(A_{1,\infty}^{(123)})$ is 0.041 above the predicted value of 0.507. The statistical problem of determining whether this particular difference is significant is a complex matter and we shall not undertake it here. However, it

Table 4 Theoretical and Observed Asymptotic Choice Proportions for Paired-Comparison Learning Experiment

| | Predicted | Observed |
|--------------------|-----------|----------|
| $\Pr(A_1)$ | 0.464 | 0.473 |
| $\Pr(A_2)$ | 0.302 | 0.294 |
| $\Pr(A_3)$ | 0.234 | 0.233 |
| $\Pr(A_1^{(12)})$ | 0.643 | 0.651 |
| $\Pr(A_1^{(13)})$ | 0.706 | 0.700 |
| $\Pr(A_2^{(23)})$ | 0.571 | 0.561 |
| $\Pr(A_1^{(123)})$ | 0.507 | 0.548 |
| $\Pr(A_2^{(123)})$ | 0.282 | 0.258 |
| $\Pr(A_3^{(123)})$ | 0.211 | 0.194 |

should be noted that similar discrepancies have been found in other studies dealing with three or more responses (see Gardner, 1957; Detambel, 1955), and it may be necessary, in subsequent developments of the theory, to consider some reinterpretation of reinforcing events in the multiple-response case.

In order to make predictions for more complex aspects of the data, it is necessary to obtain estimates of c , N , and N^* . Estimation procedures of the sort referred to in Sec. 2.2 are applicable, but the analysis becomes tedious and such details are not appropriate here. However, some comparisons can be made between sequential statistics that do not depend on parameter values. For example, certain nonparametric comparisons can be made between statistics where each depends on c and N but where the difference is independent of these parameters. Such comparisons are particularly helpful when they permit us to discriminate among different models without introducing the complicating factor of having to estimate parameters.

To indicate the types of comparisons that are possible, we may consider the subsequence of trials on which (A_1A_2) is presented and, in particular, the expression

$$\Pr(A_{1,n+1}^{(12)} \mid W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)});$$

that is, the probability of an A_1 -response on the $(n + 1)$ st presentation of (A_1A_2) , given that on the n th presentation of (A_1A_2) an A_1 occurred and was followed by a win and that on the $(n - 1)$ st presentation of (A_1A_2) an A_2 occurred, followed by a win. To compute this probability, we note that

$$\Pr(A_{1,n+1}^{(12)} \mid W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) = \frac{\Pr(A_{1,n+1}^{(12)} W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)})}{\Pr(W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)})}.$$

Now our problem is to compute the two quantities on the right-hand side of this equation. We first observe that

$$\begin{aligned} \Pr(A_{1,n+1}^{(12)} W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) \\ = \sum_{i,j} \Pr(A_{1,n+1}^{(12)} C_{j,n+1}^{(12)} W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)} C_{i,n-1}^{(12)}), \end{aligned}$$

where $C_{i,n}^{(12)}$ denotes the conditioning state for set S_{12} in which i elements are conditioned to A_1 and $N - i$ to A_2 on the n th presentation of (A_1A_2) . Conditionalizing and applying the axioms, we may expand the last expression into

$$\begin{aligned} \sum_{i,j} \Pr(A_{1,n+1}^{(12)} \mid C_{j,n+1}^{(12)}) \Pr(C_{j,n+1}^{(12)} \mid W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)} C_{i,n-1}^{(12)}) \\ \cdot (1 - \lambda_1) \Pr(A_{1,n}^{(12)} \mid W_{n-1}^{(12)} A_{2,n-1}^{(12)} C_{i,n-1}^{(12)}) (1 - \lambda_2) \\ \cdot \Pr(A_{2,n-1}^{(12)} \mid C_{i,n-1}^{(12)}) \Pr(C_{i,n-1}^{(12)}). \end{aligned}$$

Further, the sampling and response axioms permit the simplifications

$$\Pr(A_{1,n+1}^{(12)} \mid C_{j,n+1}^{(12)}) = \frac{j}{N},$$

$$\Pr(A_{1,n}^{(12)} \mid W_{n-1}^{(12)} A_{2,n-1}^{(12)} C_{i,n-1}^{(12)}) = \frac{i}{N},$$

and

$$\Pr(A_{2,n-1}^{(12)} \mid C_{i,n-1}^{(12)}) = \frac{N - i}{N}.$$

Finally, in order to carry out the summation, we make use of the relation

$$\Pr(C_{j,n+1}^{(12)} \mid W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)} C_{i,n-1}^{(12)}) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

which expresses the fact that no change in the conditioning state can occur if the pattern sampled leads to a win (see Axiom C2). Combining these results and simplifying, we have

$$\begin{aligned} \Pr(A_{1,n+1}^{(12)} W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) \\ = (1 - \lambda_1)(1 - \lambda_2) \sum_i \left(\frac{i}{N}\right)^2 \left(\frac{N - i}{N}\right) \Pr(C_{i,n-1}^{(12)}). \quad (50a) \end{aligned}$$

Similarly, we obtain

$$\Pr (W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) = (1 - \lambda_1)(1 - \lambda_2) \sum_i \frac{i}{N} \binom{N-i}{N} \Pr (C_{i,n-1}^{(12)}), \quad (50b)$$

and, finally, taking the quotient of the last two expressions,

$$\Pr (A_{1,n+1}^{(12)} | W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) = \frac{\sum_i \left(\frac{i}{N}\right)^2 \binom{N-i}{N} \Pr (C_{i,n-1}^{(12)})}{\sum_i \frac{i}{N} \binom{N-i}{N} \Pr (C_{i,n-1}^{(12)})}. \quad (50c)$$

We next consider the same sequential statistic but with the responses reversed on trials n and $n - 1$; namely,

$$\Pr (A_{1,n+1}^{(12)} | W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)})$$

Interestingly enough, if we compute

$$\Pr (A_{1,n+1}^{(12)} W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)})$$

and

$$\Pr (W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)}),$$

they turn out to be expressed by the right sides of Eq. 50a and 50b, respectively. Hence, for all n ,

$$\Pr (A_{1,n+1}^{(12)} | W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) = \Pr (A_{1,n+1}^{(12)} | W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)}). \quad (51)$$

Comparable predictions, of course, hold for the subsequences of trials on which $(A_1 A_3)$ or $(A_2 A_3)$ are presented.

Equation 51 provides a test of the theory which does not depend on parameter estimates. Further, it is a prediction that differentiates between this model and many other models. For example, in the next section we consider a certain class of linear models, and it can be shown that they generate the same predictions for the quantities in Table 4 as the pattern model. However, the sequential equality displayed in Eq. 51 does not hold for the linear model.

To check these predictions, we shall utilize the data over all trials of the $(A_1 A_2)$ subsequence and not restrict the analysis to asymptotic performance. Specifically, we define

$$\begin{aligned} \zeta_{112} &= \sum_n \Pr (A_{1,n+1}^{(12)} W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) \\ \zeta_{121} &= \sum_n \Pr (A_{1,n+1}^{(12)} W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)}) \\ \zeta_{12} &= \sum_n \Pr (W_n^{(12)} A_{1,n}^{(12)} W_{n-1}^{(12)} A_{2,n-1}^{(12)}) \\ \zeta_{21} &= \sum_n \Pr (W_n^{(12)} A_{2,n}^{(12)} W_{n-1}^{(12)} A_{1,n-1}^{(12)}). \end{aligned}$$

But by the results just obtained we have $\zeta_{121} = \zeta_{112}$ and $\zeta_{21} = \zeta_{12}$ for any given subject. Further, if we define ζ_{ijk} as the sum of the ζ_{ijk} 's over all subjects, then it follows that $\zeta_{121} = \zeta_{112}$, independent of intersubject differences in c and N . Similarly, $\zeta_{12} = \zeta_{21}$. Thus we have a set of predictions that are not only nonparametric but that require no restrictive assumptions on variability between subjects. Observed frequencies corresponding to these theoretical quantities are as follows:

$$\begin{array}{ll} \zeta_{121} = 140 & \zeta_{112} = 138 \\ \zeta_{21} = 243 & \zeta_{12} = 244 \\ \frac{\zeta_{121}}{\zeta_{21}} = 0.576 & \frac{\zeta_{112}}{\zeta_{12}} = 0.566. \end{array}$$

Similarly, for the (A_1A_3) subsequence,

$$\begin{array}{ll} \zeta_{131} = 67 & \zeta_{113} = 64 \\ \zeta_{31} = 120 & \zeta_{13} = 122 \\ \frac{\zeta_{131}}{\zeta_{31}} = 0.558 & \frac{\zeta_{113}}{\zeta_{13}} = 0.525. \end{array}$$

Finally, for the (A_2A_3) subsequence,

$$\begin{array}{ll} \zeta_{232} = 45 & \zeta_{223} = 49 \\ \zeta_{32} = 82 & \zeta_{23} = 87 \\ \frac{\zeta_{232}}{\zeta_{32}} = 0.549 & \frac{\zeta_{223}}{\zeta_{23}} = 0.563. \end{array}$$

Further analyses will be required to determine whether the pattern model gives an entirely satisfactory interpretation of paired-comparison learning. It is already apparent, however, that it may be difficult indeed to find another theory with equally simple machinery that will take us further in this direction than the pattern model.

3. A COMPONENT MODEL FOR STIMULUS COMPOUNDING AND GENERALIZATION

3.1 Basic Concepts; Conditioning and Response Axioms

In the preceding section we simplified our analysis of learning in terms of the N -element pattern model by assuming that all of the patterns

involved in a given experiment are disjoint or, at any rate, that generalization effects from one stimulus pattern to another are negligible. Now we shall go to the other extreme and treat problems of simple transfer of training between different stimulus situations that have elements in common, and make no reference to a learning process occurring over trials. Again the basic mathematical apparatus is that of sets and elements but with a reinterpretation that needs to be clearly distinguished from that of the pattern model. In Secs. 1 and 2 we regarded the pattern of stimulation effective on any trial as a single element sampled from a larger set of such patterns; now we shall consider the trial pattern as itself constituting a set of elements, the elements representing the various components or aspects of the stimulus situation that may be sampled by the subject in differing combinations on different trials. We proceed first to give the two basic axioms that establish the dependence of response probability on the conditioning state of the stimulus sample. Then some theorems that specify relationships between response probabilities in overlapping stimulus samples are derived and are illustrated in terms of applications to experiments on simple stimulus compounding. Consideration of the process by which trial samples are drawn from a larger stimulus population is deferred to Sec. 3.3.

The basic axioms of the component model are as follows:

Basic Axioms

- C1. *The sample s of stimulation effective on any trial is partitioned into subsets s_i ($i = 1, 2, \dots, r$, where r is the number of response alternatives), the i th subset containing the elements conditioned to (or "connected to") response A_i .*
- C2. *The probability of response A_i in the presence of the stimulus sample s is given by*

$$\Pr(A_i | s) = \frac{N(s_i)}{N(s)},$$

where $N(x)$ denotes the number of elements in the set x .

In Axiom C1 we modify the usual definition of a partition to the extent of permitting some of the subsets to be empty; that is, there may be some response alternatives that are conditioned to none of the elements of s . We do mean to assume, however, that each element of s is conditioned to exactly one response. The substance of Axiom C2 is, then, to make the probability that a given response will be evoked by s equal to the proportion of elements of s that are conditioned to that response.

3.2 Stimulus Compounding

An elementary transfer situation arises if two responses are reinforced, each in the presence of a different stimulus sample, and all or part of one sample is combined with all or part of the other to form a new test situation. To begin with a special case, let us consider an experiment conducted in the laboratory of one of the writers (W.K.E.).⁹ In one stage of the experiment a number of disjoint samples of three distinct cues drawn from a large population were used as the stimulus members of paired-associate items, and by the usual method of paired presentation one response was reinforced in the presence of some of these samples and a different response in the presence of others. The constituent cues, intended to serve as the empirical counterparts of stimulus elements, were various typewriter symbols, which for present purposes we designate by small letters *a*, *b*, *c*, etc.; the responses were the numbers "one" and "two," spoken aloud. Instructions to the subjects indicated that the cues represented symptoms and the numbers diseases with which the symptoms were associated. Following the training trials, new combinations of "symptoms" were formed, and the subjects were instructed to make their best guesses at the correct diagnoses.

Suppose now that response A_1 had been reinforced in the presence of the sample (*abc*) and response A_2 in the presence of the sample (*def*). If a test trial were given subsequently with the sample (*abd*), direct application of Axiom C2 yields the prediction that response A_1 should occur with probability $\frac{2}{3}$. Similarly, if a test were given with the sample (*ade*), response A_1 would be predicted to occur with probability $\frac{1}{3}$. Results obtained with 40 subjects, each given 24 tests of each type, were as follows:

| | | |
|--|-------|-------|
| percentage overlap of training and test sets | 0.667 | 0.333 |
| percentage response 1 to test set | 0.669 | 0.332 |

Success in bringing off a priori predictions of this sort depends not only on the basic soundness of the theory but also on one's success in realizing various simplifying assumptions in the experimental situation. As we have mentioned, it was our intention in designing this experiment to choose cues, *a*, *b*, *c*, etc., which would take on the role of stimulus elements. Actually, in order to justify our theoretical predictions, it was necessary only that the cues behave as equal-sized sets of elements. To bring out the

⁹ This experiment was conducted at Indiana University with the assistance of Miss Joan SeBreny.

importance of the equal N assumption, let us suppose that the individual cues actually correspond to sets s_a, s_b , etc., of elements. Then, given the same training (response A_1 reinforced to the combination abc and response A_2 to def) and assuming the training effective in conditioning all elements of each subset to the reinforced response, application of Axiom C2 yields for the probability of response A_1 to abd

$$\Pr(A_1 | s_a s_b s_d) = \frac{N_a + N_b}{N_a + N_b + N_d},$$

where we have used the obvious abbreviation $N(s_i) = N_i$. This equation reduces to $\Pr(A_1 | s_a s_b s_d) = \frac{2}{3}$ if $N_a = N_b = N_d$.

In this experiment we depended on common-sense considerations to choose cues that could be expected to satisfy the equal- N requirement and also counterbalanced the design of the experiment so that minor deviations might be expected to average out. Sometimes it may not be possible to depend on common-sense considerations. In that case a preliminary experiment can be utilized to check on the simplifying assumptions. Suppose, for example, we had been in doubt as to whether cues a and b would behave as equal-sized sets. To check on them, we could have run a preliminary experiment in which we reinforced, say, response A_1 to a and response A_2 to b , then tested with the compound ab . Probability of response A_1 to ab is, according to the model, given by

$$\Pr(A_1 | s_a s_b) = \frac{N_a}{N_a + N_b},$$

which should deviate in the appropriate direction from $\frac{1}{2}$ if N_a and N_b are not equal. By means of calibration experiments of this sort sets of cues satisfying the equal- N assumption can be assembled for use in further research involving applications of the model.

The expressions we have obtained for probabilities of response to stimulus compounds can readily be generalized with respect both to set sizes and to level of training. Suppose that a collection of cues a, b, c, \dots corresponds to a collection of stimulus sets s_a, s_b, s_c, \dots of sizes N_a, N_b, N_c, \dots and that some response A_j is conditioned to a proportion p_{aj} of the elements in s_a , a proportion p_{bj} of the elements in s_b , and so on. Then probability of response A_j to a compound of these cues is, by Axiom C2, expressed by the relation

$$\Pr(A_j | s_a, s_b, s_c, \dots) = \frac{N_a p_{aj} + N_b p_{bj} + N_c p_{cj} + \dots}{N_a + N_b + N_c + \dots}. \quad (52)$$

Application of Eq. 52 can be illustrated in terms of a study of probabilistic discrimination learning reported in Estes, Burke, Atkinson, & Frankmann (1957). In this study the individual cues were lights that differed

from each other only in their positions on a panel. The first stage of the experiment consisted in discrimination training according to a routine that we shall not describe here except to say that on theoretical grounds it was predicted that at the end of training the proportion of elements in a sample associated with the i th light conditioned to the first of two alternative responses would be given by $p_{i1} = i/13$. Following this training, the subjects were given compounding tests with various triads of lights. Considering, say, the triad of lights 1, 2, and 3, the values of p_{i1} should be $p_{11} = \frac{1}{13}$, $p_{21} = \frac{2}{13}$, and $p_{31} = \frac{3}{13}$, assuming $N_1 = N_2 = N_3 = N$, and substituting these values into Eq. 52, we obtain

$$\Pr(A_1 | 1, 2, 3) = \frac{N/13 + 2N/13 + 3N/13}{3N} = \frac{2}{13} = 0.15$$

as the predicted probability of response 1 to the compound 1, 2, 3. Theoretical values similarly computed for a number of triads are compared with the empirical test proportions reported by Estes et al. in Table 5.

Table 5 Theoretical and Observed Proportions of Response A_1 to Triads of Lights in Stimulus Compounding Test

| Triad | Theoretical | Observed |
|------------|-------------|----------|
| 1, 2, 3 | 0.15 | 0.22 |
| 4, 5, 6 | 0.38 | 0.31 |
| 1, 3, 11 | 0.38 | 0.41 |
| 7, 8, 9 | 0.62 | 0.59 |
| 2, 10, 12 | 0.62 | 0.58 |
| 10, 11, 12 | 0.85 | 0.77 |

An important consideration in applications of models for stimulus compounding is the question whether the experimental situation contains an appreciable amount of background stimulation in addition to the controlled stimuli manipulated by the experimenter. Suppose, for example, we are interested in the problem that a compound of two conditioned stimuli, say a light and a tone, each of which has been paired with the same unconditioned stimulus, may have a higher probability of evoking a conditioned response (CR) than either of the stimuli presented separately. To analyze this problem in terms of the present model, we may represent the light and the tone by stimulus sets s_L and s_T . Assuming that as a result of the previous reinforcement the proportions of conditioned elements in s_L and s_T (and therefore the probabilities of CR's to the stimuli taken separately) are p_L and p_T , respectively, application of Axiom C2

yields for the probability of a CR to the compound of light and tone presented together, neglecting any possible background stimulation,

$$\Pr(CR | L, T) = \frac{N_L p_L + N_T p_T}{N_L + N_T}.$$

Clearly, the probability of a CR to the compound is simply a weighted mean of p_L and p_T , and therefore its value must fall between the probabilities of a CR to the two conditioned stimuli taken separately. No "summation" effect is predicted.

Often, however, it may be unrealistic to assume that background stimulation from the apparatus and surroundings is negligible. In fact, the experimenter may have to count on an appreciable amount of background stimulation, predominantly conditioned to behaviors incompatible with the CR, to prevent "spontaneous" occurrences of the to-be-conditioned response during intervals between presentations of the experimentally controlled stimuli. Let us now expand our representation of the conditioning situation by defining a set s_b of background elements, a proportion p_b of which are conditioned to the CR. For simplicity, we shall consider only the special case of $p_b = 0$. Then the theoretical probabilities of evocation of the CR by the light, the tone, and the compound of light and sound (together with background stimulation in each case) are given by

$$\Pr(CR | L) = \frac{N_L p_L}{N_L + N_b},$$

$$\Pr(CR | T) = \frac{N_T p_T}{N_T + N_b},$$

and

$$\Pr(CR | L, T) = \frac{N_T p_T + N_L p_L}{N_T + N_L + N_b},$$

respectively. Under these conditions it is possible to obtain a summation effect. Assume, for example, that $N_T = N_L = N_b$ and $p_T > p_L$, so $\Pr(CR | T) > \Pr(CR | L)$. Taking the difference between the probability of a CR to the compound and probability of a CR to the tone alone, we have

$$\begin{aligned} \Pr(CR | L, T) - \Pr(CR | T) &= \frac{p_T + p_L}{3} - \frac{p_T}{2} \\ &= \frac{2p_T + 2p_L - 3p_T}{6} \\ &= \frac{2p_L - p_T}{6}, \end{aligned}$$

which is positive if the inequality $2p_L > p_T$ holds. Thus, in this case, probability of a *CR* to the compound will exceed probability of a *CR* to either conditioned stimulus alone, provided that p_T is not more than twice p_L .

The role of background stimuli has been particularly important in the interpretation of drive stimuli. It has been assumed (Estes, 1958, 1961a) that in simple animal learning experiments (e.g., those involving the learning of running or bar-pressing responses with food or water reward) the stimulus sample to which the animal responds at any time is compounded from several sources: the experimentally controlled conditioned stimulus (*CS*) or equivalent; stimuli, perhaps largely intra-organismic in origin, controlled by the level of food or water deprivation; and extraneous stimuli that are not systematically correlated with reward of the response undergoing training and therefore remain for the most part connected to competing responses. It is assumed further that the sizes of samples of elements associated with the *CS* and with extraneous sources s_C and s_E are independent of drive but that the size of the sample of drive-stimulus elements, s_D , increases as a function of deprivation. In most simple reward-learning experiments conditioning of the *CS* and drive cues would proceed concurrently, and it might be expected that at a given stage of learning the proportions of elements in samples from these sources conditioned to the rewarded response *R* would be equal, that is, $p_C = p_D$. If this were the case, then probability of the rewarded response would be independent of deprivation; for, letting D and D' correspond to levels of deprivation such that $N_D < N_{D'}$, we have as the theoretical probabilities of response *R* at the two deprivations,

$$\Pr(R | CS, D) = \frac{N_C p_C + N_D p_D}{N_C + N_D}$$

and

$$\Pr(R | CS, D') = \frac{N_C p_C + N_{D'} p_{D'}}{N_C + N_{D'}}.$$

If the same training were given at the two drive levels, then we would have $p_D = p_{D'}$ as well as $p_C = p_D$; in this case the difference between the two expressions is zero. Considering the same assumptions, but with extraneous cues taken explicitly into account, we arrive at a quite different picture. In this case the two expressions for response probability are

$$\Pr(R | CS, D, E) = \frac{N_C p_C + N_D p_D + N_E p_E}{N_C + N_D + N_E}$$

and

$$\Pr(R | CS, D', E) = \frac{N_C p_C + N_{D'} p_{D'} + N_E p_E}{N_C + N_{D'} + N_E}.$$

Now, letting $p_C = p_D = p_{D'} = p$ and, for simplicity, taking $p_E = 0$, we obtain for the difference

$$\begin{aligned} \Pr(R \mid CS, D', E) - \Pr(R \mid CS, D, E) \\ &= p \left[\frac{N_C + N_{D'}}{N_C + N_{D'} + N_E} - \frac{N_C + N_D}{N_C + N_D + N_E} \right] \\ &= p \frac{N_E(N_{D'} - N_D)}{(N_C + N_{D'} + N_E)(N_C + N_D + N_E)}, \end{aligned}$$

which is obviously greater than zero, given the assumption $N_{D'} > N_D$. Thus, in this theory, the principal reason why probability of the rewarded response tends, other things being equal, to be higher at higher deprivations is that the larger the sample of drive stimuli, the more effective it is in outweighing the effects of extraneous stimuli.

3.3 Sampling Axioms and Major Response Theorem of Fixed Sample Size Model

In Sec. 3.2 we considered some transfer effects which can be derived within a component model by considering only relationships among stimulus samples that have had different reinforcement histories. Generally, however, it is desirable to take account of the fact that there may not always be a one-to-one correspondence between the experimental stimulus display and the stimulation actually influencing the subject's behavior. Because of a number of factors, for example, variations in receptor-orienting responses, fluctuations in the environmental situation, or variations in excitatory states or thresholds of receptors, the subject often may sample only a portion of the stimulation made available by the experimenter. One of the chief problems of statistical learning theories has been to formulate conceptual representations of the stimulus sampling process and to develop their implications for learning phenomena. With respect to specific mathematical properties of the sampling process, component models that have appeared in the literature may be classified into two main types: (1) models assuming fixed sampling probabilities for the individual elements of a stimulus population, in which case sample size varies randomly from trial to trial; and (2) models assuming a fixed ratio between sample size and population size. The first type was first discussed by Estes and Burke (1953), the second by Estes (1950), and some detailed comparisons of the two types have been presented by Estes (1959b). In this section we shall limit consideration to models of the second type, since these are in most respects easier to work with.

In the remainder of this section we shall distinguish stimulus populations and samples by using S , with subscripts as needed, for a population and s for a sample. The sampling axioms to be utilized are as follows:

Sampling Axioms

- S1. *For any fixed, experimenter-defined stimulating situation, sample size and population size are constant over trials.*
 S2. *All samples of the same size have equal probabilities.*

A prerequisite to nearly all applications of the model is a theorem relating response probability to the state of conditioning of a stimulus population. We derive this theorem in terms of a stimulus situation S containing N elements from which a sample of size $N(s) = \sigma$ is drawn on each trial. Assuming that some number N_i of the elements of S is conditioned to response A_i , we wish to obtain an expression for the expected proportion of elements conditioned to A_i in samples drawn from S , since this proportion will, by Axiom C2, be equal to the probability of evocation of response A_i by samples from S . We begin, as usual, with the probability in which we are interested; then, using the axioms of the model as appropriate, we proceed to expand in terms of the state of conditioning and possible stimulus samples:

$$\Pr(A_i | S) = \sum_s \Pr(A_i | s) \Pr(s | S),$$

the summation being over all samples of size σ that can be drawn from S . Next, substituting expressions for the conditioned probabilities, we obtain

$$\Pr(A_i | S) = \sum_{N(s_i)=0}^{\sigma} \frac{N(s_i)}{\sigma} \frac{\binom{N_i}{N(s_i)} \binom{N - N_i}{\sigma - N(s_i)}}{\binom{N}{\sigma}}.$$

In the expression on the right $N(s_i)/\sigma$ represents the probability of A_i in the presence of a sample of size σ containing a subset s_i of elements conditioned to A_i ; the product of binomial coefficients denotes the number of ways of obtaining exactly $N(s_i)$ elements conditioned to A_i in a sample of size σ , so that the ratio of this product to the number of ways of drawing a sample of size σ is the probability of obtaining the given value of $N(s_i)/\sigma$. The resulting formula will be recognized as the familiar expression for the mean of a hypergeometric distribution (Feller, 1957, p. 218), and we have the pleasingly simple outcome that the probability of a response to the stimulating situation represented by a set S is equal to the proportion of elements of S that are conditioned to the given response:

$$\Pr(A_i | S) = \frac{N_i}{N}. \quad (53)$$

This result may seem too intuitively obvious to have needed a proof, but it should be noted that the same theorem does not hold in general for component models with fixed sampling probabilities for the elements (cf. Estes & Suppes, 1959b).

3.4 Interpretation of Stimulus Generalization

Our approach to the problem of stimulus generalization is to represent the similarity between two stimuli by the amount of overlap between two sets of elements.¹⁰ In the simplest experimental paradigm for exhibiting generalization we begin with two stimulus situations, represented by sets S_a and S_b , neither of which has any of its elements conditioned to a reference response A_1 . Training is given by reinforcement of A_1 in the presence of S_a only until the probability of A_1 in that situation reaches some value $p_{a1} > 0$. Then test trials are given in the presence of S_b , and if p_{b1} now proves to be greater than zero we say that stimulus generalization has occurred. If the axioms of the component model are satisfied, the value of p_{b1} provides, in fact, a measure of the overlap of S_a and S_b ; for, by Eq. 53, we have, immediately,

$$p_{b1} = \frac{N(S_a \cap S_b)p_{a1}}{N(S_b)},$$

where $S_a \cap S_b$ denotes the set of elements common to S_a and S_b , since the numerator of this fraction is simply the number of elements in S_b that are now conditioned to response A_1 . More generally, if the proportion of elements of S_b conditioned to A_1 before the experiment were equal to g_{b1} , not necessarily zero, the probability of response A_1 to stimulus S_b after training in S_a would be given by

$$p_{b1} = \frac{N(S_a \cap S_b)p_{a1} + [N(S_b) - N(S_a \cap S_b)]g_{b1}}{N(S_b)},$$

or, with the more compact notation $N_{ab} = N(S_a \cap S_b)$, etc.,

$$p_{b1} = \frac{N_{ab}p_{a1} + (N_b - N_{ab})g_{b1}}{N_b}. \quad (54a)$$

This relation can be put in still more convenient form by letting $N_{ab}/N_b = w_{ab}$, namely,

$$p_{b1} = w_{ab}p_{a1} + (1 - w_{ab})g_{b1}.$$

This equation may be rearranged to read

$$p_{b1} = w_{ab}(p_{a1} - g_{b1}) + g_{b1}, \quad (54b)$$

and we see that the difference ($p_{a1} - g_{b1}$) between the posttraining probability of A_1 in S_a and the pretraining probability in S_b can be regarded

¹⁰ A model similar in most essentials has been presented in Bush & Mosteller (1951b).

as the slope parameter of a linear "gradient" of generalization, in which p_{b1} is the dependent variable and the proportion of overlap between S_a and S_b is the independent variable. If we hold g_{b1} constant and let p_{a1} vary as the parameter, we generate a family of generalization gradients which have their greatest disparities at $w_{ab} = 1$ (i.e., when the test stimulus S_b is identical with S_a) and converge as the overlap between S_b and S_a decreases, until the gradients meet at $p_{b1} = g_{b1}$ when $w_{ab} = 0$. Thus the family of gradients shown in Fig. 9 illustrates the picture to be expected if a series of generalization tests is given at each of several different stages of training in S_a , or, alternatively, at several different stages of extinction following training in S_a , as was done, for example, by Guttman and Kalish (1956). The problem of "calibrating" a physical stimulus dimension to obtain a series of values that represent equal differences in the value of w_{ab} has been discussed by Carterette (1961).

The parameter w_{ab} might be regarded as an index of the similarity of S_a to S_b . In general, similarity is not a symmetrical relation, for w_{ab} is not equal to w_{ba} (w_{ab} being given by N_{ab}/N_b and the w_{ba} by N_{ab}/N_a) except in the special case $N_a = N_b$. When $N_a \neq N_b$, generalization from training with the larger set to a test with the smaller set will be greater than general-

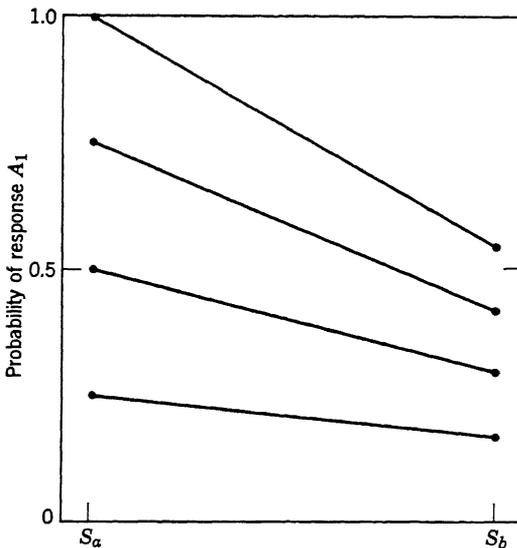


Fig. 9. Generalization from a training stimulus, S_a , to a test stimulus, S_b , at several stages of training. The parameters are $w_{ab} = 0.5$, the proportion of overlap between S_a and S_b , and $g_{b1} = 0.1$, the probability of response A_1 to S_b before training on S_a .

ization from training with the smaller set to a test with the larger set (assuming that the reinforcement given the reference response A_1 in the presence of the training set S_i establishes the same value of p_{i1} in each case before testing in S_j). We shall give no formal assumption relating size of a stimulus set to observable properties; however, it is reasonable to expect that larger sets will be associated with more intense (where the notion of intensity is applicable) or attention-getting stimuli. Thus, if S_a and S_b represent tones a and b of the same frequency but with tone a more intense than b , we should predict greater generalization if we train the reference response to a given level with a and test with b than if we train to the same level with b and test with a .

Although in the psychological literature the notion of stimulus generalization has nearly always been taken to refer to generalization along some physical continuum, such as wavelength of light or intensity of sound, it is worth noting that the set-theoretical model is not restricted to such cases. Predictions of generalization in the case of complex stimuli may be generated by first evaluating the overlap parameter w_{ab} for a given pair of situations a and b from a set of observations obtained with some particular combination of values of p_{a1} and g_{b1} and then computing theoretical values of p_{b1} for new conditions involving different levels of p_{a1} and g_{b1} . The problem of treating a simple "stimulus dimension" is of special interest, however, and we conclude our discussion of generalization by sketching one approach to this problem.¹¹

We shall consider the type of stimulus dimension that Stevens (1957) has termed *substitutive* or *metathetic*, that is, one which involves the notion of a simple ordering of stimuli along a dimension without variation in intensity or magnitude. Let us denote by Z a physical dimension of this sort, for example, wavelength of visible light, which we wish to represent by a sequence of stimulus sets. First we shall outline the properties that we wish this representation to have and then spell out the assumptions of the model more rigorously.

It is part of the intuitive basis of a substitutive dimension that one moves from point to point by exchanging some of the elements of one stimulus for new ones belonging to the next. Consequently, we assume that as values of Z change by constant increments each successive stimulus set should be generated by deleting a constant number of elements from the preceding set and adding the same number of new elements to form the

¹¹ We follow, in most respects, the treatment given by W. K. Estes and D. L. LaBerge in unpublished notes prepared for the 1957 SSRC Summer Institute in Social Science for College Teachers of Mathematics. For an approach combining essentially the same set-theoretical model with somewhat different learning assumptions, the reader is referred to Restle (1961).

next set; but, to ensure that the organism's behavior can reflect the ordering of stimuli along the Z -scale without ambiguity, we need also to assume that once an element is deleted as we go along the Z -scale it must not reappear in the set corresponding to any higher Z -value. Further, in view of the abundant empirical evidence that generalization declines in an orderly fashion as the distance between two stimuli on such a dimension increases, we must assume that (at least up to the point at which sets corresponding to larger differences in Z are disjoint) the overlap between two stimulus sets is directly related to the interval between the corresponding stimuli on the Z -scale. These properties, taken together, enable us to establish an intuitively reasonable correspondence between characteristics of a sequence of stimulus sets and the empirical notion of generalization along a dimension.

These ideas are incorporated more formally in the following set of axioms. The basis for these axioms is a stimulus dimension Z , which may be either continuous or discontinuous, a collection S_* of stimulus sets, and a function $x(Z)$ with a finite number of consecutive integers in its range. The mapping of the set (x) of scaled stimulus values onto the subsets S_i of S_* must satisfy the following axioms:

Generalization Axioms

- G1. For all $i \leq j \leq k$ in (x) , $S_i \cap S_k \subseteq S_j$.
 G2. For all $i \leq j \leq k$ in (x) , if $S_i \cap S_k \neq \emptyset$, where \emptyset is the null set, then $S_j \subseteq (S_i \cup S_k)$.
 G3. For all $h < i, j < k$ in (x) , if $i - h = k - j$, then $N_{hi} = N_{jk}$; and for all i in (x) , $N_{ii} = N$.

The set (x) may simply be a set of Z scale values or it may be a set of Z -values rescaled by some transformation. The reasons for introducing (x) are twofold. First, for mathematical simplicity we find it advisable to restrict ourselves, at least for present purposes, to a finite set of Z -values and therefore to a finite collection of stimulus sets. Second, there is no reason to believe that equal distances along physical dimensions will in general correspond to equal overlaps between stimulus sets. All that is required, however, to make the theory workable is that for any given physical dimension, wavelength of light, frequency of a tone, or whatever, we can find experimentally a transformation x such that equal distances on the x -scale do correspond to equal overlaps.

Axiom G1 states that if an element belongs to any two sets it also belongs to all sets that fall between these two sets on the x -scale. Axiom G2 states that if two sets have any common elements then all of the elements of any set falling between them belong to one or the other (or both) of the given

sets; this property ensures that the elements drop out of the sets in order as we move along the dimension. Axiom G3 describes the property that distinguishes a simple substitutive dimension from an additive, or intensity (in Stevens' terminology, *prothetic*), dimension. It should be noted that only if the number of values in the range of $x(Z)$ is no greater than $N(S_*) - N + 1$ can Axiom G3 be satisfied. This restriction is necessary in order to obtain a one-to-one mapping of the x -values into the subsets S_i of S_* .

One advantage in having the axioms set forth explicitly is that it then becomes relatively easy to design experiments bearing on various aspects of the model. Thus, to obtain evidence concerning the empirical tenability of Axiom G1, we might choose a response A_1 and a set (x) of stimuli, including a pair i and k such that $\Pr(A_1 | i) = \Pr(A_1 | k) = 0$, then train subjects with stimulus i only until $\Pr(A_1 | i) = 1$, and finally test with stimulus k . If $\Pr(A_1 | k)$ is found to be greater than zero, it must be concluded, in terms of the model, that $S_i \cap S_k \neq \emptyset$; that is, the sets corresponding to i and k have some elements in common. Given

$$\Pr(A_1 | k) > 0,$$

it must be predicted that for every stimulus j in (x) , such that $i < j < k$, $\Pr(A_1 | j) \geq \Pr(A_1 | k)$. Axiom G1 ensures that all of the elements of S_k which are now conditioned to A_1 by virtue of belonging also to S_i must be included in S_j , possibly augmented by other elements of S_i which are not in S_k .

To deal similarly with Axiom G2, we proceed in the same way to locate two members i and k of a set (x) such that $S_i \cap S_k \neq \emptyset$. Then we train subjects on both stimulus i and stimulus k until $\Pr(A_1 | i) = \Pr(A_1 | k) = 1$, response A_1 being one that before this training had probability of less than unity to all stimuli in (x) . Now, by G2, if any stimulus j falls between i and k , the set S_j must be contained entirely in the union $S_i \cup S_k$; consequently, we must predict that we will now find $\Pr(A_1 | j) = 1$ for any stimulus j such that $i \leq j \leq k$.

To evaluate Axiom G3 empirically, we require four stimuli $h < i, j < k$ such that $i - h = k - j$. If the four stimuli are all different, we can simply train subjects on h and test generalization to i , then train subjects to an equal degree on j and test generalization to k . If the amount of generalization, as measured by the probability of the test response, is the same in the two cases, then the axiom is supported. In the special case in which $h = i$ and $j = k$ we would be testing the assertion that the sets associated with different values of x are of equal size. To accomplish this test, we need only take any two neighboring values of x , say i and j , train subjects to some criterion on i and test on j , then reverse the procedure by training (different) subjects to the same criterion on j and testing on i . If the axiom is

satisfied, the amount of generalization should be the same in both directions.

Once we have introduced the notion of a dimension, it is natural to inquire whether the parameter that represents the degree of communality between pairs of stimulus sets might not be related in some simple way to a measure of distance along the dimension. With one qualification, which we mention later, the quantity $d_{ij} = 1 - w_{ij}$ could serve as a suitable measure of the distance between stimuli i and j . We can check to see whether the familiar axioms for a metric are satisfied. These axioms are

1. $d_{ij} = 0$ if and only if $i = j$,
2. $d_{ij} \geq 0$,
3. $d_{ij} = d_{ji}$,
4. $d_{ij} + d_{jk} \geq d_{ik}$,

where it is understood that i, j , and k are any members of the set (x) associated with a given dimension. The first three obviously hold, but the fourth requires a bit of analysis. To carry out a proof, we use the notation N_{ij} for the number of elements common to S_i and S_j , N_{ijk} for the number of elements in both S_i and S_j but not in S_k , and so on. The difference between the two sides of the inequality we wish to establish can be expanded in terms of this notation:

$$\begin{aligned}
 d_{ij} + d_{jk} - d_{ik} &= \left(1 - \frac{N_{ij}}{N}\right) + \left(1 - \frac{N_{jk}}{N}\right) - \left(1 - \frac{N_{ik}}{N}\right) \\
 &= \frac{1}{N} (N - N_{ij} - N_{jk} + N_{ik}) \\
 &= \frac{1}{N} (N_{ijk} + N_{ij\bar{k}} + N_{i\bar{j}k} + N_{i\bar{j}\bar{k}} - N_{i\bar{j}k} - N_{ij\bar{k}} - N_{i\bar{j}k} \\
 &\quad - N_{i\bar{j}k} + N_{i\bar{j}k} + N_{i\bar{j}\bar{k}}) \\
 &= \frac{1}{N} (N_{i\bar{j}\bar{k}} + N_{i\bar{j}k}).
 \end{aligned}$$

The last expression on the right is nonnegative, which establishes the desired inequality. To find the restrictions under which d is additive, let us assume that stimuli i, j , and k fall in the order $i < j < k$ on the dimension. Then, by Axiom G1, we know that $N_{i\bar{j}k} = 0$. However it is only in the special cases in which S_i and S_k are either overlapping or adjacent that $N_{i\bar{j}k} = 0$

and, therefore, that $d_{ij} + d_{jk} = d_{ik}$. It is possible to define an additive distance measure that is not subject to this restriction, but such extensions raise new problems and we are not able to pursue them here.

In concluding this section, we should like to emphasize one difference between the model for generalization sketched here and some of those already familiar in the literature (see, e.g., Spence, 1936; Hull, 1943). We do not postulate a particular form for generalization of response strength or excitatory tendency. Rather, we introduce certain assumptions about the properties of the set of stimuli associated with a sensory dimension; then we take these together with learning assumptions and information about reinforcement schedules as a basis for deriving theoretical gradients of generalization for particular types of experiments. Under the special conditions assumed in the example we have considered, the theory predicts that a family of linear gradients with simple properties will be observed when response probability is plotted as a function of distance from the point of reinforcement. Predictions of this sort may reasonably be tested by means of experiments in which suitable measures are taken to meet the conditions assumed in the derivations (see, e.g., Carterette, 1961); but, to deal with experiments involving different training conditions or response measures other than relative frequencies, further theoretical analysis is called for, and we must be prepared to find substantial differences in the phenotypic properties of generalization gradients derived from the same basic theory for different experimental situations.

4. COMPONENT AND LINEAR MODELS FOR SIMPLE LEARNING

In this section we combine, in a sense, the theories discussed in the preceding sections. Until now it was convenient for expositional purposes to treat the problems of learning and generalization separately. We first considered a type of learning model in which the different possible samples of stimulation from trial to trial were assumed to be entirely distinct and then turned to an analysis of generalization, or transfer, effects that could be measured on an isolated test trial following a series of learning trials. Prediction of these transfer effects depended on information concerning the state of the stimulus population just before the test trial but did not depend on information about the course of learning over preceding training trials. However, in many (perhaps most) learning situations it is not reasonable to assume that the samples, or patterns, of stimulation affecting the organism on different trials of a series are entirely disjoint; rather, they must overlap to various intermediate degrees, thus generating transfer

effects throughout the learning series. In the "component models" of stimulus sampling theory one simply takes the learning assumptions of the pattern model (Sec. 2) together with the sampling axioms and response rule of the generalization model (Sec. 3) to generate an account of learning for this more general case.

4.1 Component Models with Fixed Sample Size

As indicated earlier, the analysis of a simple learning experiment in terms of a component model is based on the representation of the stimulus as a set S of N stimulus elements from which the subject draws a sample on each trial. At any time, each element in the set S is conditioned to exactly one of the r response alternatives A_1, \dots, A_r ; by the response axiom of Sec. 3.1 the probability of a response is equal to the proportion of elements in the trial sample conditioned to that response. At the termination of a trial, if reinforcing event E_i ($i \neq 0$) occurs, then with probability c all elements in the trial sample become conditioned to response A_i . If E_0 occurs, the conditioned status of elements in the sample does not change. The conditioning parameter c plays the same role here as in the pattern model. It should be noted that in the early literature of stimulus sampling theory this parameter was usually assumed to be equal to unity.

Two general types of component models can be distinguished. For the *fixed-sample-size* model we assume that the sample size is a fixed number s throughout any given experiment. For the *independent-sampling* model we assume that the elements of the stimulus set S are sampled independently on each trial, each element having some fixed probability θ of being drawn. In this section we discuss the fixed-sample-size model and consider the case in which all possible samples of size s are sampled with equal probability.

FORMULATION FOR *RTT* EXPERIMENTS. To illustrate the model, we first consider an experimental procedure in which a particular stimulus item is given a single reinforced trial, followed by two consecutive non-reinforced test trials. The design may be conveniently symbolized RT_1T_2 . Procedures and results for a number of experiments using an *RTT* design have been reported elsewhere (Estes, 1960a; Estes, Hopkins, & Crothers, 1960; Estes, 1961b; Crothers, 1961). For simplicity, suppose we select a situation in which the probability of a correct response is zero before the first reinforcement (and in which the likelihood of a subject's obtaining correct responses by guessing is negligible on all trials). In terms of the fixed-sample-size model we can readily generate predictions for the probabilities p_{ij} of various combinations of response i on T_1 and response j

on T_2 . If $i, j = 0$ denote correct responses and $i, j = 1$ denote errors, then

$$\begin{aligned} p_{00} &= c \left(\frac{s}{N} \right)^2 \\ p_{01} &= c \left(\frac{s}{N} \right) \left(1 - \frac{s}{N} \right) \\ p_{10} &= c \left(1 - \frac{s}{N} \right) \frac{s}{N} \\ p_{11} &= 1 - c + c \left(1 - \frac{s}{N} \right)^2. \end{aligned} \tag{55}$$

To obtain the first result, we note that the correct response can occur on either trial only if conditioning occurs on the reinforced trial, which has probability c . On occasions when conditioning occurs, the whole sample of s elements becomes conditioned to the correct response and the probability of this response on each of the test trials is s/N . On occasions when conditioning does not occur on the reinforced trial, probability of a correct response remains at zero over both test trials. Note that when $s = N = 1$ this model is equivalent to the one-element model discussed in Sec. 1.1. If more than one reinforcement is given prior to T_1 , the predictions are essentially unchanged. In general, for k preceding reinforcements, the expected proportion of elements conditioned to the correct response (i.e., the probability of a correct response) at the time of the first test is

$$p_0 = 1 - \left(1 - \frac{cs}{N} \right)^k,$$

and the probability of correct responses on both T_1 and T_2 is given by

$$p_{00} = \sum_{i=1}^k \binom{k}{i} c^i (1-c)^{k-i} \left[1 - \left(1 - \frac{s}{N} \right)^i \right]^2.$$

To obtain this last expression, we note that a subject for whom i of the k reinforcements have been effective will have probability $\{1 - [1 - (s/N)]^i\}$ of making a correct response on each test, and the probability that exactly i reinforcements will be effective is $\binom{k}{i} c^i (1-c)^{k-i}$. Similarly,

$$p_{10} = p_{01} = \sum_{i=1}^k \binom{k}{i} c^i (1-c)^{k-i} \left[1 - \left(1 - \frac{s}{N} \right)^i \right] \left(1 - \frac{s}{N} \right)^i,$$

and

$$p_{11} = (1-c)^k + \sum_{i=1}^k \binom{k}{i} c^i (1-c)^{k-i} \left(1 - \frac{s}{N} \right)^{2i}.$$

If $s = N$, these expressions reduce to

$$\begin{aligned} p_{00} &= 1 - (1 - c)^k \\ p_{10} &= p_{01} = 0 \\ p_{11} &= (1 - c)^k. \end{aligned}$$

This special case appears well suited to the interpretation of data obtained by G. H. Bower (personal communication) from a study in which the T_1T_2 procedure was applied following various numbers of presentations of word-word paired-associates. For 32 subjects, each tested on 10 items, Bower reports observed proportions of $p_{00} = 0.894$, $p_{10} = p_{01} = 0.003$, and $p_{11} = 0.100$.

When applied to other *RTT* experiments, this model has, however, not yielded consistently accurate predictions. The difficulty apparently stems from the fact that our assumptions do not take account of the retention loss that is usually observed from T_1 to T_2 (see, e.g., Estes, 1961b). An extension of the model that is capable of handling retention decrement as well as the acquisition process is discussed in Sec. 4.2 below.

For *RTT* experiments, in which the probability of successful guessing is not negligible (as in paired-associate tasks involving a fixed list of responses which are known to the subject from the start), some additional considerations arise. Perhaps the most natural extension of the preceding treatment is to assume that the subject will start the experiment with a proportion $1/r$ of the elements of a given set S_i connected to the correct response and a proportion $[1 - (1/r)]$ connected to incorrect responses, r being the number of alternative responses. Then, for a fixed-sample-size model, the probability p_0 of a correct response to a given item on the first test trial after a single reinforcement is

$$\begin{aligned} p_0 &= (1 - c) \frac{1}{r} + c \left[\frac{s + (N - s)/r}{N} \right] \\ &= \left(1 - \frac{cs}{N} \right) \frac{1}{r} + \frac{cs}{N}, \end{aligned}$$

the bracketed quantity being the proportion of elements connected to the correct response in the event that the reinforcement is effective. The probabilities of various combinations of correct and incorrect responses on the two test trials are given by

$$\begin{aligned} p_{00} &= (1 - c) \frac{1}{r^2} + c\phi^2 \\ p_{10} &= p_{01} = (1 - c) \frac{1}{r} \left(1 - \frac{1}{r} \right) + c\phi(1 - \phi) \\ p_{11} &= (1 - c) \left(1 - \frac{1}{r} \right)^2 + c(1 - \phi)^2, \end{aligned} \tag{56}$$

where

$$\phi = \frac{s}{N} + \left(1 - \frac{s}{N}\right)\frac{1}{r}.$$

An alternative approach to the type of experiment in which the subject guesses on unlearned items is to assume that initially all elements are neutral, that is, are connected neither to correct nor to incorrect responses. In the presence of a sample containing only neutral elements the subject guesses, with probability $1/r$ of being correct. If the sample contains any conditioned elements, then the proportion of conditioned elements in the sample connected to the correct response determines its probability (e.g., if the sample contains nine elements, three conditioned to the correct response, two conditioned to an incorrect response, and four unconditioned, then the probability of a correct response is simply $3/5$). These assumptions seem in some respects more intuitively satisfactory than those we have considered. Perhaps the most important difference with respect to empirical implications lies in the fact that with the latter set of assumptions exposure time on test trials must be taken into account. If the stimulus exposure time is just long enough to permit a response (in terms of the theory, just long enough to permit the subject to draw a single sample of stimulus elements), then the probabilities of correct and incorrect response combinations on T_1 and T_2 are

$$\begin{aligned} p_{00} &= (1 - c)\frac{1}{r^2} + c\phi'^2, \\ p_{10} = p_{01} &= (1 - c)\frac{1}{r}\left(1 - \frac{1}{r}\right) + c\phi'(1 - \phi'), \\ p_{11} &= (1 - c)\left(1 - \frac{1}{r}\right)^2 + c(1 - \phi')^2, \end{aligned} \quad (57)$$

where

$$\phi' = 1 - \left(1 - \frac{1}{r}\right) \frac{\binom{N-s}{s}}{\binom{N}{s}}.$$

The factor $\binom{N-s}{s} / \binom{N}{s}$ is the probability that the subject will draw a sample containing none of the s elements that became conditioned on the reinforced trial; therefore $1 - \phi'$ represents the probability that a subject for whom the reinforced trial was effective nevertheless draws a sample

containing no conditioned elements and makes an incorrect guess, whereas ϕ' is the probability that such a subject will make a correct response on either test trial.

The two sets of equations (56 and 57) are formally identical and thus cannot be distinguished in application to *RTT* data. Like Eq. 55, they have the limitation of not allowing adequately for the retention loss usually observed (see, e.g., Estes, Hopkins, & Crothers, 1960); we return to this point in Sec. 4.2.

If exposure time is long enough on the test trials, then we assume that the subject continues to draw successive random samples from S and makes a response only when he finally draws a sample containing at least one conditioned element. Thus in cases in which the reinforcement has been effective on a previous trial (so that S contains a subset of s conditioned elements) the subject will eventually draw a sample containing one or more conditioned elements and will respond on the basis of these elements, thereby making a correct response with probability 1. Therefore, for the case of unlimited exposure time, $\phi' = 1$ and Eq. 57 reduces to

$$\begin{aligned} p_{00} &= (1 - c) \frac{1}{r^2} + c, \\ p_{10} = p_{01} &= (1 - c) \frac{1}{r} \left(1 - \frac{1}{r}\right), \\ p_{11} &= (1 - c) \left(1 - \frac{1}{r}\right)^2, \end{aligned} \tag{58}$$

which are identical with the corresponding equations for the one-element model of Sec. 1.2.

GENERAL FORMULATION. We turn now to the problem of deriving from the fixed-sample-size model predictions concerning the course of learning over an experiment consisting of a sequence of trials run under some prescribed reinforcement schedule. We shall limit consideration to the case in which each element in S is conditioned to exactly one of the two response alternatives, A_1 or A_2 , so that there are $N + 1$ conditioning states. Again, we let C_i ($i = 0, \dots, N$) denote the state in which i elements of the set S are conditioned to A_1 and $N - i$ to A_2 . As in the pattern model, the transition probabilities among conditioning states are functions of the reinforcement schedules and the set-theoretical parameters c , s , and N . Following our approach in Sec. 2.1, we restrict the analysis to cases in which the probability of reinforcement depends at most on the response on the given trial; we thereby guarantee that all elements in the transition

matrix for conditioning states are constant over trials. Thus the sequence of conditioning states can again be conceived as a Markov chain.

Transition Probabilities. Let $s_{i,n}$ denote the event of drawing a sample on trial n with i elements conditioned to A_1 and $s - i$ conditioned to A_2 . Then the probability of a one-step transition from state C_j to state C_{j+v} is given by

$$q_{i,j+v} = c \frac{\binom{N-j}{v} \binom{j}{s-v}}{\binom{N}{s}} \Pr(E_1 | s_{s-v} C_j), \quad (59a)$$

where $\Pr(E_1 | s_{s-v} C_j)$ is the probability of an E_1 -event, given conditioning state C_j and a sample with v elements conditioned to A_2 . To obtain Eq. 59a, we note that an E_1 must occur and that the subject must sample exactly v elements from the $N - j$ elements not already conditioned to A_1 ; the probability of the latter event is the number of ways of drawing samples with v elements conditioned to A_2 divided by the total number of ways of drawing samples of size s . Similarly

$$q_{i,j-v} = c \frac{\binom{N-j}{s-v} \binom{j}{v}}{\binom{N}{s}} \Pr(E_2 | s_v C_j) \quad (59b)$$

and

$$q_{i,j} = 1 - c + c \left[\frac{\binom{j}{s}}{\binom{N}{s}} \Pr(E_1 | s_s C_j) + \frac{\binom{N-j}{s}}{\binom{N}{s}} \Pr(E_2 | s_0 C_j) + \Pr(E_0 | C_j) \right]. \quad (59c)$$

Although it is an obvious conclusion, it is important for the reader to realize that the pattern model discussed in Sec. 2 is identical to the fixed-sample-size model when $s = 1$. This correspondence between the two models is indicated by the fact that Eqs. 59a, b, c reduce to Eq. 23a, b, c when we let $s = 1$.

For the simple noncontingent schedule in which only the two events E_1 and E_2 occur (with probabilities π and $1 - \pi$, respectively) Eqs. 59a, b, c

simplify to

$$q_{j,j+v} = c\pi \frac{\binom{N-j}{v} \binom{j}{s-v}}{\binom{N}{s}}, \quad (60a)$$

$$q_{j,j-v} = c(1-\pi) \frac{\binom{N-j}{s-v} \binom{j}{v}}{\binom{N}{s}}, \quad (60b)$$

$$q_{j,j} = 1 - c + c \left[\pi \frac{\binom{j}{s}}{\binom{N}{s}} + (1-\pi) \frac{\binom{N-j}{s}}{\binom{N}{s}} \right]. \quad (60c)$$

It is apparent that state C_N is an absorbing state when $\pi = 1$ and that C_0 is an absorbing state when $\pi = 0$. Otherwise, all states are ergodic.

Mean Learning Curve. Following the same techniques used in connection with Eq. 27, we obtain for the component model in the simple, noncontingent case

$$\Pr(A_{1,n}) = \pi - [\pi - \Pr(A_{1,1})] \left(1 - \frac{cs}{N}\right)^{n-1}. \quad (61)$$

This mean learning function traces out a smooth growth curve that can take any value between 0 and 1 on trial n if parameters are selected appropriately. However, it is important to note that for a given realization of the experiment the actual response probabilities for individual subjects (as opposed to expectations) can only take on the values 0, $1/N$, $2/N$, . . . , $(N-1)/N$, 1; that is, the values associated with the conditioning states. This stepwise aspect of the process is particularly important when one attempts to distinguish between this model and models that assume gradual continuous increments in the strength or probability of a response over time (Hull, 1943; Bush & Mosteller, 1955; Estes & Suppes, 1959a).

To illustrate this point, we consider an experiment on avoidance learning reported by Theios (1963). Fifty rats were used as subjects. The apparatus was a modified Miller-Mowrer electric-shock box, and the animal was always placed in the black compartment. Shortly thereafter a buzzer and light came on as the door between the compartments was opened. The correct response (A_1) was to run into the other compartment within 3 seconds. If A_1 did not occur, the subject was given a high intensity shock until it escaped into the other compartment. After 20 seconds the subject was returned to the black compartment, and another trial was given.

Each rat was run until it met a criterion of 20 consecutive successful avoidance responses.

Theios analyzed the situation in terms of a component model in which $N = 2$ and $s = 1$. Further, he assumed that $\Pr(A_{1,1}) = 0$, hence on trial 1 the subject is in conditioning state C_0 . Employing Eq. 60 with $\pi = 1$, $N = 2$, and $s = 1$, we obtain the following transition matrix:

$$\begin{array}{c} C_2 \quad C_1 \quad C_0 \\ \begin{bmatrix} 1 & 0 & 0 \\ \frac{c}{2} & 1 - \frac{c}{2} & 0 \\ 0 & c & 1 - c \end{bmatrix} \end{array}$$

The expected probability of an A_1 -response on trial n is readily obtained by specialization of Eq. 61,

$$\Pr(A_{1,n}) = 1 - \left(1 - \frac{c}{2}\right)^{n-1}.$$

Applying this model, Theios estimated $c = 0.43$ and provided an impressive account of such statistics as total errors, the mean learning curve, trial number of last error, autocorrelation of errors with lags of 1, 2, 3, and 4 trials, mean number of runs, probability of no reversals, and many others. However, for our immediate purposes we are interested in only one feature of his data; namely, whether the underlying response probabilities are actually fixed at 0, $\frac{1}{2}$, and 1, as specified by the model. First we note that it is not possible to establish the exact trial on which the subject moves from C_0 to C_1 or from C_1 to C_2 . Nevertheless, if there are some trials between the first success (A_1 -response) and the last error (A_2 -response), we can be sure that the subject is in state C_1 on these trials, for, if the subject has made one success, at least one of the two stimulus elements is conditioned to the A_1 -response; if on a later trial the subject makes an error, then, up to that trial, at least one of the elements is not conditioned to the A_1 -response. Since deconditioning does not occur in the present model, the subject must be in conditioning state C_1 . Thus, according to the model, the sequence of responses after the first success and before the last error should form a sequence of Bernoulli trials with constant probability $p = q = \frac{1}{2}$ of an A_1 -response. Theios has applied several statistical tests to check this hypothesis and none suggests that the assumption is incorrect. For example, the response sequences for the trials between the first success and last error were divided into blocks of four trials and the number of A_1 -responses in each block was counted. The obtained frequencies for 0, 1, 2, 3, and 4 successes were 2, 12, 17, 15, and 4, respectively;

the predicted binomial frequencies were 3.1, 12.5, 18.5, 12.5, and 3.1. The correspondence between predicted and observed frequencies is excellent, as indicated by a χ^2 goodness-of-fit test that yielded a value of 1.47 with 4 degrees of freedom.

Theios has applied the same analysis to data from an experiment by Solomon and Wynne (1953), in which dogs were required to learn an avoidance response. The findings with regard to the binomial property on trials after the first success and before the last error are in agreement with his own data but suggest that the binomial parameter is other than $\frac{1}{2}$. From a stimulus sampling viewpoint this observation would suggest that the two elements are not sampled with equal probabilities. For a detailed discussion of this Bernoulli stepwise aspect of certain stimulus sampling models, related statistical tests, and a review of relevant experimental data the reader is referred to Suppes & Ginsberg (1963).

The mean learning curve for the fixed sample size model given by Eq. 60 is identical to the corresponding equation for the pattern model with the sampling ratio cs/N taking the role of c/N . However, we need not look far to find a difference in the predictions generated by the two models. If we define $\alpha_{2,n}$ as in Eq. 29, that is,

$$\alpha_{2,n} = \sum_{i=0}^N \frac{i^2}{N^2} \Pr(C_{i,n}),$$

then by carrying out the summation, using the same methods as in the case of Eq. 27, we obtain

$$\begin{aligned} \alpha_{2,n} = & \left[1 - \frac{2cs}{N} + \frac{cs(s-1)}{N(N-1)} \right] \alpha_{2,n-1} + \frac{c}{N} \left[\frac{s}{N} - \frac{s(s-1)}{N(N-1)} \right] \alpha_{1,n-1} \\ & + 2c\pi \left(\frac{s}{N} - \frac{s^2}{N^2} \right) \alpha_{1,n-1} + \frac{c\pi s^2}{N^2}. \end{aligned} \quad (62)$$

The asymptotic variance of the response probabilities for the component model is simply

$$\sigma_{\infty}^2 = \alpha_{2,\infty} - [\Pr(A_{1,\infty})]^2.$$

Letting $\alpha_{2,n} = \alpha_{2,n-1} = \alpha_{2,\infty}$, noting that $\Pr(A_{1,\infty}) = \pi$ and carrying out the appropriate computations, we obtain

$$\sigma_{\infty}^2 = \frac{\pi(1-\pi)}{N} \left[\frac{N+(N-2)s}{2N-s-1} \right]. \quad (63)$$

This asymptotic variance of the response probabilities depends in relatively simple ways on s and N . If we hold N fixed and differentiate with respect to s , we find that σ_∞^2 increases monotonically with s ; in particular, then, this variance for a fixed sample size model with $s > 1$ is larger than that of the pattern model with the same number of elements. If we hold the sampling ratio s/N fixed and take the partial derivative with respect to N , we find σ_∞^2 to be a decreasing function of N . In the limit, if $N \rightarrow \infty$ in such a way that $s/N = \theta$ remains constant, then

$$\sigma_\infty^2 \longrightarrow \pi(1 - \pi) \frac{\theta}{2 - \theta}, \quad (64)$$

which, we shall see later, is the variance for the linear model (Estes & Suppes, 1959a). In contrast, for the pattern model the variance of the p -values approaches 0 as N becomes large. We return to comparisons between the two models in Sec. 4.3.

Sequential Predictions. We now examine some sequential statistics for the fixed-sample-size model which later will help to clarify relationships among the various stimulus sampling models. As in previous cases (e.g., Eq. 31a), we give results only for the noncontingent case in which $\Pr(E_{0,n}) = 0$ and $r = 2$.

Consider, first, $\Pr(A_{1,n+1} | E_{1,n})$. By taking account of the conditioning states on trial $n + 1$ and trial n and also the sample on trial n we may write

$$\Pr(A_{1,n+1} | E_{1,n}) = \frac{1}{\Pr(E_{1,n})} \sum_{i,j,k} \Pr(A_{1,n+1} C_{j,n+1} E_{1,n} s_{i,n} C_{k,n}),$$

where, as before, $s_{i,n}$ denotes the event of drawing a sample on trial n with i elements conditioned to A_1 and $s - i$ conditioned to A_2 . Conditionalizing, with our learning axioms in mind, we obtain

$$\begin{aligned} \Pr(A_{1,n+1} | E_{1,n}) &= \frac{1}{\Pr(E_{1,n})} \sum_{i,j,k} \Pr(A_{1,n+1} | C_{j,n+1}) \Pr(C_{j,n+1} | E_{1,n} s_{i,n} C_{k,n}) \\ &\quad \cdot \Pr(E_{1,n} | s_{i,n} C_{k,n}) \Pr(s_{i,n} | C_{k,n}) \Pr(C_{k,n}). \end{aligned}$$

But for our reinforcement procedures $\Pr(E_{1,n}) = \Pr(E_{1,n} | s_{i,n} C_{k,n})$. Further

$$\Pr(C_{j,n+1} | E_{1,n} s_{i,n} C_{k,n}) = \begin{cases} c & \text{if } j = k + s - i, \\ 1 - c & \text{if } j = k, \\ 0 & \text{otherwise;} \end{cases}$$

that is, the $s - i$ elements in the sample originally conditioned to A_2 now become conditioned to A_1 with probability c , hence a move from state C_k to C_{k+s-i} occurs. Also, as noted with regard to Eq. 59,

$$\Pr(s_{i,n} | C_{k,n}) = \frac{\binom{k}{i} \binom{N-k}{s-i}}{\binom{N}{s}}.$$

Substitution of these results in our last expression for $\Pr(A_{1,n+1} | E_{1,n})$ yields

$$\Pr(A_{1,n+1} | E_{1,n}) = \sum_{i,k} \left[c \frac{k+s-i}{N} + (1-c) \frac{k}{N} \right] \frac{\binom{k}{i} \binom{N-k}{s-i}}{\binom{N}{s}} \Pr(C_{k,n}).$$

We now need the fact that the first raw moment of the hypergeometric distribution is

$$\sum_{i=0}^k i \frac{\binom{k}{i} \binom{N-k}{s-i}}{\binom{N}{s}} = \frac{sk}{N},$$

permitting the simplification

$$\Pr(A_{1,n+1} | E_{1,n}) = \sum_k \left[\frac{cs}{N} + \frac{k}{N} \left(1 - \frac{cs}{N} \right) \right] \Pr(C_{k,n});$$

but, by definition,

$$\Pr(A_{1,n}) = \sum_k \frac{k}{N} \Pr(C_{k,n}),$$

whence

$$\Pr(A_{1,n+1} | E_{1,n}) = \left(1 - \frac{cs}{N} \right) \Pr(A_{1,n}) + \frac{cs}{N}. \quad (65a)$$

By the same method of proof we may show that

$$\Pr(A_{1,n+1} | E_{2,n}) = \left(1 - \frac{cs}{N} \right) \Pr(A_{1,n}) \quad (65b)$$

Finally, for comparison with other models, we present the expressions for

$\Pr(A_{k,n+1}E_{j,n}A_{i,n})$. Derivations of these probabilities are based on the same methods used in connection with Eq. 61a.

$$\Pr(A_{1,n+1}E_{1,n}A_{1,n}) = \pi \left\{ \left[1 - \frac{c(s-1)}{N-1} \right] \alpha_{2,n} + \frac{c(s-1)}{N-1} \alpha_{1,n} \right\}. \quad (66a)$$

$$\Pr(A_{1,n+1}E_{1,n}A_{2,n}) = \pi \left\{ \frac{cs}{N} (1 - \alpha_{1,n}) + \left[1 - \frac{c(s-1)}{N-1} \right] (\alpha_{1,n} - \alpha_{2,n}) \right\}. \quad (66b)$$

$$\Pr(A_{1,n+1}E_{2,n}A_{1,n}) = (1 - \pi) \left\{ \left[1 - \frac{c(s-1)}{N-1} \right] \alpha_{2,n} - \left[\frac{cs}{N} - \frac{c(s-1)}{N-1} \right] \alpha_{1,n} \right\}. \quad (66c)$$

$$\Pr(A_{1,n+1}E_{2,n}A_{2,n}) = (1 - \pi) \left[1 - \frac{c(s-1)}{N-1} \right] (\alpha_{1,n} - \alpha_{2,n}). \quad (66d)$$

$$\Pr(A_{2,n+1}E_{1,n}A_{1,n}) = \pi \left[1 - \frac{c(s-1)}{N-1} \right] (\alpha_{1,n} - \alpha_{2,n}). \quad (66e)$$

$$\Pr(A_{2,n+1}E_{1,n}A_{2,n}) = \pi \left\{ \left(1 - \frac{cs}{N} \right) (1 - \alpha_{1,n}) - \left[1 - \frac{c(s-1)}{N-1} \right] (\alpha_{1,n} - \alpha_{2,n}) \right\}. \quad (66f)$$

$$\Pr(A_{2,n+1}E_{2,n}A_{1,n}) = (1 - \pi) \left\{ \left[1 + \frac{cs}{N} - \frac{c(s-1)}{N-1} \right] \alpha_{1,n} - \left[1 - \frac{c(s-1)}{N-1} \right] \alpha_{2,n} \right\}. \quad (66g)$$

$$\Pr(A_{2,n+1}E_{2,n}A_{2,n}) = (1 - \pi) \left\{ 1 - \alpha_{1,n} - \left[1 - \frac{c(s-1)}{N-1} \right] (\alpha_{1,n} - \alpha_{2,n}) \right\}. \quad (66h)$$

Application of these equations to the corresponding set of trigram proportions for a preasymptotic trial block is not particularly rewarding. The difficulty is that certain combinations of parameters, for example, $\{1 - [c(s-1)/N-1]\}(\alpha_{1,n} - \alpha_{2,n})$ and cs/N , behave as units; consequently, the basic parameters c , s , and N cannot be estimated individually and, as a result, the predictions available from the simpler N -element pattern model via Eq. 32 cannot be improved upon by use of Eq. 66. For

asymptotic data the situation is somewhat different. By substituting the limiting values for $\alpha_{1,n}$ and $\alpha_{2,n}$ in Eq. 66, that is, $\alpha_1 = \pi$ and from Eq. 63

$$\begin{aligned}\alpha_2 &= \sigma_\alpha^2 + \pi^2 = \frac{\pi(1-\pi)}{N} \left[\frac{N + (N-2)s}{2N-s-1} \right] + \pi^2 \\ &= \frac{\pi[N-2s + Ns + 2\pi(N-s)(N-1)]}{N(2N-s-1)},\end{aligned}$$

we can express the trigram probabilities $\Pr(A_{k,\infty}E_{j,\infty}A_{i,\infty})$ in terms of the basic parameters of the model. The resulting expressions are somewhat cumbersome, however, and we shall not pursue this line of analysis here.

4.2 Component Models with Stimulus Fluctuation

In Sec. 4.1, as in most of the literature on stimulus sampling models for learning, we restricted attention to the special case in which the stimulation effective on successive trials of an experiment may be considered to represent independent random samples from the population of elements available under the given experimental conditions. More generally, we would expect that the independence of successive samples would depend on the interval between trials. The concept of stimulus sampling in the model corresponds to the process of stimulation in the empirical situation. Thus sampling and resampling from a stimulus population must take time; and, if the interval between trials is sufficiently short, there will not be time to draw a completely new sample. We should expect the correlation, or degree of overlap, between successive stimulus samples to vary inversely with the intertrial interval, running from perfect overlap in the limiting case (not necessarily empirically realizable) of a zero interval to independence at sufficiently long intervals. These notions have been embodied in the *stimulus fluctuation model* (Estes, 1955a, 1955b, 1959a). In this section we shall develop the assumption of stimulus fluctuation in connection with fixed-sample-size models; consequently, the expressions derived will differ in minor respects from those of the earlier presentations (cited above) that were not restricted to the case of fixed sample size.

ASSUMPTIONS AND DERIVATION OF RETENTION CURVES. Following the convention of previous articles on stimulus fluctuation models, we denote by S^* the set of stimulus elements potentially available for sampling under a given set of experimental conditions, by S the subset of elements available for sampling at any given time, and by S' the subset of elements that are temporarily unavailable (so that $S^* = S \cup S'$). The trial sample s is in turn a subset of S ; however, in this presentation we assume for simplicity that all of the temporarily available elements are sampled on

each trial (i.e., $S = s$). We denote by N , N' , and N^* , respectively, the numbers of elements in s , S' , and S^* .

The interchange between the stimulus sample and the remainder of the population, that is, between s and S' , is assumed to occur at a constant rate over time. Specifically, we assume that during an interval Δt , which is just long enough to permit the interchange of a single element between s and S' , there is probability g that such an interchange will occur, the parameter g being constant over time. We shall limit consideration to the special case in which all stimulus elements are equally likely to participate in an interchange. With this restriction, the fluctuation process can be characterized by the difference equation

$$\begin{aligned} f(t+1) &= (1-g)f(t) + g\left\{f(t)\left(1 - \frac{1}{N}\right) + [1-f(t)]\frac{1}{N'}\right\} \\ &= \left[1 - g\left(\frac{1}{N} + \frac{1}{N'}\right)\right]f(t) + \frac{g}{N'}, \end{aligned} \quad (67)$$

where $f(t)$ denotes the probability that any given element of S^* is in s at time t . This recursion can be solved by standard methods to yield the explicit formula

$$\begin{aligned} f(t) &= \frac{N}{N^*} - \left[\frac{N}{N^*} - f(0)\right] \left[1 - g\left(\frac{1}{N} + \frac{1}{N'}\right)\right]^t \\ &= J - [J - f(0)]a^t, \end{aligned} \quad (68)$$

where $J = N/N^*$, the proportion of all the elements in the sample, and $a = 1 - g(1/N + 1/N')$.

Equation 68 can now serve as the basis for deriving numerous expressions of experimental interest. Suppose, for example, that at the end of a conditioning (or extinction) period there were j_0 conditioned elements in S and k_0 conditioned elements in S' , the momentary probability of a conditioned response thus being $p_0 = j_0/N$. To obtain an expression for probability of a conditioned response after a rest interval of duration t , we proceed as follows. For each conditioned element in S at the beginning of the interval, we need only set $f(0) = 1$ in Eq. 68 to obtain the probability that the element is in S at time t . Similarly, for a conditioned element initially in S' we set $f(0) = 0$ in Eq. 68. Combining the two types, we obtain for the expected number of conditioned elements in S at time t

$$j_0[J - (J - 1)a^t] + k_0J(1 - a^t) = (j_0 + k_0)J - [(j_0 + k_0)J - j_0]a^t.$$

Dividing by N (and noting that $J = N/N^*$) we have, then, for the probability of a conditioned response at time t

$$\begin{aligned} p_t &= \frac{j_0 + k_0}{N^*} - \left[\frac{j_0 + k_0}{N^*} - p_0\right]a^t \\ &= p_0^* - (p_0^* - p_0)a^t, \end{aligned} \quad (69)$$

where p_0^* and p_0 denote the proportion of conditioned elements in the total population S^* and the initial proportion in S , respectively. If the rest interval begins after a conditioning period, we will ordinarily have $p_0 > p_0^*$ in which case Eq. 69 describes a decreasing function (forgetting, or spontaneous regression). If the rest interval begins after an extinction period, we will have $p_0 < p_0^*$, in which case Eq. 69 describes an increasing function (spontaneous recovery). The manner in which cases of spontaneous regression or recovery depend on the amount and spacing of previous acquisition or extinction has been discussed in detail elsewhere (Estes, 1955a).

APPLICATION TO THE *RTT* EXPERIMENT. We noted in the preceding section that the fixed-sample-size model could not provide a generally satisfactory account of *RTT* experiments because it did not allow for the retention loss usually observed between the first and second tests. It seems reasonable that this defect might be remedied by removing the restriction on independent sampling. To illustrate application of the more general model with provision for stimulus fluctuation, we again consider the case of an *RTT* experiment in which the probability of a correct response is negligible before the reinforced trial (and also on later trials if learning has not occurred). Letting t_1 and t_2 denote the intervals between R and T_1 and between T_1 and T_2 , respectively, we may obtain the following basic expressions by setting $f(0)$ equal to 1 or 0, as appropriate, in Eq. 68: For the probability that an element sampled on R is sampled again on T_1 ,

$$f_1 = J + (1 - J)a^{t_1};$$

for the probability that an element sampled on T_1 is sampled again on T_2 ,

$$f_2 = J + (1 - J)a^{t_2};$$

and for the probability that an element not sampled on T_1 is sampled on T_2 ,

$$f_3 = J(1 - a^{t_2}).$$

Assuming now that $N = 1$, so that we are dealing with a generalized form of the pattern model, we can write the probabilities of the four combinations of correct and incorrect responses on T_1 and T_2 in terms of the conditioning parameter c and the parameters f_i :

$$\begin{aligned} p_{00} &= cf_1f_2, \\ p_{01} &= cf_1(1 - f_2), \\ p_{10} &= c(1 - f_1)f_3, \\ p_{11} &= 1 - c + c(1 - f_1)(1 - f_3), \end{aligned} \tag{70}$$

where, as before, the subscripts 0 and 1 denote correct responses and errors, respectively. As they stand, Eqs. 70 are not suitable for application

to data because there are too many parameters to be estimated. This difficulty could be surmounted by adding a third test trial, for then the resulting eight observation equations

$$\begin{aligned} p_{000} &= cf_1f_2^2, \\ p_{001} &= cf_1f_2(1 - f_2), \\ p_{010} &= cf_1(1 - f_2)f_3, \end{aligned}$$

etc., would permit overdetermination of the four parameters. In the case of some published studies (e.g., Estes, 1961b) the data can be handled quite well on the assumption that f_1 is approximately unity, in which case Eqs. 70 reduce to

$$\begin{aligned} p_{00} &= cf_2, \\ p_{01} &= c(1 - f_2), \\ p_{10} &= 0, \\ p_{11} &= 1 - c. \end{aligned}$$

In the general case of Eqs. 70 some predictions can be made without knowing the exact parameter values. It has been noted in published studies (Estes, Hopkins, & Crothers, 1960; Estes, 1961b) that the observed proportion p_{01} is generally larger than p_{10} . Taking the difference between the theoretical expressions for these quantities, we have

$$\begin{aligned} p_{01} - p_{10} &= cf_1(1 - f_2) - c(1 - f_1)f_3 \\ &= c[J + (1 - J)a^{t_1}](1 - J)(1 - a^{t_2}) \\ &\quad - c(1 - J)(1 - a^{t_1})J(1 - a^{t_2}) \\ &= c(1 - J)(1 - a^{t_2})[J + (1 - J)a^{t_1} - J(1 - a^{t_1})] \\ &= c(1 - J)(1 - a^{t_2})a^{t_1}, \end{aligned}$$

which obviously must be equal to or greater than zero. The experiments cited above have in all cases had $t_1 < t_2$ and therefore $f_1 > f_2$. Since f_2 , which is directly estimated by the proportions of instances in which correct responses on T_1 are repeated on T_2 , has ranged from about 0.6 to 0.9 in these experiments (and f_1 must be larger), it is clear that p_{10} , the probability of an incorrect followed by a correct response, should be relatively small. This theoretical prediction accords well with observation.

Numerous predictions can be generated concerning the effects of varying the durations of t_1 and t_2 . The probability of repeating a correct response from T_1 to T_2 , for example, should depend solely on the parameter f_2 , decreasing as t_2 increases (and f_2 therefore decreases). The probability of a correct response on T_2 following an incorrect response on T_1 should depend most strongly on f_3 , increasing as t_2 (and therefore f_3) increases.

The over-all proportion correct per test should, of course, decrease from T_1 to T_2 (although the difference between proportions on T_1 and T_2 tends to zero as t_1 becomes large). Data relevant to these and other predictions are available in studies by Estes, Hopkins, and Crothers (1960), Peterson, Saltzman, Hillner, and Land (1962), and Witte (R. Witte, personal communication). The predictions concerning effects of variation of t_2 are well confirmed by these studies. Results bearing on predictions concerning variation in t_1 are not consistent over the set of experiments, possibly because of artifacts arising from item selection (discussed by Peterson et al., 1962).

APPLICATION TO THE SIMPLE NONCONTINGENT CASE. We restrict consideration to the special case of $N = 1$; thus we are dealing with a variant of the pattern model in which the pattern sampled on any trial is the one most likely to be sampled on the next trial. No new concepts are required beyond those introduced in connection with the *RTT* experiment, but it is convenient to denote by a single symbol, say g , the probability that the stimulus pattern sampled on any trial n is exchanged for another pattern on trial $n + 1$. In terms of this notation,

$$g = 1 - f_1 = (1 - J)(1 - a^t) = \left(1 - \frac{1}{N^*}\right)(1 - a^t),$$

where t is now taken to denote the intertrial interval. Also, we denote by $u_{1m,n}$ the probability of the state of the organism in which m stimulus patterns are conditioned to the A_1 -response and one of these is sampled and by $u_{0m,n}$ the probability that m patterns are conditioned to A_1 but a pattern conditioned to A_2 is sampled. Obviously

$$p_n = \sum_{m=0}^{N^*} u_{1m,n},$$

where, as usual, p_n denotes the probability of the A_1 -response on trial n .

Now we can write expressions for trigram probabilities, following essentially the same reasoning used before in the case of the pattern model with independent sampling. For the joint event $A_1E_1A_1$ we obtain

$$\begin{aligned} \Pr(A_{1,n+1}E_{1,n}A_{1,n}) &= \pi \sum_m u_{1m,n} \left[1 - g + g \frac{m-1}{N'}\right] \\ &= \pi \left[\left(1 - g - \frac{g}{N'}\right) p_n + g \sum_m u_{1m,n} \frac{m}{N'} \right], \end{aligned}$$

for if an element conditioned to A_1 is sampled on trial n then with probability $1 - g$ it is resampled and with probability $g[(m-1)/N']$

it is replaced by another element conditioned to A_1 ; in either event an A_1 -response must occur on trial $n + 1$. If the abbreviations $U_n = \sum_m u_{1m,n}(m/N')$ and $V_n = \sum_m u_{0m,n}(m/N')$ are used, the trigram probabilities can be written in relatively compact form:

$$\begin{aligned}
 \Pr(A_{1,n+1}E_{1,n}A_{1,n}) &= \pi \left[\left(1 - g - \frac{g}{N'} \right) p_n + gU_n \right], \\
 \Pr(A_{1,n+1}E_{2,n}A_{1,n}) &= (1 - \pi) \left[\left((1 - c)(1 - g) - \frac{g}{N'} \right) p_n + gU_n \right], \\
 \Pr(A_{1,n+1}E_{1,n}A_{2,n}) &= \pi [c(1 - g)(1 - p_n) + gV_n], \\
 \Pr(A_{1,n+1}E_{2,n}A_{2,n}) &= (1 - \pi)gV_n, \\
 \Pr(A_{2,n+1}E_{1,n}A_{1,n}) &= \pi g \left[\left(1 + \frac{1}{N'} \right) p_n - U_n \right], \\
 \Pr(A_{2,n+1}E_{2,n}A_{1,n}) &= (1 - \pi) \left[\left(c - cg + g + \frac{g}{N'} \right) p_n - gU_n \right], \\
 \Pr(A_{2,n+1}E_{1,n}A_{2,n}) &= \pi [(1 - c + cg)(1 - p_n) - gV_n], \\
 \Pr(A_{2,n+1}E_{2,n}A_{2,n}) &= (1 - \pi)[1 - p_n - gV_n].
 \end{aligned} \tag{71}$$

The chief difference between these expressions and the corresponding ones for the independent sampling models is that sequential effects now depend on the intertrial interval. Consider, for example, the first two of Eqs. 71, involving repetitions of response A_1 . It will be noted that both expressions represent linear combinations of p_n and U_n , with the relative contribution of p_n increasing as the intertrial interval (and therefore g) decreases. Also, it is apparent from the defining equations for p_n and U_n that $p_n \geq U_n$, with equality obtaining only in the special cases in which both are equal to unity or both equal to zero. Therefore, the probability of a repetition is inversely related to the intertrial interval. In particular, the probability that a correct A_1 - or A_2 -response will be repeated tends to unity in the limit as the intertrial interval goes to zero. When the intertrial interval becomes large, the parameter g approaches $1 - 1/N^*$ and Eqs. 71 reduce to those of a pattern model with N elements and independent sampling.

Summing the first four of Eqs. 71, we obtain a recursion for probability of the A_1 -response:

$$p_{n+1} = \left(1 - c - g - \frac{g}{N'} + cg \right) p_n + c(1 - g)\pi + g(U_n + V_n).$$

Now, although a full proof would be quite involved, it is not hard to

show heuristically that the asymptote is independent of the intertrial interval. We note first that asymptotically we have

$$\begin{aligned} U_n &= \sum_m u_{1m} \frac{m}{N'} \\ &= \sum_m u_m \frac{m}{N^*} \frac{m}{N'} \\ &= \frac{N^*}{N'} \sum_m \left(\frac{m}{N^*} \right)^2 u_m \\ &= \frac{N^*}{N'} \alpha_{2,n}, \end{aligned}$$

where u_m is the probability that m elements are conditioned to A_1 . The substitution of $u_m(m/N^*)$ for u_{1m} is possible in view of the intuitively evident fact that, asymptotically, the probability that an element conditioned to A_1 will constitute the trial sample is simply equal to the proportion of such elements in the total population. Substituting into the recursion for p_n in terms of this relation, and the analogous one for V_n ,

$$V_n = \frac{N^*}{N'} (p_n - \alpha_{2,n}),$$

we obtain

$$\begin{aligned} p_{n+1} &= \left(1 - c - g - \frac{g}{N'} + cg \right) p_n + c(1 - g)\pi + g \frac{N^*}{N'} p_n \\ &= (1 - c + cg)p_n + c(1 - g)\pi, \end{aligned}$$

the simplification in the last line having been effected by means of the identity

$$-g - \frac{g}{N'} = -g \left(\frac{N' + 1}{N'} \right) = -g \frac{N^*}{N'}.$$

Setting $p_{n+1} = p_n = p_\infty$ and solving for p_∞ , we arrive at the tidy outcome

$$p_\infty = (1 - c + cg)p_\infty + c(1 - g)\pi,$$

whence

$$p_\infty = \pi.$$

The recursion in p_n can be solved, but the resulting formula expressing p_n as a function of n and the parameters is too cumbersome to yield much useful information by visual inspection. It seems intuitively obvious that for $g < 1 - 1/N^*$ (i.e., for any but very long intertrial intervals) the learning curve will rise more sharply on early trials than the corresponding curve for the independent sampling case. This is so because only sampled elements can undergo conditioning, and, once sampled, an element is more likely to be resampled the shorter the intertrial interval. However,

the curves for longer and shorter intervals must cross ultimately, with the curve for the longer interval approaching asymptote more rapidly on later trials (Estes, 1955b). If $\pi = 1$, the total number of errors expected during learning must be independent of the intertrial interval because each initially unconditioned element will continue to produce an error each time it is sampled until it is finally conditioned, and the probability of any specified number of errors before conditioning depends only on the value of the conditioning parameter c . Similarly, if π is set equal to 0 after a conditioning session, the total number of conditioned responses during extinction is independent of the intertrial interval.

4.3 The Linear Model as a Limiting Case

For those experiments in which the available stimuli are the same on all trials the possibility arises of using a model that suppresses the concept of stimuli. In such a "pure" reinforcement model the learning assumptions specify directly how response probability changes on a reinforced trial. By all odds the most popular models of this sort are those which assume probability of a response on a given trial to be a linear function of the probability of that response on the previous trial.¹²

The so-called "linear models" received their first systematic treatment by Bush and Mosteller (1951a, 1955) and have been investigated and developed further by many others. We shall be concerned only with a certain class of linear models based on a single learning parameter θ . A more extensive analysis of this class of linear models has been given in Estes & Suppes (1959a).

The linear theory is formulated for the probability of a response on trial $n + 1$, given the entire preceding sequence of responses and reinforcements.¹³ Let x_n be the sequence of responses and reinforcements of a given subject through trial n ; that is, x_n is a sequence of length $2n$ with entries in the odd positions indicating responses and entries in the even positions indicating reinforcements. The axioms of the linear model are as follows.

Linear Axioms

For every i, i' and k such that $1 \leq i, i' \leq r$ and $0 \leq k \leq r$:

L1. If $\Pr(E_{i,n}A_{i',n}x_{n-1}) > 0$, then

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

¹² For a discussion of this general class of "incremental" models see Chapter 9 by Sternberg in this volume.

¹³ In the language of stochastic processes we have a chain of infinite order.

- L2. If $\Pr(E_{k,n}A_{i',n}x_{n-1}) > 0$, $k \neq i$ and $k \neq 0$, then

$$\Pr(A_{i,n+1} | E_{k,n}A_{i',n}x_{n-1}) = (1 - \theta) \Pr(A_{i,n} | x_{n-1}).$$
- L3. If $\Pr(E_{0,n}A_{i',n}x_{n-1}) > 0$, then

$$\Pr(A_{i,n+1} | E_{0,n}A_{i',n}x_{n-1}) = \Pr(A_{i,n} | x_{n-1}).$$

By Axiom L1, if the reinforcing event E_i , corresponding to response A_i , occurs on trial n , then (regardless of the response occurring on trial n) the probability of A_i increases by a linear transform of the old value. By L2, if some reinforcing event other than E_i occurs on trial n , then the probability of A_i decreases by a linear transform of its old value; and by L3 occurrence of the "neutral" event E_0 leaves response probabilities unchanged. The axioms may be written more compactly in terms of the probability $p_{xi,n}$ that a subject identified with sequence x makes an A_i response on trial n :

1. If the subject receives an E_i -event on trial n ,

$$p_{xi,n+1} = (1 - \theta)p_{xi,n} + \theta.$$

2. If the subject receives an E_k -event ($k \neq i$ and $k \neq 0$) on trial n ,

$$p_{xi,n+1} = (1 - \theta)p_{xi,n}.$$

3. If the subject receives an E_0 -event on trial n ,

$$p_{xi,n+1} = p_{xi,n}.$$

From a mathematical standpoint it is important to note that for the linear model the response probability associated with a particular subject is free to vary continuously over the entire interval from 0 to 1, since this probability undergoes linear transformations as a result of reinforcement. Consequently, if we wish to interpret changes in response probability as transitions among states of a Markov process, we must deal with a continuous-state space. Thus the Markov interpretation is of little practical value for calculational purposes. In stimulus sampling models response probability is defined in terms of the proportion of stimuli conditioned; since the set of stimuli is finite, so also is the set of values taken on by the response probability of any individual subject. It is this finite character of stimulus sampling models that makes possible the extremely useful interpretation of the models as finite Markov chains.

An inspection of the three axioms for the linear model indicates that they have the same general form as Eqs. 65, which describe changes in response probability for the fixed-sample-size component model; that is, if we let $\theta = cs/N$, then the two sets of rules are similar. As might be expected from this observation, many of the predictions generated by the

two models are identical when $\theta = cs/N$. For example, in the simple noncontingent situation the mean learning curve for the linear model is

$$\Pr(A_{1,n}) = \pi - [\pi - \Pr(A_{1,1})](1 - \theta)^{n-1}, \quad (72)$$

which is the same as that of the component model (see Estes & Suppes, 1959a, for a derivation of results for the linear model). However, the two models are not identical in all respects, as is indicated by a comparison of the asymptotic variances of the response distributions. For the linear model

$$\sigma_{\infty}^2 = \pi(1 - \pi) \frac{\theta}{2 - \theta},$$

as contrasted to Eq. 63 for the component model. However, as already noted in connection with Eq. 63, in the limit (as $N \rightarrow \infty$) the σ_{∞}^2 for the component model equals the predicted value for the linear model.

The last result suggests that the component model may converge to the linear process as $N \rightarrow \infty$. This conjecture is substantially correct; it can be shown that in the limit both the fixed-sample-size model and the independent sampling model approach the linear model for an extremely broad class of assumptions governing the sampling of elements. The derivation of the linear model from component models holds for any reinforcement schedule, for any finite number r of responses, and for every trial n , not simply at asymptote. The proof of this convergence theorem is lengthy and it is not presented here. However, the proof depends on the fact that the variance of the sampling distribution for any statistic of the trial sample approaches 0 as N becomes large. A proof of the convergence theorem is given by Estes and Suppes (1959b). Kemeny and Snell (1957) also have considered the problem but their proof is restricted to the two-choice noncontingent situation at asymptote.

COMPARISON OF THE LINEAR AND PATTERN MODELS. The same limiting result does not, of course, hold for the pattern model discussed in Sec. 2. For the pattern model only one element is sampled on each trial, and it is obvious that as $N \rightarrow \infty$ the learning effect of this sampling scheme would diminish to zero. For experimental situations in which both the linear model and the pattern model appear to be applicable it is important to derive differential predictions from the two models that, on empirical grounds, will permit the researcher to choose between them. To this end we display a few predictions for the linear model applied to both the *RTT* situation and the simple two-response noncontingent situation; these results will be compared with the corresponding equations for the pattern model.

For simplicity let us assume that in the case of the *RTT* situation the likelihood of a correct response by guessing is negligible on all trials. Then,

according to the linear model, the probability of a reinforced response changes in accordance with the equation

$$p_{n+1} = (1 - \theta)p_n + \theta.$$

In the present application the probability of a correct response on the first trial (the R trial) is zero, hence the probability of a correct response on the first test trial is simply θ . No reinforcement is given on T_1 , and consequently the probability of a correct response does not change between T_1 and T_2 . Therefore, p_{00} , the probability of a correct response on both T_1 and T_2 (as defined in connection with Eq. 55) is θ^2 . Similarly, we obtain $p_{01} = p_{10} = \theta(1 - \theta)$ and $p_{11} = (1 - \theta)^2$. Some relevant data are presented in Table 6 (from Estes, 1961b). They represent joint response

Table 6 Observed Joint Response Proportions for RTT Experiment and Predictions from Linear Retention-Loss Model and Sampling Model

| | Observed Proportion | Retention-Loss Model | Sampling Model |
|----------|------------------------|-------------------------|-------------------|
| p_{00} | 0.238 | 0.238 | 0.238 |
| p_{01} | 0.147 | 0.238 | 0.152 |
| p_{10} | 0.017 | 0.018 | 0 |
| p_{11} | 0.598 | 0.506 | 0.610 |

proportions for 40 subjects, each tested on 15 paired associate items of the type described in Sec. 1.1, the RTT design applied to each item. In order to minimize the probability of correct responses occurring by guessing, these items were introduced (one per trial) into a larger list, the composition of which changed from trial to trial. A critical item introduced on trial n received one reinforcement (paired presentation of stimulus and response members), followed by a test (presentation of stimulus alone) on trial n and trial $n + 1$, after which it was dropped from the list.

From an inspection of the data column of Table 6 it is obvious that the simple linear model cannot handle these proportions. It suffices to note that the model requires $p_{01} = p_{10}$, whereas the difference between these two entries in the data column is quite large.

One might try to preserve the linear model by arguing that the pattern of observed results in Table 6 could have arisen as an artifact. If, for example, there are differences in difficulty among items (or, equivalently, differences in learning rate among subjects), then the instances of incorrect response on T_1 would predominately represent smaller θ -values than instances of correct responses. On this account it might be expected that

the predicted proportion of correct following incorrect responses would be smaller than that allowed for under the "equal θ " assumption and therefore that the linear model might not actually be incompatible with the data of Table 6. We can easily check the validity of such an argument. Suppose that parameter θ_i is associated with a proportion f_i of the items (or subjects). Then in each case in which θ_i is applicable the probability of a correct response on T_1 followed by an error on T_2 is $\theta_i(1 - \theta_i)$. Clearly, then, p_{01} estimated from a group of items described by differences in θ would be

$$p_{01} = \sum_i f_i \theta_i (1 - \theta_i).$$

But a similar argument yields

$$p_{10} = \sum_i f_i (1 - \theta_i) \theta_i.$$

Since, again, the expressions for p_{10} and p_{01} are equal for all distributions of θ_i , it is clear that individual differences in learning rates alone could not account for the observed results.

A related hypothesis that might seem to merit consideration is that of individual differences in rates of forgetting. Since the proportion of correct responses on T_2 is less than that on T_1 , there is evidently some retention loss, and differences among subjects, or items, in susceptibility to this retention loss might be a source of bias in the data. The hypothesis can be formulated in the linear model as follows: the probability of the correct response on T_1 is equal to θ ; if, however, there is a retention loss, then the probability of a correct response on T_2 will have declined to some value ρ , such that $\rho < \theta$. If there are individual differences in amount of retention loss, then we should again categorize the population of subjects and items into subgroups, with a proportion f_i of the subjects characterized by retention parameter ρ_i . Theoretical expressions for p_{jk} can be derived for such a population by the same method used in the preceding case; the results are

$$p_{00} = \theta \sum_i f_i \rho_i,$$

$$p_{01} = \theta \sum_i f_i (1 - \rho_i),$$

$$p_{10} = (1 - \theta) \sum_i f_i \rho_i,$$

$$p_{11} = (1 - \theta) \sum_i f_i (1 - \rho_i).$$

This time the expressions for p_{10} and p_{01} are different; with a suitable choice of parameter values, they could accommodate the difference between the observed proportions p_{01} and p_{10} . However, another difficulty remains. To obtain a near-zero value for p_{10} would require either a θ near unity, which would be incompatible with the observed proportion of 0.385 correct on T_1 , or a value of $\sum_i f_i \rho_i$ near zero, which would be incompatible

with the observed proportion of 0.255 correct on T_2 . Thus we have no support for the hypothesis that individual differences in amount of retention loss might account for the pattern of empirical values.

We could go on in a similar fashion and examine the results of supplementing the original linear model by hypotheses involving more complex combinations or interactions of possible sources of bias (see Estes, 1961b). For example, we might assume that there are large individual differences in both learning and retention parameters. But, even with this latitude, it would not be easy to adjust the linear model to the RTT data. Suppose that we admit different learning parameters, θ_1 and θ_2 , and different retention parameters, ρ_1 and ρ_2 , the combination $\theta_1\rho_1$ obtaining for half the items and the combination $\theta_2\rho_2$ for the other half. Now the p_{ij} formulas become

$$p_{00} = \frac{\theta_1\rho_1 + \theta_2\rho_2}{2},$$

$$p_{01} = \frac{\theta_1(1 - \rho_1) + \theta_2(1 - \rho_2)}{2},$$

$$p_{10} = \frac{(1 - \theta_1)\rho_1 + (1 - \theta_2)\rho_2}{2},$$

$$p_{11} = \frac{(1 - \theta_1)(1 - \rho_1) + (1 - \theta_2)(1 - \rho_2)}{2}.$$

From the data column of Table 6 the proportions of correct responses on the first and second test trials are $p_{0-} = 0.385$ and $p_{-0} = 0.255$, respectively. Adding the first and second of the foregoing equations to obtain the theoretical expression for p_{0-} and the first and third equations to get p_{-0} , we have

$$p_{0-} = \frac{\theta_1 + \theta_2}{2}$$

and

$$p_{-0} = \frac{\rho_1 + \rho_2}{2}.$$

Equating theoretical and observed values, we obtain the constraints

$$\theta_1 + \theta_2 = 0.770$$

$$\rho_1 + \rho_2 = 0.510,$$

which should be satisfied by the parameter values. If the proportion p_{00} in Table 6 is to be predicted correctly, we must have

$$\frac{\theta_1\rho_1 + \theta_2\rho_2}{2} = 0.238,$$

or, substituting from the two preceding equations,

$$\theta_1 \rho_1 + (0.77 - \theta_1)(0.51 - \rho_1) = 0.476,$$

which may be solved for θ_1 :

$$\theta_1 = \frac{0.083 + 0.77\rho_1}{2\rho_1 - 0.51}.$$

Now the admissible range of parameter values can be further reduced. For the right-hand side of this last equation to have a value between 0 and 1, ρ_1 must be greater than 0.48; so we have the relatively narrow bounds on the parameters ρ_i

$$0.48 \leq \rho_1 \leq 0.51$$

$$0 \leq \rho_2 \leq 0.03.$$

Using these bounds on ρ_1 , we find from the equation expressing θ_1 as a function of ρ_1 that θ_1 must in turn satisfy $0.93 \leq \theta_1 \leq 1.0$. But now the model is in trouble, for, in order to satisfy the constraint $\theta_1 + \theta_2 = 0.77$, θ_2 would have to be negative (and the correct response probabilities for half of the items on T_1 would also be negative). About the best we can do, without allowing "negative probabilities," is to use the limits we have obtained for ρ_1 , ρ_2 , and θ_1 and arbitrarily assign a zero or small positive value to θ_2 . Choosing the combination $\theta_1 = 0.95$, $\theta_2 = 0.01$, $\rho_1 = 0.5$, and $\rho_2 = 0.01$, we obtain the theoretical values listed for the linear model in Table 6. By introducing additional assumptions or additional parameters, we could improve the fit of the linear model to these data, but there would seem to be little point in doing so. The refractoriness of the data to description by any reasonably simple form of the model suggests that perhaps the learning process is simply not well represented by the type of growth function embodied in the linear model.

By contrast, these data can be quite readily handled by the stimulus fluctuation model developed in the preceding section. Letting $f_1 = 1$ in Eqs. 70 and using the estimates $c = 0.39$ and $f_2 = 0.61$, we obtain the theoretical values listed under "Sampling Model" in Table 6. We would not, of course, claim that the sampling model had been rigorously tested, since two parameters had to be estimated and there are only three degrees of freedom in this set of data. However, the model does seem more promising than any of the variants of the linear model that have been investigated. More stringent tests of the sampling model can readily be obtained by running similar experiments with longer sequences of test trials, since predictions concerning joint response proportions over blocks of three or more test trials can be generated without additional assumptions.

ADDITIONAL COMPARISONS BETWEEN THE LINEAR AND PATTERN MODEL. We now turn to a few comparisons between the linear model and the multi-element pattern model for the simple noncontingent situation. First of all, we note that the mean learning curves for the two models (as given in Eq. 37 and Eq. 72) are identical if we let $c/N = \theta$. However, the expressions for the variance of the asymptotic response distribution are different; for the linear model $\sigma_\infty^2 = \pi(1 - \pi)[\theta/(2 - \theta)]$, whereas for the pattern model $\sigma_\infty^2 = \pi(1 - \pi)(1/N)$. This difference is reflected in another prediction that provides a more direct experimental test of the two models. It concerns the asymptotic variance of the distribution of the number of A_1 -responses in a block of K trials which we denote $\text{Var}(\bar{A}_K)$. For the linear model (cf. Estes & Suppes, 1959a),

$$\text{Var}(\bar{A}_K) = \pi(1 - \pi) \left\{ \frac{K(4 - 3\theta)}{2 - \theta} - \frac{2(1 - \theta)}{(2 - \theta)\theta} [1 - (1 - \theta)^K] \right\}.$$

For the pattern model, by Eq. 42,

$$\text{Var}(\bar{A}_K) = \pi(1 - \pi) \left\{ K + \frac{2K(1 - c)}{c} - \frac{2(1 - c)N}{c^2} \left[1 - \left(1 - \frac{c}{N} \right)^K \right] \right\}.$$

Note that, for $c = \theta$, the variance for the pattern model is larger than for the linear model. However, for the case of $\theta = c/N$ the variance for the pattern model can be larger or smaller than for the linear model depending on the particular values of c and N .

Finally, we present certain asymptotic sequential predictions for the linear model in the noncontingent situation; namely

$$\begin{aligned} \lim \Pr(A_{1,n+1} \mid E_{1,n}A_{1,n}) &= (1 - \theta)a + \theta \\ \lim \Pr(A_{1,n+1} \mid E_{2,n}A_{1,n}) &= (1 - \theta)a \\ \lim \Pr(A_{1,n+1} \mid E_{1,n}A_{2,n}) &= (1 - \theta)b + \theta \\ \lim \Pr(A_{1,n+1} \mid E_{2,n}A_{2,n}) &= (1 - \theta)b \end{aligned}$$

where

$$a = \pi + \frac{\theta(1 - \pi)}{2 - \theta} \quad \text{and} \quad b = \pi - \frac{\theta\pi}{2 - \theta}.$$

These predictions are to be compared with Eq. 34 for the pattern model. In the case of the pattern model we note that $\Pr(A_1 \mid E_1A_1)$ and $\Pr(A_1 \mid E_2A_2)$ depend only on π and N , whereas $\Pr(A_1 \mid E_2A_1)$ and $\Pr(A_1 \mid E_1A_2)$ depend on π , N , and c . In contrast, all four sequential probabilities depend on π and θ in the linear model. For comparisons between the linear model and the pattern model in application to two-choice data, the reader is referred to Suppes & Atkinson (1960).

4.4 Applications to Multiperson Interactions

In this section we apply the linear model to experimental situations involving multiperson interactions in which the reinforcement for any given subject depends both on his response and on the responses of other subjects. Several recent investigations have provided evidence indicating the fruitfulness of this line of development. For example, Bush and Mosteller (1955) have analyzed a study of imitative behavior in terms of their linear model, and Estes (1957a), Burke (1959, 1960), and Atkinson and Suppes (1958) have derived and tested predictions from linear models for behavior in two- and three-person games. Suppes and Atkinson (1960) have also provided a comparison between pattern models and linear models for multiperson experiments and have extended the analysis to situations involving communication between subjects, monetary payoff, social pressure, economic oligopolies, and related variables.

The simple two-person game has particular advantages for expository purposes, and we use this situation to illustrate the technique of extending the linear model to multiperson interactions. We consider a situation which, from the standpoint of game theory (see, e.g., Luce & Raiffa, 1957), may be characterized as a game in normal form with a finite number of strategies available to each player. Each play of the game constitutes a trial, and a player's choice of a strategy for a given trial corresponds to the selection of a response. To avoid problems having to do with the measurement of utility (or from the viewpoint of learning theory, problems of reward magnitude), we assume a unit reward that is assigned on an all-or-none basis. Rules of the game require the two players to exhibit their choices simultaneously on all trials (as in a game of matching pennies), and each player is informed that, given the choice of the other player on the trial, there is exactly one choice leading to the unit reward.

We designate the two players as A and B and let A_i ($i = 1, \dots, r$) and B_j ($j = 1, \dots, r'$) denote the responses available to the two players. The set of reinforcement probabilities prescribed by the experimenter may be represented in a matrix (a_{ij}, b_{ij}) analogous to the "payoff matrix" familiar in game theory. The number a_{ij} represents the probability of Player A being correct on any trial of the experiment, given the response pair $A_i B_j$; similarly, b_{ij} is the probability of Player B being correct, given the response pair $A_i B_j$. For example, consider the matrix

$$\begin{array}{c} \begin{array}{cc} & B_1 & B_2 \\ A_1 & \left[\frac{1}{2}, \frac{1}{2} \right] & 1, 0 \\ A_2 & 1, 0 & 0, 1 \end{array} \end{array}$$

When both subjects make Response 1, each has probability $\frac{1}{2}$ of receiving reward; when both make Response 2, then only Player *B* receives reward; when either of the other possible response pairs occurs (i.e., A_2B_1 or A_1B_2), then only Player *A* receives reward. It should be emphasized that, although one usually thinks of one player winning and the other losing on any given play of a game, this is not a necessary restriction on the model. In theory, and in experimental tests of the theory, it is quite possible to permit both or neither of the players to be rewarded on any trial. However, to provide a relatively simple theoretical interpretation of reinforcing events, it is essential that on a nonrewarded trial the player be informed (or led to infer) that some other choice, had he made it under the same circumstances, would have been successful. We return to this point later.

Let $E_i^{(A)}$ denote the event of reinforcing the A_i response for Player *A* and $E_j^{(B)}$ the event of reinforcing the B_j response for Player *B*. To simplify our analysis, we consider the case in which each subject has only two response alternatives, and we define the probability of occurrence of a particular reinforcing event in terms of the payoff parameters as follows (for $i \neq i'$ and $j \neq j'$):

$$\begin{aligned}
 a_{ij} &= \Pr(E_i^{(A)} \mid A_{i,n}B_{j,n}) & b_{ij} &= \Pr(E_j^{(B)} \mid A_{i,n}B_{j,n}) \\
 1 - a_{ij} &= \Pr(E_{i'}^{(A)} \mid A_{i,n}B_{j,n}) & 1 - b_{ij} &= \Pr(E_{j'}^{(B)} \mid A_{i,n}B_{j,n}).
 \end{aligned}
 \tag{73}$$

For example, if Player *A* makes an A_1 -response and is rewarded, then an $E_1^{(A)}$ occurs; however, if an A_1 is made and no reward occurs, then we assume that the other response is reinforced, that is, an $E_2^{(A)}$ occurs.

Finally, one last definition to simplify notation. We denote Player *A*'s response probability by α and Player *B*'s by β , and we denote by γ the joint probability of an A_1 - and B_1 -response. Specifically,

$$\alpha_n = \Pr(A_{1,n}), \quad \beta_n = \Pr(B_{1,n}), \quad \gamma_n = \Pr(A_{1,n}B_{1,n}). \tag{74}$$

We now derive a theorem that provides recursive expressions for α_n and β_n and points up a property of the model that greatly complicates the mathematics, namely, that both α_{n+1} and β_{n+1} depend on the joint probability $\gamma_n = \Pr(A_{1,n}B_{1,n})$. The statement of the theorem is as follows:

$$\begin{aligned}
 \alpha_{n+1} &= [1 - \theta_A(2 - a_{12} - a_{22})]\alpha_n + \theta_A(a_{22} - a_{21})\beta_n \\
 &\quad + \theta_A(a_{11} + a_{21} - a_{12} - a_{22})\gamma_n + \theta_A(1 - a_{22})
 \end{aligned}
 \tag{75a}$$

$$\begin{aligned}
 \beta_{n+1} &= [1 - \theta_B(2 - b_{21} - b_{22})]\beta_n + \theta_B(b_{22} - b_{12})\alpha_n \\
 &\quad + \theta_B(b_{11} + b_{12} - b_{21} - b_{22})\gamma_n + \theta_B(1 - b_{22}),
 \end{aligned}
 \tag{75b}$$

where θ_A and θ_B are the learning parameters for players *A* and *B*. In the proof of this theorem it will suffice to derive the difference equation for α_{n+1} , since the derivation for β_{n+1} is identical. To begin with, from

Axioms L1 and L2 we can easily show that the general form of a recursion for α_n is

$$\alpha_{n+1} = (1 - \theta_A)\alpha_n + \theta_A \Pr(E_{1,n}^{(A)}).$$

The term $\Pr(E_{1,n}^{(A)})$ can then be expanded to

$$\begin{aligned} \Pr(E_{1,n}^{(A)}) &= \sum_{i,j} \Pr(E_{1,n}^{(A)} A_{i,n} B_{j,n}) \\ &= \sum_{i,j} \Pr(E_{1,n}^{(A)} | A_{i,n} B_{j,n}) \Pr(A_{i,n} B_{j,n}) \end{aligned}$$

and by Eqs. 73

$$\begin{aligned} \Pr(E_{1,n}^{(A)}) &= a_{11} \Pr(A_{1,n} B_{1,n}) + a_{12} \Pr(A_{1,n} B_{2,n}) \\ &\quad + (1 - a_{21}) \Pr(A_{2,n} B_{1,n}) + (1 - a_{22}) \Pr(A_{2,n} B_{2,n}). \end{aligned} \quad (76)$$

Next we observe that

$$\begin{aligned} \Pr(A_{1,n} B_{2,n}) &= \Pr(B_{2,n} | A_{1,n}) \Pr(A_{1,n}) \\ &= [1 - \Pr(B_{1,n} | A_{1,n})] \Pr(A_{1,n}) \\ &= \Pr(A_{1,n}) - \Pr(A_{1,n} B_{1,n}). \end{aligned} \quad (77a)$$

Similarly,

$$\Pr(A_{2,n} B_{1,n}) = \Pr(B_{1,n}) - \Pr(A_{1,n} B_{1,n}), \quad (77b)$$

and

$$\begin{aligned} \Pr(A_{2,n} B_{2,n}) &= \Pr(A_{2,n} | B_{2,n}) \Pr(B_{2,n}) \\ &= [1 - \Pr(A_{1,n} | B_{2,n})] \Pr(B_{2,n}) \\ &= \Pr(B_{2,n}) - \Pr(A_{1,n} B_{2,n}) \\ &= 1 - \Pr(B_{1,n}) - \Pr(A_{1,n}) + \Pr(A_{1,n} B_{1,n}). \end{aligned} \quad (77c)$$

Substituting into Eq. 76 from Eqs. 77a, 77b, and 77c and simplifying by means of the definitions of α , β , and γ , we obtain

$$\begin{aligned} \Pr(E_{1,n}^{(A)}) &= a_{11}\gamma_n + a_{12}(\alpha_n - \gamma_n) + (1 - a_{21})(\beta_n - \gamma_n) \\ &\quad + (1 - a_{22})(1 - \alpha_n - \beta_n + \gamma_n) \\ &= -(1 - a_{12} - a_{22})\alpha_n + (a_{22} - a_{21})\beta_n \\ &\quad + (a_{11} + a_{21} - a_{12} - a_{22})\gamma_n + (1 - a_{22}). \end{aligned}$$

Substitution of this expression into the general recursion for α_n yields the desired result, which completes the proof.

It has been shown by Lamperti and Suppes (1959) that the limits α , β , and γ exist, whence (letting $\alpha_{n+1} = \alpha_n = \alpha$, $\beta_{n+1} = \beta_n = \beta$ and $\gamma_n = \gamma$ in Eqs. 75a and 75b) we have two linear relations that are independent of θ_A and θ_B , namely,

$$a\alpha = b\beta + c\gamma + d, \quad e\beta = f\alpha + g\gamma + h, \quad (78)$$

where

$$\begin{aligned}
 a &= 2 - a_{12} - a_{22} & b &= a_{22} - a_{21} \\
 c &= a_{11} + a_{21} - a_{12} - a_{22} & d &= 1 - a_{22} \\
 e &= 2 - b_{21} - b_{22} & f &= b_{22} - b_{12} \\
 g &= b_{11} + b_{12} - b_{21} - b_{22} & h &= 1 - b_{22}.
 \end{aligned}
 \tag{79}$$

By eliminating γ from Eqs. 78 we obtain the following linear relation in α and β :

$$(-ag - ce)\alpha + (bg + cf)\beta = ch - dg. \tag{80}$$

Unfortunately, this relationship is one of the few quantitative results that can be directly computed for the linear model. It has, however, the advantageous feature that it is independent of the learning parameters θ_A and θ_B and therefore may be compared directly with experimental data. Application of this result can be illustrated in terms of the game cited earlier in which the payoff matrix takes the form

$$\begin{array}{cc}
 & \begin{array}{c} B_1 \quad B_2 \end{array} \\
 \begin{array}{c} A_1 \\ A_2 \end{array} & \begin{bmatrix} \frac{1}{2}, \frac{1}{2} & 1, 0 \\ 1, 0 & 0, 1 \end{bmatrix}
 \end{array}$$

From Eqs. 79 we obtain

$$\begin{array}{cccc}
 a = 1 & b = -1 & c = \frac{1}{2} & d = 1 \\
 e = 1 & f = 1 & g = -\frac{1}{2} & h = 0
 \end{array}$$

and Eq. 80 becomes

$$\left(\frac{1}{2} - \frac{1}{2}\right)\alpha + \left(\frac{1}{2} + \frac{1}{2}\right)\beta = \frac{1}{2}$$

or $\beta = \frac{1}{2}$. From this result we predict immediately that the long-run proportion of B_1 -responses will tend to $\frac{1}{2}$. To derive a prediction for Player A , we substitute the known values of the parameters into the first part of Eq. 78 to obtain

$$\begin{aligned}
 \alpha &= -\beta + \frac{1}{2}\gamma + 1 \\
 &= \frac{1}{2} + \frac{1}{2}\gamma.
 \end{aligned}$$

Unfortunately we cannot compute γ , the asymptotic probability of the A_1B_1 -response pair. However, we know γ is positive, and, since only one half of Player B 's responses are B_1 's, γ cannot be greater than $\frac{1}{2}$. Therefore we have $0 \leq \gamma \leq \frac{1}{2}$ and as a result can set definite bounds on the long-run probability of an A_1 -response, namely,

$$\frac{1}{2} \leq \alpha \leq \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}.$$

Thus we have the basis for a rather exacting experimental test, since the asymptotic predictions for both subjects are parameter-free; that is, they do not depend on the θ -values of either subject or on the initial response probabilities.

Of course, by imposing restrictions on the experimentally determined parameters a_{ij} and b_{ij} a variety of results can be obtained. We limit ourselves to the consideration of one such case: choice of the parameters so that the coefficients of γ_n will vanish in the recursive equations (75a) and (75b). Specifically, if we let $c = g = 0$ and $af - be \neq 0$, then

$$\begin{aligned}\alpha_{n+1} &= a\alpha_n + b\beta_n + d \\ \beta_{n+1} &= e\beta_n + f\alpha_n + h.\end{aligned}\tag{81}$$

Solutions for this system are well known and can be obtained by a number of different techniques; for a detailed discussion of the problem of obtaining explicit expressions of α_n and β_n for arbitrary n the reader is referred to an article by Burke (1960). We do know, however, that the limits for α_n and β_n exist and are independent of both the initial conditions and θ_A and θ_B . By substituting $\alpha = \alpha_{n+1} = \alpha_n$ and $\beta = \beta_{n+1} = \beta_n$ into the two recursions we obtain

$$\alpha = \frac{bh + df}{af - be}$$

and

$$\beta = \frac{ah + de}{af - be}.$$

The fact that α and β are independent of θ_A and θ_B under the restrictions imposed on the parameters in no way implies that γ is also independent of these quantities.

Equations 81 provide a precise test of the model, and the necessary conditions for this test involve only experimentally manipulable parameters. A great deal of experimental work has been conducted on this restricted problem, and, in general, the correspondence between predicted and observed values has been good; for accounts of this work see Atkinson & Suppes (1958), Burke (1959, 1960), and Suppes & Atkinson (1960).

In conclusion we should mention that all of the predictions presented in this section are identical to those that can be derived from the pattern model of Sec. 2. However, in general, only the grosser predictions, such as those for α_n and β_n , are the same for the two models.

5. DISCRIMINATION LEARNING¹⁴

The distinction between simple learning and discrimination learning is somewhat arbitrary. By discrimination we refer, roughly speaking, to the

¹⁴ Using the terminology proposed by Bush, Galanter, and Luce in Chapter 2, the class of problems considered in this section would be called "identification-learning" experiments.

process whereby the subject learns to make one response to one of a pair of stimuli and a different response to the other. But there is an element of discrimination in any learning situation. Even in the simplest conditioning experiment the subject learns to make a conditioned response only when the conditioned stimulus is presented, and therefore to do something else when that stimulus is absent. In the paired-associate situation (referred to several times in preceding sections) the subject learns to associate the appropriate member of a response set with each member of a set of stimuli and therefore to "discriminate" the stimuli. The principal basis for differentiation between the two categories of learning seems to be that in the case of discrimination learning the similarity, or communality, between stimuli is a major independent variable; in the case of simple learning stimulus similarity is an extraneous factor to be minimized experimentally and neglected in theory as far as possible.

One of the general strategic assumptions of the type of stimulus-response theory, which has been associated with the development of stimulus sampling models, is that discrimination learning involves a combination of processes, each of which can be studied independently in simpler situations—the learning aspect in experiments on acquisition or extinction and the stimulus relationships in experiments on stimulus generalization or transfer of training. Thus there will be nothing new at the conceptual level in our treatment of discrimination. There is adequate scope for analysis of different types of discriminative situations, but, since our main concern in this section is with methods rather than content, we shall not go far in this direction. We propose only to show how the processes of association and generalization treated in preceding sections enter into discrimination learning, and this can be accomplished by formulating assumptions and deriving results of general interest for a few important cases.

5.1 The Pattern Model for Discrimination Learning

As in the cases of simple acquisition and probability learning, it is sometimes useful in the treatment of discriminative situations to ignore generalization effects among the stimuli involved in an experiment and to regard each stimulus display as a unique pattern. Thus behavior elicited by the stimulus display will depend only on the subject's reinforcement history with respect to that particular pattern. Two important variants of the model arise, depending on whether experimental arrangements do or do not ensure that the subject will sample the entire stimulus display presented on each trial.

Case 1. All cues presented are sampled on each trial. For a classical two-stimulus, two-response discrimination problem (e.g., a Lashley situation in which the rat is differentially rewarded for jumping to a black card and avoiding a grey card) our conceptualization requires a distinction among three types of cues: we denote by S_1 the set of component cues present only in the stimulus situation associated with reinforcement of response A_1 , by S_2 the set of cues present only in the situation associated with reinforcement of response A_2 , and by S_c the set of cues present in both situations. In the example of the Lashley situation A_1 might be the response of jumping to the left-hand window; A_2 , the response of jumping to the right-hand window; S_1 , the stimulation present only on trials with black cards; S_2 , the stimulation present only on trials with grey cards; and S_c , the stimulation common to both types of trials. We denote by N_1 , N_2 , and N_c the number of cues in each of these subsets. In standard experiments the "cues" refer to experimentally manipulable aspects of the situation, such as tones, objects, colors, or symbols, and it is reasonably well known just how many different combinations of these cues will be responded to by subjects as distinct patterns. In some instances, however, the experimenter may have no a priori knowledge of the patterns distinguishable by the subject; in such instances the N_i may be treated as unknown parameters to be estimated from data, and the model may thus serve as a tool in securing evidence concerning the subject's perceptions of the physical situation.

Suppose, now, that the experimenter's procedure is to present on some trials (T_1 -trials) a set of cues including m_1 from S_1 and m_c from S_c and on the remaining trials (T_2 -trials) m_2 cues from S_2 and m_c from S_c . Further, let the two types of trials occur with equal frequencies in random sequence.

On trials of type T_1 there will be $\binom{N_1}{m_1} \binom{N_c}{m_c}$ different patterns of cues available. Assuming that these patterns are all equally probable and letting $b_{1c} = \left[\binom{N_1}{m_1} \binom{N_c}{m_c} \right]^{-1}$, we can obtain an expression for probability of a correct response on a T_1 -trial simply by appropriate substitution into Eq. 28, namely,

$$\Pr(A_{1,n_1} | T_{1,n_1}) = 1 - [1 - \Pr(A_{1,1} | T_{1,1})](1 - cb_{1c})^{n_1-1}, \quad (82)$$

where n_1 is the ordinal number of the T_1 -trial. The corresponding function for T_2 -trials is obtained similarly with parameter $b_{2c} = \left[\binom{N_2}{m_2} \binom{N_c}{m_c} \right]^{-1}$.

In the discrimination literature cues in the sets S_1 and S_2 are commonly referred to as *relevant* and those in S_c as *irrelevant*, since S_1 and S_2 are associated with reinforcing events, whereas the S_c are not. It is apparent

by inspection of Eq. 82 that (for the foregoing specified experimental conditions) the pattern model predicts that probability of correct responding will go asymptotically to unity regardless of the numbers of relevant and irrelevant cues, provided only that neither m_1 nor m_2 is equal to zero. Rate of approach to asymptote on each type of trial is inversely related to the total number of patterns available for sampling. Therefore, other things being equal, rate of learning is decreased (and total errors to criterion increased) by the addition of either relevant or irrelevant cues.

Case 2. Only a subset of the cues presented on each trial is sampled. We consider now the situation that arises if the number of cues presented per trial is too large, or the exposure time too short, for the entire stimulus display to be sampled by the subject. Let us suppose that there are only two stimulus displays. The display on T_1 -trials comprises the N_1 cues of S_1 together with the N_c cues of S_c , and that on T_2 -trials, the N_2 cues of S_2 together with the N_c cues of S_c ; further, to simplify the analysis let $N_1 = N_2 = N$. For a given fixed exposure time we assume a fixed sample size s , with all samples of exactly s cues being equiprobable. On T_1 -trials, then, there will be $\binom{N}{s_1} \binom{N_c}{s - s_1}$ ways of filling the sample with s_1 cues from S_1 and the remainder from S_c . The asymptote of discriminative performance will depend on the size of s in relation to N_c . If $s \leq N_c$, so that the entire sample can come from the set of irrelevant cues, then the asymptotic probability of a correct response will be less than unity.

In Case 2 two types of patterns need to be distinguished for each type of trial. We can limit consideration to T_1 -trials, since analogous arguments hold for T_2 . There may be some patterns that include only cues from S_c and learning with respect to them will be on a simple random reinforcement schedule. The proportion of such patterns, w_c , is given by

$$w_c = \frac{\binom{N_c}{s}}{\binom{N + N_c}{s}},$$

which is equal to zero if $s > N_c$. If T_1 - and T_2 -trials have equal probabilities, then the probability, to be denoted V_n , that a pattern containing only cues from S_c will be conditioned to the A_1 -response on trial n can be obtained from Eq. 28 by setting $\pi_{12} = \pi_{21} = \frac{1}{2}$:

$$\Pr(A_{1,n}) = V_n \quad \text{and} \quad \frac{c}{N} = \frac{cw_c}{\binom{N_c}{s}} = cb,$$

where

$$b = \binom{N + N_c}{s}^{-1},$$

that is,

$$V_n = \frac{1}{2} - (\frac{1}{2} - V_1)(1 - cb)^{n-1}. \quad (83)$$

The remaining patterns available on T_1 -trials all contain at least one cue from S_1 and thus occur only on trials when response A_1 is reinforced. The probability, to be denoted U_n , that any one of these is conditioned to A_1 on trial n may be similarly obtained by rewriting Eq. 28, this time with $\pi_{12} = 0$, $\pi_{21} = 1$, $\Pr(A_{1,n}) = U_n$, and $c/N = \frac{1}{2}cb$, that is,

$$U_n = 1 - (1 - U_1)(1 - \frac{1}{2}cb)^{n-1}, \quad (84)$$

where the factor $\frac{1}{2}$ enters because these patterns are available for sampling on only one half of the trials.

Now, to obtain the probability of an A_1 -response if a T_1 -display is presented on trial n , we need only combine Eqs. 83 and 84, weighting each by the probability of the appropriate type of pattern, namely,

$$\begin{aligned} \Pr(A_{1,n} | T_{1,n}) &= (1 - w_c)U_n + w_cV_n \\ &= 1 - w_c + \frac{1}{2}w_c - (1 - w_c)(1 - U_1)(1 - \frac{1}{2}cb)^{n-1} \\ &\quad - w_c(\frac{1}{2} - V_1)(1 - cb)^{n-1}, \end{aligned} \quad (85a)$$

which may be simplified, if $U_1 = V_1 = \frac{1}{2}$, to

$$\Pr(A_{1,n} | T_{1,n}) = 1 - \frac{1}{2}w_c - \frac{1}{2}(1 - w_c)(1 - \frac{1}{2}cb)^{n-1}. \quad (85b)$$

The resulting expression for probability of a correct response has a number of interesting general properties. The asymptote, as anticipated, depends in a simple way on w_c , the proportion of "irrelevant patterns." When $w_c = 0$, the asymptotic probability of a correct response is unity; when $w_c = 1$, the whole process reduces to simple random reinforcement. Between these extremes, asymptotic performance varies inversely with w_c , so that the terminal proportion of correct responses on either type of trial provides a simple estimate of this parameter from data. The slope parameter cb could then be estimated from total errors over a series of trials. As in Case 1, the rate of approach to asymptote proves to depend only on the conditioning parameters and total number of patterns available for sampling; thus it is a joint function of the total number of cues $N + N_c$ and the sample size s but does not depend on the relative proportions of relevant and irrelevant cues. The last result may seem implausible, but it should be noted that the result depends on the simplifying assumption of the pattern model that there are no transfer effects from

learning on one pattern to performance on another pattern that has component cues in common with the first. The situation in this regard is different for the "mixed model" to be discussed next.

5.2 A Mixed Model

The pattern model may provide a relatively complete account of discrimination data in situations involving only distinct, readily discriminable patterns of stimulation, as, for example the "paired-comparison" experiment discussed in Sec. 2.3 or the verbal discrimination experiment treated by Bower (1962). Also, this model may account for some aspects of the data (e.g., asymptotic performance level, trials to criterion) even in discrimination experiments in which similarity, or communality, among stimuli is a major variable. But, to account for other aspects of the data in cases of the latter type, it is necessary to deal with transfer effects throughout the course of learning. The approach to this problem which we now wish to consider employs no new conceptual apparatus but simply a combination of ideas developed in preceding sections.

In the *mixed model* the conceptualization of the discriminative situation and the learning assumptions is exactly the same as that of the pattern model discussed in Sec. 5.1. The only change is in the response rule and that is altered in only one respect. As before, we assume that once a stimulus pattern has become conditioned to a response it will evoke that response on each subsequent occurrence (unless on some later trial the pattern becomes reconditioned to a different response, as, for example, during reversal of a discrimination). The new feature concerns patterns which have not yet become conditioned to any of the response alternatives of the given experimental situation but which have component cues in common with other patterns that have been so conditioned. Our assumption is simply that transfer occurs from a conditioned to an unconditioned pattern in accordance with the assumptions utilized in our earlier treatment of compounding and generalization (specifically, by axiom C2, together with a modified version of C1, of Sec. 3.1).

Before the assumptions about transfer can be employed unambiguously in connection with the mixed model, the notion of conditioned status of a component cue needs to be clarified. We shall say that a cue is conditioned to response A_i if it is a component of a stimulus pattern that has become conditioned to response A_i . If a cue belongs to two patterns, one of which is conditioned to response A_i and one to response A_j ($i \neq j$), then the conditioning status of the cue follows that of the more recently conditioned pattern. If a cue belongs to no conditioned pattern, then it is

said to be in the unconditioned, or "guessing," state. Note that a pattern may be unconditioned even though all of its cues are conditioned. Suppose for example, that a pattern consisting of cues x , y , and z in a particular arrangement has never been presented during the first n trials of an experiment but that each of the cues has appeared in other patterns, say wxy and wrz , which have been presented and conditioned. Then all of the cues of pattern xyz would be conditioned, but the pattern would still be in the unconditioned state. Consequently, if wxy had been conditioned to response A_1 and wrz to A_2 , the probability of A_1 in the presence of pattern xyz would be $\frac{2}{3}$; but, if response A_1 were effectively reinforced in the presence of xyz , its probability of evocation by that pattern would henceforth be unity.

The only new complication arises if an unconditioned pattern includes cues that are still in the unconditioned state. Several alternative ways of formulating the response rule for this case have some plausibility, and it is by no means sure that any one choice will prove to hold for all types of situations. We shall limit consideration to the formulation suggested by a recent study of discrimination and transfer which has been analyzed in terms of the mixed model (Estes & Hopkins, 1961). The amended response rule is a direct generalization of Axiom C2 of Sec. 3.1; specifically, for a situation involving r response alternatives the following assumptions will apply:

1. If all cues in a pattern are unconditioned, the probability of any response A_i is equal to $1/r$.

2. If a pattern (sample) comprises m cues conditioned to response A_i , m' cues conditioned to other responses, and m'' unconditioned cues, then the probability that A_i will be evoked by this pattern is given by

$$\Pr(A_i) = \frac{m + (m''/r)}{m + m' + m''}.$$

In other words, Axiom C2 holds but with each unconditioned cue contributing "weight" $1/r$ toward the evocation of each of the alternative responses.

To illustrate these assumptions in operation, let us consider a simple classical discrimination experiment involving three cues, a , b , and c , and two responses, A_1 and A_2 . We shall assume that the pattern ac is presented on half of the trials, with A_1 reinforced, and bc on the other half of the trials, with A_2 reinforced, the two types of trials occurring in random sequence. We assume further that conditions are such as to ensure the subject's sampling both cues presented on each trial. In a tabulation of the possible conditioning states of each pattern a 1, 2, or 0, respectively, in a state column indicates that the pattern is conditioned to A_1 , conditioned to A_2 , or unconditioned. For each pair of values under States, the associated

A_1 -probabilities, computed according to the modified response rule, are given in the corresponding positions under A_1 -probability. To reduce algebraic complications, we shall carry out derivations for the special case in which the subject starts the experiment with both patterns unconditioned. Then, under the conditions of reinforcement specified, only

| States | | A_1 -Probability to Each Pattern | |
|--------|---|---------------------------------------|---------------|
| | | ac | bc |
| 1 | 2 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 2 | 2 | 0 | 0 |
| 2 | 1 | 0 | 1 |
| 0 | 1 | $\frac{3}{4}$ | 1 |
| 0 | 2 | $\frac{1}{4}$ | 0 |
| 1 | 0 | 1 | $\frac{3}{4}$ |
| 2 | 0 | 0 | $\frac{1}{4}$ |
| 0 | 0 | $\frac{1}{2}$ | $\frac{1}{2}$ |

the states represented in the first, seventh, sixth, and ninth rows of the table are available to the subject, and for brevity we number these states 3, 2, 1, and 0, in the order just listed; that is,

State 3 = pattern ac conditioned to A_1 , and pattern bc conditioned to A_2 .

State 2 = pattern ac conditioned to A_1 , and pattern bc unconditioned.

State 1 = pattern ac unconditioned, and pattern bc conditioned to A_2 .

State 0 = both patterns ac and bc are unconditioned.

Now, these states can be interpreted as the states of a Markov chain, since the probability of transition from any one of them to any other on a given trial is independent of the preceding history. The matrix of probabilities for one-step transitions among the four states takes the following form:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{c}{2} & 1 - \frac{c}{2} & 0 & 0 \\ \frac{c}{2} & 0 & 1 - \frac{c}{2} & 0 \\ 0 & \frac{c}{2} & \frac{c}{2} & 1 - c \end{bmatrix}, \quad (86)$$

where the states are ordered 3, 2, 1, 0 from top to bottom and left to right. Thus State 3 (in which ac is conditioned to A_1 and bc to A_2) is an absorbing state, and the process must terminate in this state, with asymptotic probability of a correct response to each pattern equal to unity. In State 2 pattern ac is conditioned to A_1 , but bc is still unconditioned. This state can be reached only from State 0, in which both patterns are unconditioned; the probability of the transition is $\frac{1}{2}$ (the probability that pattern ac will be presented) times c (the probability that the reinforcing event will produce conditioning); thus the entry in the second cell of the bottom row is $c/2$. From State 2 the subject can go only to State 3, and this transition again has probability $c/2$. The other cells are filled in similarly.

Now the probability $u_{i,n}$ of being in state i on trial n can be derived quite easily for each state. The subject is assumed to start the experiment in State 0 and has probability c of leaving this state on each trial; hence

$$u_{0,n} = (1 - c)^{n-1}.$$

For State 1 we can write a recursion,

$$u_{1,n} = \left(1 - \frac{c}{2}\right)^{n-2} \frac{c}{2} + \left(1 - \frac{c}{2}\right)^{n-3} (1 - c) \frac{c}{2} + \dots + (1 - c)^{n-2} \frac{c}{2},$$

which holds if $n \geq 2$. To be in State 1 on trial n the subject must have entered at the end of trial 1, which has probability $c/2$, and then remained for $n - 2$ trials, which has probability $[(1 - (c/2)]^{n-2}$; have entered at the end of trial 2, which has probability $(1 - c)(c/2)$, and then remained for $n - 3$ trials, which has probability $[1 - (c/2)]^{n-3}$; ...; or have entered at the end of trial $n - 1$, which has probability $(1 - c)^{n-2}(c/2)$. The right-hand side of this recursion can be summed to yield

$$\begin{aligned} u_{1,n} &= \frac{c}{2} (1 - c)^{n-2} \sum_{v=0}^{n-2} \left[\frac{1 - (c/2)}{1 - c} \right]^v \\ &= (1 - c)^{n-1} \left\{ \left[\frac{2 - c}{2(1 - c)} \right]^{n-1} - 1 \right\} \\ &= \left(1 - \frac{c}{2}\right)^{n-1} - (1 - c)^{n-1}. \end{aligned}$$

By an identical argument we obtain

$$u_{2,n} = \left(1 - \frac{c}{2}\right)^{n-1} - (1 - c)^{n-1},$$

and then by subtraction

$$\begin{aligned} u_{3,n} &= 1 - u_{2,n} - u_{1,n} - u_{0,n} \\ &= 1 - 2 \left(1 - \frac{c}{2}\right)^{n-1} + (1 - c)^{n-1}. \end{aligned}$$

From the tabulation of states and response probabilities we know that the probability of response A_1 to pattern ac is equal to 1, 1 , $\frac{1}{4}$, and $\frac{1}{2}$, respectively, when the subject is in State 3, 2, 1, or 0. Consequently the probability of a correct (A_1) response to ac is obtained simply by summing these response probabilities, each weighted by the state probability, namely,

$$\begin{aligned} \Pr(A_{1,n} | ac) &= u_{3,n} + u_{2,n} + \frac{1}{4}u_{1,n} + \frac{1}{2}u_{0,n} \\ &= 1 - 2\left(1 - \frac{c}{2}\right)^{n-1} + (1 - c)^{n-1} + \left(1 - \frac{c}{2}\right)^{n-1} \\ &\quad - (1 - c)^{n-1} + \frac{1}{4}\left(1 - \frac{c}{2}\right)^{n-1} - \frac{1}{4}(1 - c)^{n-1} \\ &\quad + \frac{1}{2}(1 - c)^{n-1} \\ &= 1 - \frac{3}{4}\left(1 - \frac{c}{2}\right)^{n-1} + \frac{1}{4}(1 - c)^{n-1}. \end{aligned} \quad (87)$$

Equation 87 is written for the probability of an A_1 -response to ac on trial n ; however, the expression for probability of an A_2 -response to bc is identical, and consequently Eq. 87 expresses also the probability p_n of a correct response on any trial, without regard to the stimulus pattern presented. A simple estimator of the conditioning parameter c is now obtainable by summing the error probability over trials. Letting e denote the expected total errors during learning, we have

$$\begin{aligned} e &= \sum_{n=1}^{\infty} (1 - p_n) \\ &= \frac{3}{4} \sum_{n=1}^{\infty} \left(1 - \frac{c}{2}\right)^{n-1} - \frac{1}{4} \sum_{n=1}^{\infty} (1 - c)^{n-1} \\ &= \frac{3}{4} \frac{2}{c} - \frac{1}{4} \frac{1}{c} \\ &= \frac{5}{4c}. \end{aligned}$$

An example of the sort of prediction involving a relatively direct assessment of transfer effects is the following. Suppose the first stimulus pattern to appear is ac ; the probability of a correct response to it is, by hypothesis, $\frac{1}{2}$, and if there were no transfer between patterns the probability of a correct response to bc when it first appeared on a later trial should be $\frac{1}{2}$ also. Under the assumptions of the mixed model, however, the probability of a

correct response to bc , if it first appeared on trial 2, should be

$$\frac{[1 - \frac{1}{2}(1 - c) - c] + \frac{1}{2}}{2} = \frac{1}{2} - \frac{c}{4};$$

if it first appeared on trial 3, it should be

$$\frac{\frac{1}{2}(1 - c)^2 + \frac{1}{2}}{2} = \frac{1}{2} - \frac{c}{2}\left(1 - \frac{c}{2}\right);$$

and so on, tending to $\frac{1}{4}$ after a sufficiently long prior sequence of ac trials.

Simply by inspection of the transition matrix we can develop an interesting prediction concerning behavior during the presolution period of the experiment. By presolution period we mean the sequence of trials before the last error for any given subject. We know that the subject cannot be in State 3 on any trial before the last error. On all trials of the presolution period the probability of a correct response should be equal either to $\frac{1}{2}$ (if no conditioning has occurred) or to $\frac{5}{8}$ (if exactly one of the two stimulus patterns has been conditioned to its correct response). Thus the proportion, which we denote by P_{ps} , of correct responses over the presolution trial sequence should fall in the interval

$$\frac{1}{2} \leq P_{ps} \leq \frac{5}{8},$$

and, in fact, the same bounds obtain for any subset of trials within the presolution sequence. Clearly, predictions from this model concerning presolution responding differ sharply from those derivable from any model that assumes a continuous increase in probability of correct responding during the presolution period; this model also differs, though not so sharply, from a pure "insight" model that assumes no learning on presolution trials. As far as we know, no data relevant to these differential predictions are available in the literature (though similar predictions have been tested in somewhat different situations: Suppes & Ginsberg, 1963; Theios, 1963). Now that the predictions are in hand, it seems likely that pertinent analyses will be forthcoming.

The development in this section was for the case in which there were only three cues, a , b , and c . For the more general case we could assume that there are N_a cues associated with stimulus a , N_b with stimulus b , and N_c with stimulus c . If we assume, as we have in this section, that experimental conditions are such to ensure the subject's sampling all cues presented on each trial, then Eq. 87 may be rewritten as

$$\Pr(A_{1,n} | ac) = 1 - \frac{1}{2}(1 + w_1)\left(1 - \frac{c}{2}\right)^{n-1} + \frac{1}{2}w_1(1 - c)^{n-1}$$

$$\Pr(A_{2,n} | bc) = 1 - \frac{1}{2}(1 + w_2)\left(1 - \frac{c}{2}\right)^{n-1} + \frac{1}{2}w_2(1 - c)^{n-1},$$

where

$$w_1 = \frac{N_c}{N_a + N_c} \quad \text{and} \quad w_2 = \frac{N_c}{N_b + N_c}.$$

Further,

$$\begin{aligned} e &= \sum_{n=1}^{\infty} \left\{ \frac{1}{2} [1 - \Pr(A_{1,n} | ac)] + \frac{1}{2} [1 - \Pr(A_{2,n} | bc)] \right\} \\ &= \frac{1}{c} \left(1 + \frac{1}{2} \bar{w} \right), \end{aligned}$$

where $\bar{w} = \frac{1}{2}(w_1 + w_2)$. The parameter \bar{w} is an index of similarity between the stimuli ac and bc ; as \bar{w} approaches its maximum value of 1, the number of total errors increases. Further, the proportion of correct responses over the presolution trial sequence should fall in the interval

$$\frac{1}{2} \leq P_{ps} \leq \frac{1}{2} + \frac{1}{4}(1 - w_1)$$

or in the interval

$$\frac{1}{2} \leq P_{ps} \leq \frac{1}{2} + \frac{1}{4}(1 - w_2),$$

depending on whether ac or bc is conditioned first.

5.3 Component Models

As long as the number of stimulus patterns involved in a discrimination experiment is relatively small, an analysis in terms of an appropriate case of the mixed model can be effected along the lines indicated in Sec. 5.2. But the number of cues need become only moderately large in order to generate a number of patterns so great as to be unmanageable by these methods. However, if the number of patterns is large enough so that any particular pattern is unlikely to be sampled more than once during an experiment, the emendations of the response rule presented in Sec. 5.2 can be neglected and the process treated as a simple extension of the component model of Sec. 4.1.

Suppose, for example, that a classical discrimination involved a set, S_1 , of cues available only on trials when A_1 is reinforced, a set, S_2 , of cues available only on trials when A_2 is reinforced, and a set, S_c , of cues available on all trials; further, assume that a constant fraction of each set presented is sampled by the subject on any trial. If the two types of trials occur with equal probabilities and if the numbers of cues in the various sets are large enough so that the number of possible trial samples is larger than the number of trials in the experiment, then we may apply Eq. 53 of Sec. 3.3 to obtain approximate expressions for response probabilities. For example, asymptotically all of the N_1 elements of S_1 and half of the N_c elements of S_c

(on the average) would be conditioned to response A_1 , and therefore probability of A_1 on a trial when S_1 was presented would be predicted by the component model to be

$$\Pr(A_1 | S_1) = \frac{N_1 + \frac{1}{2}N_c}{N_1 + N_c},$$

which will, in general, have a value intermediate between $\frac{1}{2}$ and unity. Functions for learning curves and other aspects of the data can be derived for various types of discrimination experiments from the assumptions of the component model. Numerous results of this sort have been published (Burke & Estes, 1957; Bush & Mosteller, 1951b; Estes, 1958, 1961a; Estes, Burke, Atkinson & Frankmann, 1957; Popper, 1959; Popper & Atkinson, 1958).

5.4 Analysis of a Signal Detection Experiment

Although, so far, we have developed stimulus sampling models only in connection with simple associative learning and discrimination learning, it should be noted that such models may have much broader areas of application. On occasion we may even see possibilities of using the concepts of stimulus sampling and association to interpret experiments that, by conventional classifications, do not fall within the area of learning. In this section we examine such a case.

The experiment to be considered fits one of the standard paradigms associated with studies of signal detection (see, e.g., Tanner & Swets, 1954; Swets, Tanner, & Birdsall, 1961; or Chapter 3, Vol. 1, by Luce). The subject's task in this experiment, like that of an observer monitoring a radar screen, is to detect the presence of a visual signal which may occur from time to time in one of several possible locations. Problems of interest in connection with theories of signal detection arise when the signals are faint enough so that the observer is unable to report them with complete accuracy on all occasions. One empirical relation that we would want to account for, in quantitative detail, is that between detection probabilities and the relative frequencies with which signals occur in different locations. Another is the improvement in detection rate that may occur over a series of trials even when the observer receives no knowledge of results.

A possible way of accounting for the "practice effect" is suggested by some rather obvious analogies between the detection experiment and the probability learning experiment considered earlier: we expect that, when the subject actually detects a signal (in terms of stimulus sampling theory, samples the corresponding stimulus element), he will make the appropriate

verbal report. Further, in the absence of any other information, this detection of the signal may act as a reinforcing event, leading to conditioning of the verbal report to other cues in the situation which may have been available for sampling before the occurrence of the signal. If so, and if signals occur in some locations more often than in others, then on the basis of the theory developed in earlier sections we should predict that the subject will come to report the signal in the preferred location more frequently than in others on trials when he fails to detect a signal and is forced to respond to background cues. These notions are made more explicit in connection with the following analysis of a visual recognition experiment reported by Kinchla (1962).

Kinchla employed a forced-choice, visual-detection situation involving a series of more than 900 discrete trials for each subject. Two areas were outlined on a uniformly illuminated milk-glass screen. Each trial began with an auditory signal, during which one of the following events occurred:

1. A fixed increment in radiant intensity occurred in area 1—a T_1 -type trial.
2. A fixed increment in radiant intensity occurred in area 2—a T_2 -type trial.
3. No change in the radiant character of either signal area occurred—a T_0 -type trial.

Subjects were told that a change in illumination would occur in one of the two areas on each trial. Following the auditory signal, the subject was required to make either an A_1 - or an A_2 -response (i.e., select one of two keys placed below the signal area) to indicate the area he believed had changed in brightness. The subject was given no information at the end of the trial as to whether his response was correct. Thus, on a given trial, one of three events occurred (T_1 , T_2 , T_0), the subject made either an A_1 - or an A_2 -response, and a short time later the next trial began.

For a fixed signal intensity, the experimenter has the option of specifying a schedule for presenting the T_i -events. Kinchla selected a simple probabilistic procedure in which $\Pr(T_{i,n}) = \xi_i$ and $\xi_1 + \xi_2 + \xi_0 = 1$. Two groups of subjects were run. For Group I, $\xi_1 + \xi_2 = 0.4$ and $\xi_0 = 0.2$. For Group II, $\xi_1 = \xi_0 = 0.2$ and $\xi_2 = 0.6$. The purpose of Kinchla's study was to determine how these event schedules influenced the likelihood of correct detections.

The model that we shall use to analyze the experiment combines two quite distinct processes: a simple perceptual process defined with regard to the signal events and a learning process associated with background cues. The stimulus situation is conceptually represented in terms of two *sensory elements*, s_1 and s_2 , corresponding to the two alternative signals,

and a set, S , of elements associated with stimulus features common to all trials. On every trial the subject is assumed to sample a single element from the background set S , and he may or may not sample one of the sensory elements. If the s_1 element is sampled, an A_1 occurs; if s_2 is sampled, an A_2 occurs. If neither sensory element is sampled, the subject makes the response to which the background element is conditioned. Conditioning of elements in S changes from trial to trial via a learning process.

The sampling of sensory elements depends on the trial type (T_1 , T_2 , T_0) and is described by a simple probabilistic model. The learning process associated with S is assumed to be the multi-element pattern model presented in Sec. 2. Specifically, the assumptions of the model are embodied in the following statements:

1. If T_i ($i = 1, 2$) occurs, then sensory element s_i will be sampled with probability h (with probability $1 - h$ neither s_1 nor s_2 will be sampled). If T_0 occurs, then neither s_1 nor s_2 will be sampled.

2. Exactly one element is sampled from S on every trial. Given the set S of N elements, the probability of sampling a particular element is $1/N$.

3. If s_i ($i = 1, 2$) is sampled on trial n , then with probability c' the element sampled from S on the trial becomes conditioned to A_i at the end of trial n . If neither s_1 nor s_2 is sampled, then with probability c the element sampled from S becomes conditioned with equal likelihood to A_1 or A_2 at the end of trial n .

4. If sensory element s_i is sampled, then A_i will occur. If neither sensory element is sampled, then the response to which the sampled element from S is conditioned will occur.

If we let p_n denote the expected proportion of elements in S conditioned to A_1 at the start of trial n , then (in terms of statements 1 and 4) we can immediately write an expression for the likelihood of an A_i -response, given a T_j -event, namely,

$$\Pr(A_{1,n} | T_{1,n}) = h + (1 - h)p_n, \quad (88a)$$

$$\Pr(A_{2,n} | T_{2,n}) = h + (1 - h)(1 - p_n), \quad (88b)$$

$$\Pr(A_{1,n} | T_{0,n}) = p_n. \quad (88c)$$

The expression for p_n can be obtained from Statements 2 and 3 by the same methods used throughout Sec. 2 of this chapter (for a derivation of this result, see Atkinson, 1963a):

$$p_n = p_\infty - (p_\infty - p_1) \left[1 - \frac{1}{N}(a + b) \right]^{n-1},$$

where $a = \xi_1 hc' + (1 - h)(c/2) + \xi_0 h(c/2)$, $b = \xi_2 hc' + (1 - h)(c/2) + \xi_0 h(c/2)$, and $p_\infty = a/(a + b)$. Division of the numerator and denominator of p_∞ by c yields the expression

$$p_\infty = \frac{\xi_1 h \psi + \frac{1}{2}(1 - h) + \xi_0 h \frac{1}{2}}{(1 - \xi_0)(1 - h + h\psi) + \xi_0}, \quad (89)$$

where $\psi = c'/c$. Thus the asymptotic expression for p_n does not depend on the absolute values of c' and c but only on their ratio.

An inspection of Kinchla's data indicates that the curves for $\Pr(A_i | T_j)$ are extremely stable over the last 400 or so trials of the experiment; consequently we shall view this portion of the data as asymptotic. Table 7

Table 7 Predicted and Observed Asymptotic Response Probabilities for Visual Detection Experiment

| | Group I | | Group II | |
|------------------|----------|-----------|----------|-----------|
| | Observed | Predicted | Observed | Predicted |
| $\Pr(A_1 T_1)$ | 0.645 | 0.645 | 0.558 | 0.565 |
| $\Pr(A_2 T_2)$ | 0.643 | 0.645 | 0.730 | 0.724 |
| $\Pr(A_1 T_0)$ | 0.494 | 0.500 | 0.388 | 0.388 |

presents the observed mean values of $\Pr(A_i | T_j)$ for the last 400 trials. The corresponding asymptotic expressions are specified in terms of Eqs. 88 and 89 and are simply

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n} | T_{1,n}) = h + (1 - h)p_\infty, \quad (90a)$$

$$\lim_{n \rightarrow \infty} \Pr(A_{2,n} | T_{2,n}) = h + (1 - h)(1 - p_\infty), \quad (90b)$$

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n} | T_{0,n}) = p_\infty. \quad (90c)$$

In order to generate asymptotic predictions, we need values for h and ψ . We first note by inspection of Eq. 89 that $p_\infty = \frac{1}{2}$ for Group I; in fact, whenever $\xi_1 = \xi_2$, we have $p_\infty = \frac{1}{2}$. Hence taking the observed asymptotic value for $\Pr(A_1 | T_1)$ in Group I (i.e., 0.645) and setting it equal to $h + (1 - h)\frac{1}{2}$ yields an estimate of $h = 0.289$. The background illumination and the increment in radiant intensity are the same for both experimental groups, and therefore we would require an estimate of h obtained from Group I to be applicable to Group II. In order to estimate ψ , we take the observed asymptotic value of $\Pr(A_1 | T_0)$ in Group II and set it equal to the right side of Eq. 89 with $h = 0.289$, $\xi_1 = \xi_0 = 0.2$, and $\xi_2 = 0.6$; solving for ψ , we obtain $\hat{\psi} = 2.8$. Use of these estimates of h and ψ in Eqs. 89 and 90 yields the asymptotic predictions given in Table 7.

Over-all, the equations give an excellent account of these particular response measures. However, a more crucial test of the model is provided by an analysis of the sequential data. To indicate the nature of the sequential predictions that can be obtained, consider the probability of an A_1 -response on a T_1 -trial, given the various trial types and responses that can occur on the preceding trial, that is,

$$\Pr(A_{1,n+1} \mid T_{1,n+1}A_{i,n}T_{j,n}),$$

where $i = 1, 2$ and $j = 0, 1, 2$. Explicit expressions for these quantities can be derived from the axioms by the same methods used throughout this chapter. To indicate their form, theoretical expressions for

$$\lim_{n \rightarrow \infty} \Pr(A_{1,n+1} \mid T_{1,n+1}A_{i,n}T_{j,n})$$

are given, and, to simplify notation, they are written as $\Pr(A_1 \mid T_1A_iT_j)$. The expressions for these quantities are as follows:

$$\Pr(A_1 \mid T_1A_1T_1) = \frac{[h + (1-h)\delta]p_\infty + (1-p_\infty)h\gamma'}{NX} + \frac{(N-1)X}{N}, \quad (91a)$$

$$\Pr(A_1 \mid T_1A_2T_1) = \frac{(1-h)\delta'(1-p_\infty)}{N(1-X)} + \frac{(N-1)X}{N}, \quad (91b)$$

$$\Pr(A_1 \mid T_1A_2T_2) = \frac{h\gamma p_\infty + [h^2 + (1-h)\delta'](1-p_\infty)}{NY} + \frac{(N-1)X}{N}, \quad (91c)$$

$$\Pr(A_1 \mid T_1A_1T_2) = \frac{(1-h)\delta p_\infty}{N(1-Y)} + \frac{(N-1)X}{N}, \quad (91d)$$

$$\Pr(A_1 \mid T_1A_1T_0) = \frac{\delta}{N} + \frac{(N-1)X}{N}, \quad (91e)$$

$$\Pr(A_1 \mid T_1A_2T_0) = \frac{\delta'}{N} + \frac{(N-1)X}{N}, \quad (91f)$$

where

$$\gamma = c'h + (1-c'),$$

$$\gamma' = c' + (1-c')h,$$

$$\delta = (c/2)h + [1 - (c/2)],$$

$$\delta' = (c/2) + [1 - (c/2)]h,$$

and

$$X = h + (1-h)p_\infty,$$

$$Y = h + (1-h)(1-p_\infty).$$

It is interesting to note that the asymptotic expressions for $\Pr(A_{1,n} \mid T_{j,n})$ depend only on h and ψ , whereas the quantities in Eq. 91 are functions of

all four parameters N , c , c' , and h . Comparable sets of equations can be written for $\Pr(A_2 | T_2A_1T_j)$ and $\Pr(A_1 | T_0A_iT_j)$.

The expressions in Eq. 91 are rather formidable, but numerical predictions can be easily calculated once values for the parameters have been obtained. Further, independent of the parameter values, certain relations among the sequential probabilities can be specified. As an example of such

Table 8 Predicted and Observed Asymptotic Sequential Response Probabilities in Visual-Detection Experiment

| | Group I | | Group II | |
|------------------------|----------|-----------|----------|-----------|
| | Observed | Predicted | Observed | Predicted |
| $\Pr(A_2 T_2A_1T_1)$ | 0.57 | 0.58 | 0.59 | 0.64 |
| $\Pr(A_2 T_2A_2T_1)$ | 0.65 | 0.69 | 0.70 | 0.76 |
| $\Pr(A_2 T_2A_2T_2)$ | 0.71 | 0.71 | 0.79 | 0.77 |
| $\Pr(A_2 T_2A_1T_2)$ | 0.61 | 0.59 | 0.69 | 0.66 |
| $\Pr(A_2 T_2A_1T_0)$ | 0.54 | 0.59 | 0.68 | 0.66 |
| $\Pr(A_2 T_2A_2T_0)$ | 0.66 | 0.70 | 0.71 | 0.76 |
| $\Pr(A_1 T_1A_1T_1)$ | 0.73 | 0.71 | 0.70 | 0.65 |
| $\Pr(A_1 T_1A_2T_1)$ | 0.62 | 0.59 | 0.59 | 0.52 |
| $\Pr(A_1 T_1A_2T_2)$ | 0.53 | 0.58 | 0.53 | 0.51 |
| $\Pr(A_1 T_1A_1T_2)$ | 0.66 | 0.70 | 0.64 | 0.64 |
| $\Pr(A_1 T_1A_1T_0)$ | 0.72 | 0.70 | 0.61 | 0.63 |
| $\Pr(A_1 T_1A_2T_0)$ | 0.61 | 0.59 | 0.48 | 0.52 |
| $\Pr(A_2 T_0A_1T_1)$ | 0.38 | 0.40 | 0.47 | 0.49 |
| $\Pr(A_2 T_0A_2T_1)$ | 0.56 | 0.58 | 0.59 | 0.66 |
| $\Pr(A_2 T_0A_2T_2)$ | 0.64 | 0.60 | 0.67 | 0.68 |
| $\Pr(A_2 T_0A_1T_2)$ | 0.47 | 0.42 | 0.51 | 0.51 |
| $\Pr(A_2 T_0A_1T_0)$ | 0.47 | 0.42 | 0.50 | 0.51 |
| $\Pr(A_2 T_0A_2T_0)$ | 0.60 | 0.58 | 0.65 | 0.66 |

a relation, it can be shown that $\Pr(A_1 | T_1A_1T_0) \geq \Pr(A_1 | T_1A_2T_0)$ for any stimulus schedule and any set of parameter values. To see this, simply subtract Eq. 91f from Eq. 91e and note that $\delta \geq \delta'$.

In Table 8 the observed values for $\Pr(A_i | T_jA_kT_l)$ are presented as reported by Kinchla. Estimates of these conditional probabilities were computed for individual subjects, using the data over the last 400 trials; the averages of these individual estimates are the quantities given in the table. Each entry is based on 24 subjects.

In order to generate theoretical predictions for the observed entries in Table 8, values for N , c , c' , and h are needed. Of course, estimates of h and $\psi = c'/c$ have already been made for this set of data, and therefore it is

necessary only to estimate N and either c or c' . We obtain our estimates of N and c by a least-squares method; that is, we select a value of N and c (where $c' = c\psi$) so that the sum of squared deviations between the 36 observed values in Table 8 and the corresponding theoretical quantities is minimized. The theoretical quantities for $\Pr(A_1 | T_1 A_i T_j)$ are computed from Eq. 91; theoretical expressions for $\Pr(A_2 | T_2 A_i T_j)$ and $\Pr(A_2 | T_0 A_i T_j)$ have not been presented here but are of the same general form as those given in Eq. 91.

With this technique, estimates of the parameters are as follows:

$$\begin{aligned} N &= 4.23 & c' &= 1.00 \\ h &= 0.289 & c &= 0.357. \end{aligned} \tag{92}$$

The predictions corresponding to these parameter values are presented in Table 8. When we note that only four of the possible 36 degrees of freedom represented in Table 8 have been utilized in estimating parameters, the close correspondence between theoretical and observed quantities may be interpreted as giving considerable support to the assumptions of the model.

A great deal of research needs to be done to explore the consequences of this approach to signal detection. In terms of the experimental problem considered in this section, much progress can be made via differential tests among alternative formulations of the model. For example, we postulated a multi-element pattern model to describe the learning process associated with background stimuli; it would be important to determine whether other formulations of the learning process such as those developed in Sec. 4 or those proposed by Bush and Mosteller (1955) would provide as good or even better theoretical fits than the ones displayed in Tables 7 and 8. Also, it would be valuable to examine variations in the scheme for sampling sensory elements along lines developed by Luce (1959, 1963) and Restle (1961).

More generally, further development of the theory is required before we can attempt to deal with the wide range of empirical phenomena encompassed in the approach to perception via decision theory proposed by Swets, Tanner, and Birdsall (1961) and others. Some theoretical work has been done by Atkinson (1963b) along the lines outlined in this section to account for the *ROC* (receiver-operating-characteristic) curves that are typically observed in detection studies and to specify the relation between forced-choice and yes-no experiments. However, this work is still quite tentative, and an evaluation of the approach will require extensive analyses of the detailed sequential properties of psychophysical data.

5.5 Multiple-Process Models

Analyses of certain behavioral situations have proved to require formulations in terms of two or more distinguishable, though possibly interdependent, learning processes that proceed simultaneously. For some situations these separate processes may be directly observable; for other situations we may find it advantageous to postulate processes that are unobservable but that determine in some well-defined fashion the sequence of observable behaviors. For example, in Restle's (1955) treatment of discrimination learning it is assumed that irrelevant stimuli may become "adapted" over a period of time and thus be rendered nonfunctional. Such an analysis entails a two-process system. One process has to do with the conditioning of stimuli to responses, whereas the other prescribes both the conditions under which cues become irrelevant and the rate at which adaptation occurs.

Another application of multiple-process models arises with regard to discrimination problems in which either a covert or a directly observable orienting response is required. One process might describe how the stimuli presented to the subject become conditioned to discriminative responses. Another might specify the acquisition and extinction of various orienting responses; these orienting responses would determine the specific subset of the environment that the subject would perceive on a given trial. For models dealing with this type of problem, see Atkinson (1958), Bush & Mosteller (1951b), Bower (1959), and Wyckoff (1952).

As another example, consider a two-process scheme developed by Atkinson (1960) to account for certain types of discrimination behavior. This model makes use of the distinction, developed in Secs. 2 and 3, between component models and pattern models and suggests that the subject may (at any instant in time) perceive the stimulus situation either as a unit pattern or as a collection of individual components. Thus two perceptual states are defined: one in which the subject responds to the pattern of stimulation and one in which he responds to the separate components of the situation. Two learning processes are also defined. One process specifies how the patterns and components become conditioned to responses, and the second process describes the conditions under which the subject shifts from one perceptual state to another. The control of the second process is governed by the reinforcing schedule, the subject's sequence of responses, and by similarity of the discriminanda. In this model neither the conditioning states nor the perceptual states are observable; nevertheless, the behavior of the subject is rigorously defined in terms of these hypothetical states.

Models of the sort described are generally difficult to work with mathematically and consequently have had only limited development and analysis. It is for this reason that we select a particularly simple example to illustrate the type of formulation that is possible. The example deals with a discrimination-learning task investigated by Atkinson (1961) in which observing responses are categorized and directly measured.

The experimental situation consists of a sequence of discrete trials. Each trial is specified in terms of the following classifications:

- T_1, T_2 : *Trial type*. Each trial is either a T_1 or a T_2 . The trial type is set by the experimenter and determines *in part* the stimulus event occurring on the trial.
- R_1, R_2 : *Observing responses*. On each trial the subject makes either an R_1 or R_2 . The particular observing response determines in part the stimulus event for that trial.
- s_1, s_b, s_2 : *Stimulus events*. Following the observing response, one and only one of these stimulus events (discriminative cues) occurs. On a T_1 -trial either s_1 or s_b can occur; on a T_2 -trial either s_2 or s_b can occur.¹⁵
- A_1, A_2 : *Discriminative responses*. On each trial the subject makes either an A_1 - or A_2 -response to the presentation of a stimulus event.
- O_1, O_2 : *Trial outcome*. Each trial is terminated with the occurrence of one of these events. An O_1 indicates that A_1 was the correct response for that trial and O_2 indicates that A_2 was correct.

The sequence of events on a trial is as follows: (1) The ready signal occurs and the subject responds with R_1 or R_2 . (2) Following the observing response, $s_1, s_2,$ or s_b is presented. (3) To the onset of the stimulus event the subject responds with either A_1 or A_2 . (4) The trial terminates with either an O_1 - or O_2 -event.

To keep the analysis simple, we consider an experimenter-controlled reinforcement schedule. On a T_1 -trial either an O_1 occurs with probability π_1 or an O_2 with probability $1 - \pi_1$; on a T_2 -trial an O_1 occurs with probability π_2 or an O_2 with probability $1 - \pi_2$. The T_1 -trial occurs with probability β and T_2 with probability $1 - \beta$. Thus a T_1 - O_1 -combination occurs with probability $\beta\pi_1$, a T_1 - O_2 , with probability $\beta(1 - \pi_1)$, and so on.

The particular stimulus event s_i ($i = 1, 2, b$) that the experimenter

¹⁵ The subscript b has been used to denote the stimulus event that may occur on *both* T_1 - and T_2 -trials; the subscripts 1 and 2 denote stimulus events unique to T_1 - and T_2 -trials, respectively.

presents on any trial depends on the trial type (T_1 or T_2) and the subject's observing response (R_1 or R_2).

1. If an R_1 is made, then
 - (a) with probability α the s_1 -event occurs on a T_1 -trial and the s_2 -event on a T_2 -trial;
 - (b) with probability $1 - \alpha$ the s_b -event occurs, regardless of the trial type.
2. If an R_2 is made, then
 - (a) with probability α the s_b -event occurs, regardless of the trial type;
 - (b) with probability $1 - \alpha$ the s_1 -event occurs on a T_1 -trial and s_2 on a T_2 -trial.

To clarify this procedure, consider the case in which $\alpha = 1$, $\pi_1 = 1$, and $\pi_2 = 0$. If the subject is to be correct on every trial, he must make an A_1 on a T_1 -trial and an A_2 on a T_2 -trial. However, the subject can ascertain the trial type only by making the appropriate observing response; that is, R_1 must be made in order to identify the trial type, for the occurrence of R_2 always leads to the presentation of s_b , regardless of the trial type. Hence for perfect responding the subject must make R_1 with probability 1 and then make A_1 to s_1 or A_2 to s_2 . The purpose of the Atkinson study was to determine how variations in π_1 , π_2 , and α would affect both the observing responses and the discriminative responses.

Our analysis of this experimental procedure is based on the axioms presented in Secs. 1 and 2. However, in order to apply the theory, we must first identify the stimulus and reinforcing events in terms of the experimental operations. The identification we offer seems quite natural to us and is in accord with the formulations given in Secs. 1 and 2.

We assume that associated with the ready signal is a set S_R of pattern elements. Each element in S_R is conditioned to the R_1 - or the R_2 -observing response; there are N' such elements. At the start of each trial (i.e., with the onset of the ready signal) an element is sampled from S_R , and the subject makes the response to which the element is conditioned.

Associated with each stimulus event, s_i ($i = 1, 2, b$), is a set, S_i , of pattern elements; elements in S_i are conditioned to the A_1 - or the A_2 -discriminative response. There are N such elements in each set, S_i , and for simplicity we assume that the sets are pairwise disjoint. When the stimulus event s_i occurs, one element is randomly sampled from S_i , and the subject makes the discriminative response to which the element is conditioned.

Thus we have two types of learning processes: one defined on the set S_R and the other defined on the sets S_1 , S_b , and S_2 . Once the reinforcing

events have been specified for these processes, we can apply our axioms. The interpretation of reinforcement for the discriminative-response process is identical to that given in Sec. 2. If a pattern element is sampled from set S_i for $i = 1, 2, b$ and is followed by an O_j outcome, then with probability c the element becomes conditioned to A_j and with probability $1 - c$ the conditioning state of the sampled element remains unchanged.

The conditioning process for the S_R set is somewhat more complex in that the reinforcing events for the observing responses are assumed to be subject-controlled. Specifically, if an element conditioned to R_i is sampled from S_R and followed by either an A_1O_1 - or A_2O_2 -event, then the element will remain conditioned to R_i ; however, if A_1O_2 or A_2O_1 occurs, then with probability c' the element will become conditioned to the *other* observing response. Otherwise stated, if an element from S_R elicits an observing response that selects a stimulus event and, in turn, the stimulus event elicits a correct discriminative response (i.e., A_1O_1 or A_2O_2), then the sampled element will remain conditioned to that observing response. However, if the observing response selects a stimulus event that gives rise to an incorrect discriminative response (i.e., A_1O_2 or A_2O_1), then there will be a decrement in the tendency to repeat that observing response on the next trial.

Given the foregoing identification of events, we can now generate a mathematical model for the experiment. To simplify the analysis, we let $N' = N = 1$; namely, we assume that there is one element in each of our stimulus sets and consequently the single element is sampled with probability 1 whenever the set is available. With this restriction we may describe the conditioning state of a subject at the start of each trial by an ordered four-tuple $\langle ijkl \rangle$:

1. The first member i is 1 or 2 and indicates whether the single element of S_R is conditioned to R_1 or R_2 .
2. The second member j is 1 or 2 and indicates whether the single element of S_1 is conditioned to A_1 or A_2 .
3. The third member k is 1 or 2 and indicates whether the element of S_b is conditioned to A_1 or A_2 .
4. The fourth member l is 1 or 2 and indicates whether the element of S_2 is conditioned to A_1 or A_2 .

Thus, if the subject is in state $\langle ijkl \rangle$, he will make the R_i observing response; then, to s_1 , s_b , or s_2 , he will make discriminative response A_j , A_k , or A_l , respectively.

From our assumptions it follows that the sequence of random variables that take the subject states $\langle ijkl \rangle$ as values is a 16-state Markov chain.

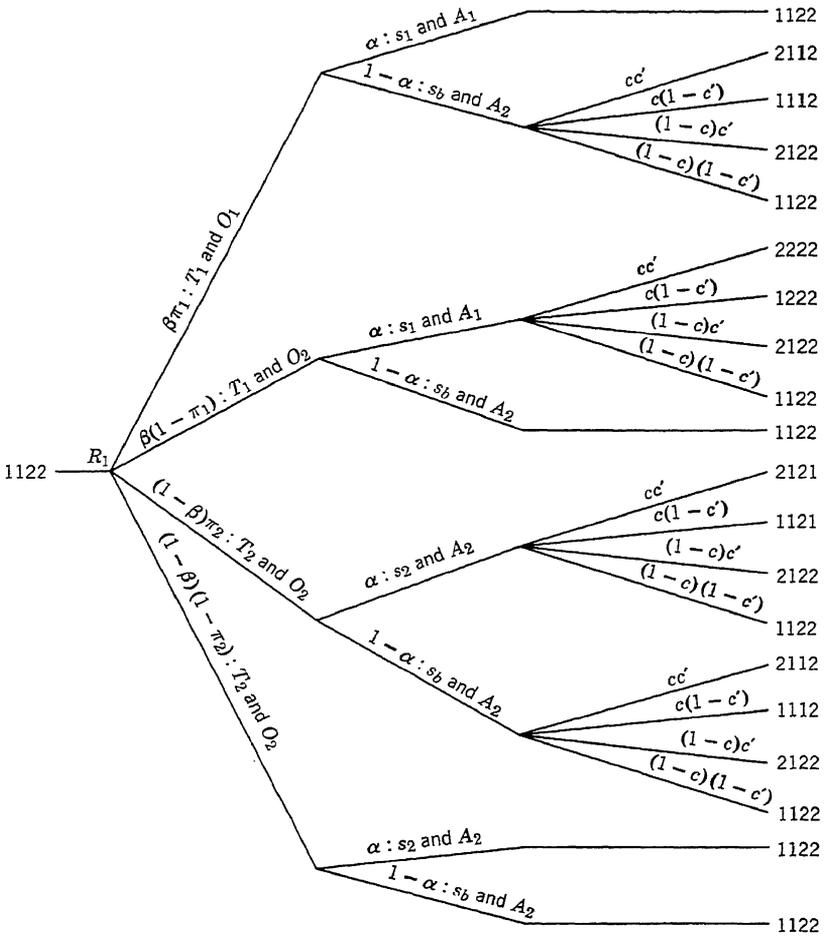


Fig. 10. Branching process, starting in state $\langle 1122 \rangle$, for a single trial in the two-process discrimination-learning model.

Figure 10 displays the possible transitions that can occur when the subject is in state $\langle 1122 \rangle$ on trial n . To clarify this tree, let us trace out the top branch. An R_1 is elicited with probability 1, and with probability $\beta\pi_1$ a T_1 -trial with an O_1 -outcome will occur; further, given an R_1 -response on a T_1 -trial, there is probability α that the s_1 -stimulus event will occur; the onset of the s_1 -event elicits a correct response, hence no change occurs in the conditioning state of any of the stimulus patterns. Now consider the next set of branches: an R_1 occurs and we have a T_1O_1 -trial; with probability $1 - \alpha$ the s_b -stimulus will be presented and an A_2 will occur; the A_2 -response is incorrect (in that it is followed by an O_1 -event); hence

with probability c the element of set S_b will become conditioned to A_1 and with independent probability c' the element of set S_R will become conditioned to the alternative observing response, namely R_2 .

From this tree we obtain probabilities corresponding to the $\langle 1122 \rangle$ row in the transition matrix. For example, the probability of going from $\langle 1122 \rangle$ to $\langle 2112 \rangle$ is simply $\beta\pi_1(1 - \alpha)cc' + (1 - \beta)\pi_2(1 - \alpha)cc'$; that is, the sum over branches 2 and 15. An inspection of the transition matrix yields important results. For example, if $\alpha = 1$, $\pi_1 = 1$, and $\pi_2 = 0$, then states $\langle 1112 \rangle$ and $\langle 1122 \rangle$ are absorbing, hence in the limit $\Pr(R_{1,n}) = 1$, $\Pr(A_{1,n} | T_{1,n}) = 1$, and $\Pr(A_{2,n} | T_{2,n}) = 1$.

As before, let $u_{ijkl}^{(n)}$ denote the probability of being in state $\langle ijkl \rangle$ on trial n ; when the limit exists, let $u_{ijkl} = \lim_{n \rightarrow \infty} u_{ijkl}^{(n)}$. Experimentally, we are interested in evaluating the following theoretical predictions:

$$\Pr(R_{1,n}) = u_{1111}^{(n)} + u_{1112}^{(n)} + u_{1121}^{(n)} + u_{1122}^{(n)} \\ + u_{1211}^{(n)} + u_{1212}^{(n)} + u_{1221}^{(n)} + u_{1222}^{(n)}, \quad (93a)$$

$$\Pr(A_{1,n} | T_{1,n}) = u_{1111}^{(n)} + u_{1112}^{(n)} + u_{2111}^{(n)} + u_{2112}^{(n)} \\ + \alpha[u_{1121}^{(n)} + u_{1122}^{(n)} + u_{2211}^{(n)} + u_{2212}^{(n)}] \\ + (1 - \alpha)[u_{1211}^{(n)} + u_{1212}^{(n)} + u_{2121}^{(n)} + u_{2122}^{(n)}], \quad (93b)$$

$$\Pr(A_{1,n} | T_{2,n}) = u_{1111}^{(n)} + u_{1211}^{(n)} + u_{2111}^{(n)} + u_{2211}^{(n)} \\ + \alpha[u_{1121}^{(n)} + u_{1221}^{(n)} + u_{2112}^{(n)} + u_{2212}^{(n)}] \\ + (1 - \alpha)[u_{1112}^{(n)} + u_{1212}^{(n)} + u_{2121}^{(n)} + u_{2221}^{(n)}], \quad (93c)$$

$$\Pr(R_{1,n} \cap A_{1,n}) = u_{1111}^{(n)} + \alpha u_{1121}^{(n)} + (1 - \alpha) u_{1212}^{(n)} \\ + \frac{1}{2}\alpha[u_{1122}^{(n)} + u_{1221}^{(n)}] \\ + (1 - \frac{1}{2}\alpha)(u_{1112}^{(n)} + u_{1211}^{(n)}), \quad (93d)$$

$$\Pr(R_{2,n} \cap A_{1,n}) = u_{2111}^{(n)} + \alpha u_{2212}^{(n)} + (1 - \alpha)u_{2121}^{(n)} \\ + \frac{1}{2}(1 - \alpha)[u_{2122}^{(n)} + u_{2221}^{(n)}] \\ + [1 - \frac{1}{2}(1 - \alpha)][u_{2112}^{(n)} + u_{2211}^{(n)}]. \quad (93e)$$

The first equation gives the probability of an R_1 -response. The second and third equations give the probability of an A_1 -response on T_1 - and T_2 -trials, respectively. Finally, the last two equations present the probability of the joint occurrence of each observing response with an A_1 -response.

In the experiment reported by Atkinson (1961) six groups with 40 subjects in each group were run. For all groups $\pi_1 = 0.9$ and $\beta = 0.5$. The groups differed with respect to the value of α and π_2 . For Groups I to III the value of $\alpha = 1$; and for Groups IV to VI $\alpha = 0.75$. For

Groups I and IV, $\pi_2 = 0.9$; for II and V, $\pi_2 = 0.5$; and for Groups III and VI, $\pi_2 = 0.1$. The design can be described by the following array:

| | | | | |
|----------|------|---------|-----|-----|
| | | π_2 | | |
| | | 0.9 | 0.5 | 0.1 |
| α | 1.0 | I | II | III |
| | 0.75 | IV | V | VI |

Given these values of π_1 , π_2 , α , and β , the 16-state Markov chain is irreducible and aperiodic. Thus $\lim u_{ijkl}^{(n)} = u_{ijkl}$ exists and can be obtained by solving the appropriate set of 16 linear equations (see Eq. 16).

Table 9 Predicted and Observed Asymptotic Response Probabilities in Observing Response Experiment

| | Group I | | | Group II | | | Group III | | |
|---------------------|----------|------|-------|----------|------|-------|-----------|------|-------|
| | Pred. | Obs. | SD | Pred. | Obs. | SD | Pred. | Obs. | SD |
| $\Pr(A_1 T_1)$ | 0.90 | 0.94 | 0.014 | 0.81 | 0.85 | 0.164 | 0.79 | 0.79 | 0.158 |
| $\Pr(A_1 T_2)$ | 0.90 | 0.94 | 0.014 | 0.59 | 0.61 | 0.134 | 0.21 | 0.23 | 0.182 |
| $\Pr(R_1)$ | 0.50 | 0.45 | 0.279 | 0.55 | 0.59 | 0.279 | 0.73 | 0.70 | 0.285 |
| $\Pr(R_1 \cap A_1)$ | 0.45 | 0.43 | 0.266 | 0.39 | 0.42 | 0.226 | 0.37 | 0.36 | 0.164 |
| $\Pr(R_2 \cap A_1)$ | 0.45 | 0.47 | 0.293 | 0.31 | 0.31 | 0.232 | 0.13 | 0.16 | 0.161 |
| | Group IV | | | Group V | | | Group VI | | |
| | Pred. | Obs. | SD | Pred. | Obs. | SD | Pred. | Obs. | SD |
| $\Pr(A_1 T_1)$ | 0.90 | 0.93 | 0.063 | 0.80 | 0.82 | 0.114 | 0.73 | 0.73 | 0.138 |
| $\Pr(A_1 T_2)$ | 0.90 | 0.95 | 0.014 | 0.60 | 0.68 | 0.114 | 0.27 | 0.25 | 0.138 |
| $\Pr(R_1)$ | 0.49 | 0.50 | 0.257 | 0.52 | 0.53 | 0.305 | 0.63 | 0.72 | 0.263 |
| $\Pr(R_1 \cap A_1)$ | 0.44 | 0.47 | 0.241 | 0.35 | 0.38 | 0.219 | 0.32 | 0.36 | 0.138 |
| $\Pr(R_2 \cap A_1)$ | 0.46 | 0.47 | 0.247 | 0.34 | 0.36 | 0.272 | 0.19 | 0.13 | 0.168 |

The values predicted by the model are given in Table 9 for the case in which $c = c'$. Values for the u_{ijkl} 's were computed and then combined by Eq. 93 to predict the response probabilities. By presenting a single value for each theoretical quantity in the table we imply that these predictions are independent of c and c' . Actually, this is not always the case. However, for the schedules employed in this experiment the dependency of these asymptotic predictions on c and c' is virtually negligible. For $c = c'$, ranging over the interval from 0.0001 to 1.0, the predicted values given in

Table 9 are affected in only the third or fourth decimal place; it is for this reason that we present theoretical values to only two decimal places.

In view of these comments it should be clear that the predictions in Table 9 are based solely on the experimental parameter values. Consequently, differences between subjects (that may be represented by inter-subject variability in c and c') do not substantially affect these predictions.

In the Atkinson study 400 trials were run and the response proportions appear to have reached a fairly stable level over the second half of the experiment. Consequently, the proportions computed over the final block of 160 trials were used as estimates of asymptotic quantities. Table 9 presents the mean and standard deviation of the 40 observed proportions obtained under each experimental condition.

Despite the fact that these gross asymptotic predictions hold up quite well, it is obvious that some of the predictions from the model will not be confirmed. The difficulty with the one-element assumption is that the fundamental theory laid down by the axioms of Sec. 2 is completely deterministic in many respects. For example, when $N' = 1$, we have

$$\Pr (R_{1,n+1} \mid O_{1,n}A_{1,n}R_{1,n}) = 1;$$

namely, if an R_1 occurs on trial n and is reinforced (i.e., followed by an A_1O_1 -event), then R_1 will recur with probability 1 on trial $n + 1$. This prediction is, of course, a consequence of the assumption that we have but one element in set S_R which necessarily is sampled on every trial. If we assume more than one element, the deterministic features of the model no longer hold, and such sequential statistics become functions of c , c' , N , and N' . Unfortunately, for elaborate experimental procedures of the sort described in this section the multi-element case leads to complicated mathematical processes for which it is extremely difficult to carry out computations. Thus the generality of the multi-element assumption may often be offset by the difficulty involved in making predictions.

Naturally, it is usually preferable to choose from the available models the one that best fits the data, but in the present state of psychological knowledge no single model is clearly superior to all others in every facet of analysis. The one-element assumption, despite some of its erroneous features, may prove to be a valuable instrument for the rapid exploration of a wide variety of complex phenomena. For most of the cases we have examined the predicted mean response probabilities are usually independent of (or only slightly dependent on) the number of elements assumed. Thus the one-element assumption may be viewed as a simple device for computing the grosser predictions of the general theory.

For exploratory work in complex situations, then, we recommend using the one-element model because of the greater difficulty of computations

for the multi-element models. In advocating this approach, we are taking a methodological position with which some scientists do not agree. Our position is in contrast to one that asserts that a model should be discarded once it is clear that certain of its predictions are in error. We do not take it to be the principal goal (or even, in many cases, an important goal) of theory construction to provide models for particular experimental situations. The assumptions of stimulus sampling theory are intended to describe processes or relationships that are common to a wide variety of learning situations but with no implication that behavior in these situations is a function solely of the variables represented in the theory. As we have attempted to illustrate by means of numerous examples, the formulation of a model within this framework for a particular experiment is a matter of selecting the relevant assumptions, or axioms, of the general theory and interpreting them in terms of the conditions of the experiment. How much of the variance in a set of data can be accounted for by a model depends jointly on the adequacy of the theoretical assumptions and on the extent to which it has been possible to realize experimentally the boundary conditions envisaged in the theory, thereby minimizing the effects of variables not represented. In our view a model, in application to a given experiment, is not to be classified as "correct" or "incorrect"; rather, the degree to which it accounts for the data may provide evidence tending either to support or to cast doubt on the theory from which it was derived.

References

- Atkinson, R. C. A stochastic model for rote serial learning. *Psychometrika*, 1957, **22**, 87-96.
- Atkinson, R. C. A Markov model for discrimination learning. *Psychometrika*, 1958, **23**, 308-322.
- Atkinson, R. C. A theory of stimulus discrimination learning. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences*. Stanford: Stanford Univer. Press, 1960. Pp. 221-241.
- Atkinson, R. C. The observing response in discrimination learning. *J. exp. Psychol.*, 1961, **62**, 253-262.
- Atkinson, R. C. Choice behavior and monetary payoffs. In J. Criswell, H. Solomon, & P. Suppes (Eds.), *Mathematical methods in small group processes*. Stanford: Stanford Univer. Press, 1962. Pp. 23-34.
- Atkinson, R. C. Mathematical models in research on perception and learning. In M. Marx (Ed.), *Psychological Theory*. (2nd ed.) New York: Macmillan, 1963, in press. (a)
- Atkinson, R. C. A variable sensitivity theory of signal detection. *Psychol. Rev.*, 1963, **70**, 91-106. (b)
- Atkinson, R. C., & Suppes, P. An analysis of two-person game situations in terms of statistical learning theory. *J. exp. Psychol.*, 1958, **55**, 369-378.

- Billingsley, P. *Statistical inference for Markov processes*. Chicago: Univer. of Chicago Press, 1961.
- Bower, G. H. Choice-point behavior. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 109-124.
- Bower, G. H. Application of a model to paired-associate learning. *Psychometrika*, 1961, **26**, 255-280.
- Bower, G. H. A model for response and training variables in paired-associate learning. *Psychol. Rev.*, 1962, **69**, 34-53.
- Burke, C. J. Applications of a linear model to two-person interactions. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 180-203.
- Burke, C. J. Some two-person interactions. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences*. Stanford: Stanford Univer. Press, 1960. Pp. 242-253.
- Burke, C. J., & Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika*, 1957, **22**, 133-145.
- Bush, R. R. A survey of mathematical learning theory. In R. D. Luce (Ed.), *Developments in mathematical psychology*. Glencoe, Illinois: The Free Press, 1960. Pp. 123-165.
- Bush, R. R., & Estes, W. K. (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959.
- Bush, R. R., & Mosteller, F. A mathematical model for simple learning. *Psychol. Rev.*, 1951, **58**, 313-323. (a)
- Bush, R. R., & Mosteller, F. A model for stimulus generalization and discrimination. *Psychol. Rev.*, 1951, **58**, 413-423. (b)
- Bush, R. R., & Mosteller, F. *Stochastic models for learning*. New York: Wiley, 1955.
- Bush, R. R., & Sternberg, S. A single-operator model. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 204-214.
- Carterette, Teresa S. An application of stimulus sampling theory to summated generalization. *J. exp. Psychol.*, 1961, **62**, 448-455.
- Crothers, E. J. *All-or-none paired associate learning with unit and compound responses*. Unpublished doctoral dissertation, Indiana University, 1961.
- Detambel, M. H. A test of a model for multiple-choice behavior. *J. exp. Psychol.*, 1955, **49**, 97-104.
- Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.
- Estes, W. K. Statistical theory of spontaneous recovery and regression. *Psychol. Rev.*, 1955, **62**, 145-154. (a)
- Estes, W. K. Statistical theory of distributional phenomena in learning. *Psychol. Rev.*, 1955, **62**, 369-377. (b)
- Estes, W. K. Of models and men. *Amer. Psychol.*, 1957, **12**, 609-617. (a)
- Estes, W. K. Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika*, 1957, **22**, 113-132. (b)
- Estes, W. K. Stimulus-response theory of drive. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Vol. 6. Lincoln, Nebraska: Univer. Nebraska Press, 1958.
- Estes, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: a study of a science*. Vol. 2. New York: McGraw-Hill, 1959. Pp. 380-491. (a)
- Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 9-52. (b)

- Estes, W. K. Learning theory and the new mental chemistry. *Psychol. Rev.*, 1960, **67**, 207-223. (a)
- Estes, W. K. A random-walk model for choice behavior. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences*. Stanford: Stanford Univer. Press, 1960. Pp. 265-276. (b)
- Estes, W. K. Growth and function of mathematical models for learning. In *Current trends in psychological theory*. Pittsburgh: Univer. of Pittsburgh Press, 1961. Pp. 134-151. (a)
- Estes, W. K. New developments in statistical behavior theory: differential tests of axioms for associative learning. *Psychometrika*, 1961, **26**, 73-84. (b)
- Estes, W. K. Learning theory. *Ann. Rev. Psychol.*, 1962, **13**, 107-144.
- Estes, W. K., & Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.
- Estes, W. K., Burke, C. J., Atkinson, R. C., & Frankmann, Judith P. Probabilistic discrimination learning. *J. exp. Psychol.*, 1957, **54**, 233-239.
- Estes, W. K., & Hopkins, B. L. Acquisition and transfer in pattern -vs.- component discrimination learning. *J. exp. Psychol.*, 1961, **61**, 322-328.
- Estes, W. K., Hopkins, B. L., & Crothers, E. J. All-or-none and conservation effects in the learning and retention of paired associates. *J. exp. Psychol.*, 1960, **60**, 329-339.
- Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.*, 1954, **47**, 225-234.
- Estes, W. K., & Suppes, P. Foundations of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 137-179. (a)
- Estes, W. K., & Suppes, P. *Foundations of statistical learning theory, II. The stimulus sampling model for simple learning*. Tech. Rept. No. 26, Psychology Series, Institute for Mathematical Studies in the Social Sciences, Stanford Univer., 1959. (b)
- Feller, W. *An introduction to probability theory and its applications*. (2nd ed.) New York: Wiley, 1957.
- Friedman, M. P., Burke, C. J., Cole, M., Estes, W. K., & Millward, R. B. *Extended training in a noncontingent two-choice situation with shifting reinforcement probabilities*. Paper given at the First Meetings of the Psychonomic Society, Chicago, Illinois, 1960.
- Gardner, R. A. Probability-learning with two and three choices. *Amer. J. Psychol.*, 1957, **70**, 174-185.
- Guttman, N., & Kalish, H. I. Discriminability and stimulus generalization. *J. exp. Psychol.*, 1956, **51**, 79-88.
- Goldberg, S. *Introduction to difference equations*. New York: Wiley, 1958.
- Hull, C. L. *Principles of behavior: an introduction to behavior theory*. New York: Appleton-Century-Crofts, 1943.
- Jarvik, M. E. Probability learning and a negative recency effect in the serial anticipation of alternating symbols. *J. exp. Psychol.*, 1951, **41**, 291-297.
- Jordan, C. *Calculus of finite differences*. New York: Chelsea, 1950.
- Kemeny, J. G., & Snell, J. L. Markov processes in learning theory. *Psychometrika*, 1957, **22**, 221-230.
- Kemeny, J. G., & Snell, J. L. *Finite Markov chains*. Princeton, N. J.: Van Nostrand, 1959.
- Kemeny, J. G., Snell, J. L., & Thompson, G. L. *Introduction to finite mathematics*. New York: Prentice Hall, 1957.
- Kinchla, R. A. *Learned factors in visual discrimination*. Unpublished doctoral dissertation, Univer. of California, Los Angeles, 1962.

- Lamperti, J., & Suppes, P. Chains of infinite order and their applications to learning theory. *Pacific J. Math.*, 1959, **9**, 739-754.
- Luce, R. D. *Individual choice behavior: a theoretical analysis*. New York: Wiley, 1959.
- Luce, R. D. A threshold theory for simple detection experiments. *Psychol. Rev.*, 1963, **70**, 61-79.
- Luce, R. D., & Raiffa, H. *Games and decisions*. New York: Wiley, 1957.
- Nicks, D. C. Prediction of sequential two-choice decisions from event runs. *J. exp. Psychol.*, 1959, **57**, 105-114.
- Peterson, L. R., Saltzman, Dorothy, Hillner, K., & Land, Vera. Recency and frequency in paired-associate learning. *J. exp. Psychol.*, 1962, **63**, 396-403.
- Popper, Juliet. Mediated generalization. In R. R. Bush & W. K. Estes, (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 94-108.
- Popper, Juliet, & Atkinson, R. C. Discrimination learning in a verbal conditioning situation. *J. exp. Psychol.*, 1958, **56**, 21-26.
- Restle, F. A theory of discrimination learning. *Psychol. Rev.*, 1955, **62**, 11-19.
- Restle, F. *Psychology of judgment and choice*. New York: Wiley, 1961.
- Solomon, R. L., & Wynne, L. C. Traumatic avoidance learning: acquisition in normal dogs. *Psychol. Monogr.*, 1953, **67**, No. 4.
- Spence, K. W. The nature of discrimination learning in animals. *Psychol. Rev.*, 1936, **43**, 427-449.
- Stevens, S. S. On the psychophysical law. *Psychol. Rev.*, 1957, **64**, 153-181.
- Suppes, P., & Atkinson, R. C. *Markov learning models for multiperson interactions*. Stanford: Stanford Univer. Press, 1960.
- Suppes, P., & Ginsberg, Rose. Application of a stimulus sampling model to children's concept formation of binary numbers with and without an overt correction response. *J. exp. Psychol.*, 1962, **63**, 330-336.
- Suppes, P., & Ginsberg, Rose. A fundamental property of all-or-none models. *Psychol. Rev.*, 1963, **70**, 139-161.
- Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processes in perception. *Psychol. Rev.*, 1961, **68**, 301-340.
- Tanner, W. P., Jr., & Swets, J. A. A decision-making theory of visual detection. *Psychol. Rev.*, 1954, **61**, 401-409.
- Theios, J. Simple conditioning as two-stage all-or-none learning. *Psychol. Rev.*, 1963, in press.
- Wyckoff, L. B., Jr. The role of observing responses in discrimination behavior. *Psychol. Rev.*, 1952, **59**, 431-442.

I I

*Introduction to the Formal Analysis of Natural Languages*¹

Noam Chomsky

Massachusetts Institute of Technology

George A. Miller

Harvard University

1. *The preparation of this Chapter was supported in part by the Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research, and in part by the National Science Foundation (Grants No. NSF G-16486 and No. NSF G-13903).*

Contents

| | |
|---|-----|
| 1. Limiting the Scope of the Discussion | 272 |
| 2. Some Algebraic Aspects of Coding | 277 |
| 3. Some Basic Concepts of Linguistics | 283 |
| 4. A Simple Class of Generative Grammars | 292 |
| 5. Transformational Grammars | 296 |
| 5.1. Some shortcomings of constituent-structure grammars, | 297 |
| 5.2. The specification of grammatical transformations, | 300 |
| 5.3. The constituent structure of transformed strings, | 303 |
| 6. Sound Structure | 306 |
| 6.1. The role of the phonological component, | 306 |
| 6.2. Phones and phonemes, | 308 |
| 6.3. Invariance and linearity conditions, | 310 |
| 6.4. Some phonological rules, | 313 |
| References | 319 |

Introduction to the Formal Analysis of Natural Languages

Language and communication play a special and important role in human affairs; they have been pondered and discussed by every variety of scholar and scientist. The psychologist's contribution has been but a small part of the total effort. In order to give a balanced picture of the larger problem that a mathematical psychologist faces when he turns his attention toward verbal behavior, therefore, this chapter and the next two must go well beyond the traditional bounds of psychology.

The fundamental fact that must be faced in any investigation of language and linguistic behavior is the following: a native speaker of a language has the ability to comprehend an immense number of sentences that he has never previously heard and to produce, on the appropriate occasion, novel utterances that are similarly understandable to other native speakers. The basic questions that must be asked are the following:

1. What is the precise nature of this ability?
2. How is it put to use?
3. How does it arise in the individual?

There have been several attempts to formulate questions of this sort in a precise and explicit form and to construct models that represent certain aspects of these achievements of a native speaker. When simple enough models can be constructed it becomes possible to undertake certain purely abstract studies of their intrinsic character and general properties. Studies of this kind are in their infancy; few aspects of language and communication have been formalized to a point at which such investigations are even thinkable. Nevertheless, there is a growing body of suggestive results. We shall survey some of those results here and try to indicate how such studies can contribute to our understanding of the nature and function of language.

The first of our three basic questions concerns the nature of language itself. In order to answer it, we must make explicit the underlying structure inherent in all natural languages. The principal attack on this problem has its origins in logic and linguistics; in recent years it has focused on the critically important concept of grammar. The justification for including this work in the present handbook is to make psychologists more realistically

aware of what it is a person has accomplished when he has learned to speak and understand a natural language. Associating vocal responses with visual stimuli—a feature that has attracted considerable psychological attention—is but one small aspect of the total language-learning process.

Our second question calls for an attempt to give a formal characterization of, or model for, the users of natural languages. Psychologists, who might have been expected to attack this question as part of their general study of behavior, have as yet provided only the most programmatic (and often implausible) kinds of answers. Some valuable ideas on this topic have originated in the field of communication engineering; their psychological implications were relatively direct and were promptly recognized. However, the engineering concepts have been largely statistical and have made little contact with what is known of the inherent structure of language.

By presenting (1) and (2) as two distinct questions we explicitly reject the common opinion that a language is nothing but a set of verbal responses. To say that a particular rule of grammar applies in some natural language is not to say that the people who use that language are able to follow the rule consistently. To specify the language is one task. To characterize its user is another. The two problems are obviously related but are not identical.

Our third question is no less important than the first two, yet far less progress has been made in formulating it in such a way as to support any abstract investigation. What goes on as a child begins to talk is still beyond the scope of our mathematical models. We can only mention the genetic issue and regret its relative neglect in the following pages.

1. LIMITING THE SCOPE OF THE DISCUSSION

The mathematical study of language and communication is a large topic. We must limit it sharply for our present purposes. It may help to orient the reader if we enumerate some of the limitations we have imposed in this and in the two chapters that follow.

The first limitation we imposed was to restrict our interest generally to the so-called natural languages. There are, of course, many formal languages developed by logicians and mathematicians; the study of those languages is a major concern of modern logic. In these pages, however, we have tried to limit our attention to the formal study of natural languages and largely ignore the study of formal languages. It is sometimes convenient to use miniature, artificial languages in order to illustrate a

particular property in a simplified context, and the programming languages developed by computer specialists are often of special interest. Nevertheless, the central focus here is on natural languages.

A further limitation was to eliminate all serious consideration of continuous systems. The acoustic signal produced by a speaker is a continuous function of time and is ordinarily represented as the sum of a Fourier series. Fourier representation is especially convenient when we study the effects of continuous linear transformations (filters). Fortunately this important topic has been frequently and thoroughly treated by both mathematicians and communication engineers; its absence here will not be critical.

Communication systems can be thought of as discrete because of the existence of what communication engineers have sometimes called a *fidelity criterion* (Shannon, 1949). A fidelity criterion determines how the set of all signals possible during a finite time interval should be partitioned into subsets of equivalent signals—equivalent for the receiver. A communication system may transmit continuous signals precisely, but if the receiver cannot (or will not) pay attention to the fine distinctions that the system is capable of registering the fidelity of the channel is wasted. Thus it is the receiver who establishes a criterion of acceptability for the system. The higher his criterion, the larger the number of distinct subsets of signals the communication system must be able to distinguish and transmit.

The receiver we wish to study is, of course, a human listener. The fidelity criterion is set by his capacities, training, and interests. On the basis of his perceptual distinctions, therefore, we can establish a finite set of categories to serve as the discrete symbols. Those sets may be alphabets, syllabaries, or vocabularies; the discrete elements of those sets are the indivisible atoms from which longer messages must be constructed. A listener's perception of those discrete units of course poses an important psychological problem; formal psychophysical aspects of the detection and recognition problem have already been discussed in Chapter 3, and we shall not repeat them here. However, certain considerations that may be unique for speech perception are mentioned briefly in Sec. 6, where we discuss the subject of sound structure, and again in Chapter 13, where we consider how our knowledge of grammar might serve to organize our perception of speech. As we shall see, the precise description of a human listener's fidelity criterion for speech is a complex thing, but for the moment the critical point is that people do partition speech sounds into equivalent subsets, so a discrete notation is justifiable.

In the realm of discrete systems, moreover, we limit ourselves to *concatenation* systems and to their further algebraic structure and their interrelations. In particular, we think of the flow of speech as a sequence of

discrete atoms that are immediately juxtaposed, or concatenated, one after the other. Simple as this limitation may sound, it has some implications worth noting.

Let L be the set of all finite sequences (including the sequence of zero length) that can be formed from the elements of some arbitrary finite set V . Now, if $\phi, \chi \in L$ and if $\phi \frown \chi$ represents the result of concatenating them in that order to form a new sequence ψ , then $\psi \in L$; that is to say, L is closed under the binary operation of concatenation. Furthermore, concatenation is associative,

$$(\phi \frown \chi) \frown \psi = \phi \frown (\chi \frown \psi),$$

and the empty sequence plays the role of a unique identity element. A set that includes an identity and is closed under an associative law of composition is called a *monoid*. Because monoids satisfy three of the four postulates of a group, they are sometimes called *semigroups*. A *group* is a monoid all of whose elements have inverses.

Although we must necessarily construct our spoken utterances by the associative operation of concatenation, the matter must be formulated carefully. Consider, for example, the ambiguous English sentence, *They are flying planes*, which is really two different sentences:

$$\text{They} \frown (\text{are} \frown (\text{flying} \frown \text{planes})). \quad (1a)$$

$$\text{They} \frown ((\text{are} \frown \text{flying}) \frown \text{planes}). \quad (1b)$$

If we think only of spelling or pronunciation, then Example 1a equals Example 1b and simple concatenation offers no difficulties. But if we think of grammatical structure or meaning, Examples 1a and 1b are distinctly different in a way that ordinarily is not phonetically or graphically indicated.

Linguists generally deal with such problems by assuming that a natural language has several distinct *levels*. In the present chapters, we think of each level as a separate concatenation system with its own elements and rules. The structure at a lower level is specified by the way in which its elements are related to the next higher level. In order to preserve associativity, therefore, we introduce several concatenation systems and study the relations between them.

Consider two different operations that we might perform on a written text. The first operation maps a sequence of written characters into a sequence of acoustic signals; let us refer to it as pronunciation. Pronunciations of segments of a message are (approximately) segments of the

pronunciation of that message. Thus pronunciation has about it a kind of linearity (cf. Sec. 6.3):

$$\text{pron } (x) \frown \text{pron } (y) = \text{pron } (x \frown y). \quad (2)$$

Although Eq. 2 is not true (e.g., it ignores intonation and articulatory transitions between successive segments), it is more nearly true than the corresponding statement for the next operation.

This operation maps the sequence of symbols into some representation of its subjective meaning; let us refer to it as comprehension. The meanings of segments of a message, however, are seldom identifiable as segments of the meaning of the message. Even if we assume that meanings can somehow be simply concatenated, in most cases we would probably find, under any reasonable interpretation of these notions, that

$$\text{comp } (x) \frown \text{comp } (y) \leq \text{comp } (x \frown y), \quad (3)$$

which is one way, perhaps, to interpret the Gestalt dictum that a meaningful whole is greater than the linear sum of its parts. Unless one is a confirmed associationist in the tradition of James Mill, it is not obvious what the concatenation of two comprehensions would mean or how such an operation could be performed. By comparison, the operations in Eq. 2 seem well defined.

To introduce the process of comprehension, however, raises many difficult issues. Recently there have been interesting proposals for studying abstractly certain denotative (Wallace, 1961) and connotative (Osgood, Suci, & Tannenbaum, 1957) aspects of natural lexicons. Important as this subject is for any general theory of psycholinguistics, we shall say little about it in these pages. Nevertheless, our hope is that by clearing away some syntactic problems we shall have helped to clarify the semantic issue if only by indicating some of the things that meaning is not.

Finally, as we have already noted, these chapters include little on the process of language learning. Although it is possible to give a formal description of certain aspects of language and although several mathematical models of the learning process have been developed, the intersection of these two theoretical endeavors remains disconcertingly vacant.

How an untutored child can so quickly attain full mastery of a language poses a challenging problem for learning theorists. With diligence, of course, an intelligent adult can use a traditional grammar and a dictionary to develop some degree of mastery of a new language; but a young child gains perfect mastery with incomparably greater ease and without any explicit instruction. Careful instruction and precise programming of reinforcement contingencies do not seem necessary. Mere exposure for a

remarkably short period is apparently all that is required for a normal child to develop the competence of a native speaker.

One way to highlight the theoretical questions involved here is to imagine that we had to construct a device capable of duplicating the child's learning (Chomsky, 1962a). It would have to include a device that accepted a sample of grammatical utterances as its input (with some restrictions, perhaps, on their order of presentation) and that would produce a grammar of the language (including the lexicon) as its output. A description of this device would represent a hypothesis about the innate intellectual equipment that a child brings to bear in acquiring a language. Of course, other input data may play an essential role in language learning. For example, corrections by the speech community are probably important. A correction is an indication that a certain linguistic expression is not a sentence. Thus the device may have a set of nonsentences, as well as a set of sentences, as an input. Furthermore, there may be indications that one item is to be considered a repetition of another, and perhaps other hints and helps. What other inputs are necessary is, of course, an important question for empirical investigation.

Equally important, however, is to specify the properties of the grammar that our universal language-learning device is supposed to produce as its output. This grammar is intended to represent certain of the abilities of a mature speaker. First, it should indicate how he is able to determine what is a properly formed sentence, and, second, it should provide information about the arrangements of the units into larger structures. The language-learning device must, for example, come to understand the difference between Examples 1*a* and 1*b*.

The characterization of a grammar that will provide an explicit enumeration of grammatical sentences, each with its own structural description, is a central concern in the pages that follow. What we seek is a formalized grammar that specifies the correct structural descriptions with a fairly small number of general principles of sentence formation and that is embedded within a theory of linguistic structure that provides a justification for the choice of this grammar over other alternatives. One task of the professional linguist is, in a sense, to make explicit the process that every normal child performs implicitly.

A practical language-learning device would have to incorporate strong assumptions about the class of potential grammars that a natural language can have. Presumably the device would have available an advance specification of the general form that a grammar might assume and also some procedure to decide whether a particular grammar is better than some alternative grammar on the basis of the sample input. Moreover, it would have to have certain phonetic capacities for recognizing and producing

sentences, and it would need to have some method, given one of the permitted grammars, to determine the structural description of any arbitrary sentence. All this would have to be built into the device in advance before it could start to learn a language. To imagine that an adequate grammar could be selected from the infinitude of conceivable alternatives by some process of pure induction on a finite corpus of utterances is to misjudge completely the magnitude of the problem.

The learning process, then, would consist in evaluating the various possible grammars in order to find the best one compatible with the input data. The device would seek a grammar that enumerated all the sentences and none of the nonsentences and assigned structural descriptions in such a way that nonrepetitions would differ at appropriate points. Of course, we would have to supply the language-learning device with some sort of heuristic principles that would enable it, given its input data and a range of possible grammars, to make a rapid selection of a few promising alternatives, which could then be submitted to a process of evaluation, or that would enable it to evaluate certain characteristics of the grammar before others. The necessary heuristic procedures could be simplified, however, by providing in advance a narrower specification of the class of potential grammars. The proper division of labor between heuristic methods and specification of form remains to be decided, of course, but too much faith should not be put in the powers of induction, even when aided by intelligent heuristics, to discover the right grammar. After all, stupid people learn to talk, but even the brightest apes do not.

2. SOME ALGEBRAIC ASPECTS OF CODING

Mapping one monoid into another is a pervasive operation in communication systems. We can refer to it rather loosely as *coding*—including in that term the various processes of encoding, recoding, decoding, and transmitting. In order to make this preliminary discussion definite, we can think of one monoid as consisting of all the strings that can be formed with the characters of a finite alphabet A and the other as consisting of the strings that can be formed by the words in a finite vocabulary V . In this section, therefore, we consider some abstract properties of concatenation systems in general, properties that apply equally to artificial and to natural codes.

A code C is a 1 : 1 mapping θ of strings in V into strings in A such that if v_i, v_j are strings in V then $\theta(v_i \frown v_j) = \theta(v_i) \frown \theta(v_j)$. θ is an isomorphism between strings in V and a subset of the strings in A ; strings in A provide the spellings for strings in V . In the following, if there is no danger of

confusion, we can simplify our notation by suppressing the symbol \frown for concatenation, thus adopting the normal convention for spelling systems.

Consider a simple example of a code C_1 . Let $A = \{0, 1\}$ and $V = \{v_1, \dots, v_4\}$. Define a mapping θ as follows:

$$\begin{aligned}\theta(v_1) &= 1, \\ \theta(v_2) &= 011, \\ \theta(v_3) &= 010, \\ \theta(v_4) &= 00.\end{aligned}$$

This particular mapping can be represented by a tree graph, as in Fig. 1. (For a formal discussion of tree graphs, see, for example, Berge, 1958.) The nodes represent choice points; a path down to the left from a node represents the selection of 1 from A and a path down to the right represents the selection of 0. Each word has a unique spelling indicated by a unique branch through the coding tree. When the end of a branch is reached and a full word has been spelled, the system returns to the top, ready to spell the next word.

In order to decode the message, of course, it is essential to maintain synchronization. For example, the string of words $v_4v_1v_4v_1v_1$ is spelled 0010011, but if the first letter of this spelling is lost it will be decoded as v_3v_2 . We use the symbol $\#$ at the beginning of a string of letters to indicate that it is known that this is the beginning of the total message; otherwise, a string of periods . . . is used.

At any particular point in a string of letters that spells some acceptable message there will be a fixed set of possible continuations that terminate

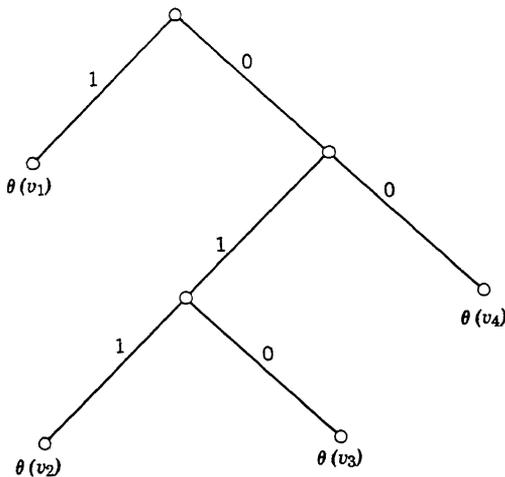


Fig. 1. Graph of coding tree for C_1 .

at the end of a word. Moreover, different initial strings may permit exactly the same continuations. In C_1 , for example, the two messages that begin #000... and #10... can be terminated by ...0#, ...10#, ...11#, or by one of those followed by other words. We say that the relation R holds between any two initial strings that permit the same continuation. We see immediately that R must be reflexive, symmetrical, and transitive and so is an equivalence relation; two initial strings of characters that permit exactly the same set of continuations are referred to as equivalent on the right. In terms of this relation, we can define the important concept of a *state*: *the set of all strings equivalent on the right constitutes a state of the coding system.*

The state of a coding system constitutes its memory of what has already occurred. Each time a letter is added to the encoded string the system can move to a new state. In C_1 there are three states: (1) S_0 when a complete word has just been spelled, (2) S_1 after #0..., and (3) S_2 after #01.... These correspond to the three nonterminal nodes in the tree graph of Fig. 1.

Following Schützenberger (1956), we can summarize the state transitions by matrices, one for each string of letters. Let rows represent the state after n letters, and let the columns represent the state after $n + 1$ letters. If a transition is possible, enter 1 in that cell; otherwise, 0. For C_1 we require two matrices, one to represent the effect of adding the letter 0, the other to represent the effect of adding the letter 1. (In general, this corresponds to a partition of the coding tree into subgraphs, one for each letter in A). To each string x associate the matrix M_x with elements m_{ij} giving the number of paths between states S_i and S_j when the string x occurs in the coded messages. For C_1 the matrices associated with the elementary strings 0 and 1 are

$$M_0 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

For a longer string the matrix is the ordered product of the matrices for the letters in the string. The product matrix $\mathcal{A}\mathcal{B}$ is interpreted in the following fashion: from S_i the system moves to S_j according to the transitions permitted by \mathcal{A} , then moves from S_j to S_k according to the transitions permitted by \mathcal{B} . The number of distinct paths from S_i to S_j to S_k is $a_{ij}b_{jk}$. The total number of paths from S_i to S_k , summing over all intervening states S_j , is $\sum_j a_{ij}b_{jk}$, the row-by-column product that gives the elements of $\mathcal{A}\mathcal{B}$. In case a particular letter cannot occur in a given state, the row of its matrix corresponding to that state will consist entirely of

zeros. Any matrix corresponding to a string in A that does not spell any part of a string in V will be a zero matrix. In general, the matrices need not possess inverses; they do not form a group, but they are adequate to provide an isomorphism with the elements of a semigroup.

If the function mapping V into A is not bi-unique, entries greater than unity will occur in the matrices or their products, signifying that a single string of n letters must be the spelling for more than one string of words. When this occurs, the received message is ambiguous and cannot be understood even though it is received without noise or distortion. There is no way of knowing which of the alternative strings of words was intended.

Because there is no phonetic boundary marker between successive words in the normal flow of speech, such ambiguities can easily arise in natural languages. Miller (1958) gives the following example in English:

The good candy came anyway.

The good can decay many ways.

The string of phonetic elements can be pronounced in a way that is relatively neutral, so that the listener has an experience roughly comparable to the visual phenomenon of reversible perspective. Ambiguities of segmentation may be even commoner in French; for example, the following couplet is well known as an instance of complete rhyme:

Gal, amant de la Reine, alla (tour magnanime),

Galamment de l'arène à la Tour Magne, à Nîmes.

Consideration of examples of this type indicates that there is far more to the production or perception of speech than merely the production or identification of successive phonetic qualities and that different kinds of information processing must go on at several levels of organization. These problems are defined more adequately for natural languages in Sec. 6.

Difficulties of segmentation can be avoided, of course. Consideration of how to avoid them leads to a simple classification of the various types of codes (Schützenberger, personal communication). *General codes* include any codes that always give a different spelling for every different string of words. A special subset of the general codes are the *tree codes* whose spelling rules can be represented graphically, as in Fig. 1, with the spelling of each word terminating at the end of a separate branch. Every tree code must be of one of two types: the *left tree codes* are those in which no spelling of any word forms an initial segment (left segment) of the spelling of any other word; the *right tree codes* are, similarly, those

in which no word forms a terminal segment (right segment) of any other word (right tree codes can be formed by reversing the spelling of all words in a left tree code). A special case is the class of codes that are both left and right tree codes simultaneously; Schützenberger has called them *anagrammatic codes*. The simplest subset of anagrammatic codes are the *uniform codes*, in which every word is spelled by the same number of letters and scansion into words is achieved at the receiver simply by counting. Uniform codes are frequently used in engineering applications of coding theory, but they have little relevance for the description of natural languages.

Another important set of codes are the self-synchronizing codes. In a self-synchronizing code, if noise or error causes some particular word boundary to be misplaced, the error will not perpetuate itself indefinitely; within a finite time the error will be absorbed and the correct synchrony will be reestablished. (C_1 is an example of a self-synchronizing code; uniform codes are not.) In a self-synchronizing tree code the word boundaries are always marked by the occurrence of a particular string of letters. If the particular string is thought of as terminating the spelling of every word, we have a left-synchronizing tree code. When the particular string consists of a single letter (which cannot then be used anywhere else), we have what has been called a *natural code*. In written language the

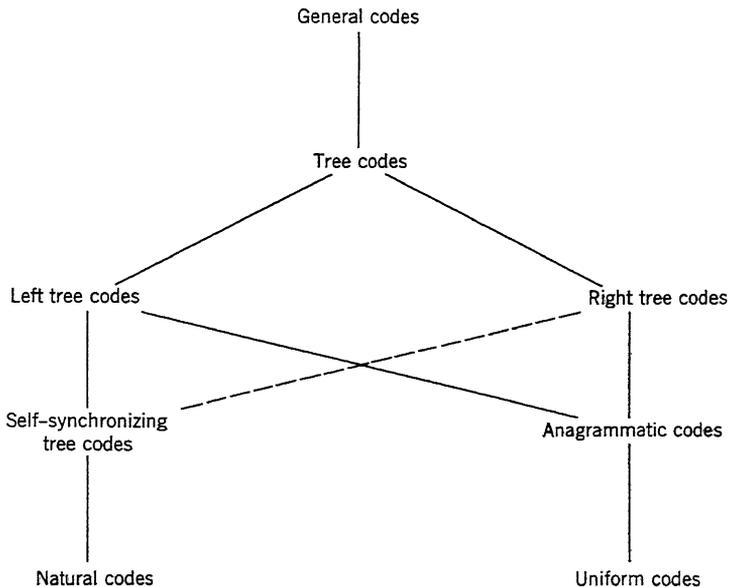


Fig. 2. Classification of coding systems.

space between words keeps the receiver synchronized. In spoken language the process is much more complex; discussion of how strings of words (morphemes) are mapped into strings of phonetic representation is postponed until Sec. 6.

In order to be certain that a particular mapping is in fact a code, it is common engineering practice to inspect it to see if it is a left tree code—to see that no spelling of any word forms an initial segment of the spelling of any other word. It is possible, however, to have general codes that are not tree codes. Schützenberger (1956) offers the following code C_2 as the simplest, nontrivial example:

$$A = \{0, 1\}, \quad V = \{v_1, \dots, v_5\},$$

and

$$\theta(v_1) = 00$$

$$\theta(v_2) = 001$$

$$\theta(v_3) = 011$$

$$\theta(v_4) = 01$$

$$\theta(v_5) = 11$$

Note that $\theta(v_1)$ is an initial segment of $\theta(v_2)$, so it is not a left tree code; and $\theta(v_4)$ is a terminal segment of $\theta(v_2)$, so it is not a right tree code.

There is an understandable desire, when constructing artificial codes, to keep the coded words as short as possible. In this connection an interesting inequality can easily be shown to hold for tree codes (Kraft, 1949). Suppose we are given a vocabulary $V = \{v_1, \dots, v_n\}$, an alphabet $A = \{a_1, \dots, a_D\}$, and a mapping θ that constitutes a left tree code. Let c_i be the length of $\theta(v_i)$. Then

$$\sum_{i=1}^n D^{-c_i} \leq 1. \quad (4)$$

This inequality can be established as follows: let w_j be the number of coded words of exactly length j . Then, since θ is a tree code, we have

$$w_1 \leq D,$$

$$w_2 \leq (D - w_1)D = D^2 - w_1D,$$

$$w_3 \leq [(D^2 - w_1D) - w_2]D = D^3 - w_1D^2 - w_2D,$$

.....

$$w_n \leq D^n - w_1D^{n-1} - w_2D^{n-2} - \dots - w_{n-1}D.$$

Dividing by D^n gives

$$\sum_{j=1}^n w_j D^{-j} \leq 1.$$

But if $n \geq c_i$ for all i , then this summation will be taken over all the coded words, which gives the relation expressed in Eq. 4. The closer Eq. 4 comes to equality for any code, the closer that code will be to minimizing the

average length of its code words. (This inequality has also been shown to hold for nontree codes: see Mandelbrot, 1954; McMillan, 1956.)

Each tree code (including, of course, all natural codes) must have some function w_j giving the number of words of coded length j . The function is of some theoretical interest, since it summarizes in a particularly simple form considerable information about the structure of the coding tree.

Finally, it should be remarked that a code can be thought of as a simple kind of automaton (cf. Chapter 12) that accepts symbols in one alphabet and produces symbols in another according to predetermined rules that depend only on the input symbol and the internal state of the device. Some of the subtler difficulties in constructing good codes will not become apparent until we assign different probabilities to the various words (cf. Chapter 13).

3. SOME BASIC CONCEPTS OF LINGUISTICS

A central concept in these pages is that of a language, so a clear definition is essential. We consider a *language* L to be a set (finite or infinite) of *sentences*, each finite in length and constructed by concatenation out of a finite set of elements. This definition includes both natural languages and the artificial languages of logic and of computer-programming theory.

In order to specify a language precisely, we must state some principle that separates the sequences of atomic elements that form sentences from those that do not. We cannot make this distinction by mere listing, since in any interesting system there is no bound on sentence length. There are two ways open to us, then, to specify a language. Either we can try to develop an operational test of some sort that will distinguish sentences from nonsentences or we can attempt to construct a recursive procedure for enumerating the infinite list of sentences. The first of these approaches has rarely been attempted and is not within the domain of this survey. The second provides one aspect of what could naturally be called a *grammar* of the language in question. We confine ourselves here to the second approach—to the investigation of grammars.

In actual investigations of natural language a proposed operational test or a proposed grammar must, of course, meet certain empirical conditions. Before the construction of such a test or grammar can begin, there must be a finite class K_1 of sequences that can, with reasonable security, be assigned to the set of sentences as well as, presumably, a class K_2 of sequences that can, with reasonable security, be excluded from this class. The empirical significance of a proposed operational test or a proposed grammar will, in large part, be determined by their success in drawing

a distinction that separates K_1 from K_2 . The question of empirical adequacy, however, and the problems to which it gives rise are beyond the scope of this survey.

We limit ourselves, then, to the study of grammars: by a *grammar* we mean a set of rules that (in particular) recursively specify the sentences of a language. In general, each of the rules we need will be of the form

$$\phi_1, \dots, \phi_n \rightarrow \phi_{n+1}, \quad (5)$$

where each of the ϕ_i is a structure of some sort and where the relation \rightarrow is to be interpreted as expressing the fact that if our process of recursive specification generates the structures ϕ_1, \dots, ϕ_n then it also generates the structure ϕ_{n+1} .

The precise specification of the kinds of rules that can be permitted in a grammar is one of the major concerns of mathematical linguistics, and it is to this question that we shall turn our attention. Consider, for the moment, the following special case of rules of the form of Rule 5. Let $n = 1$ in Rule 5 and let ϕ_1 and ϕ_2 each be a sequence of symbols of a certain alphabet (or vocabulary). Thus, if we had a finite language consisting of the sentences $\sigma_1, \dots, \sigma_n$ and an abstract element S (representing "sentence"), which we take as an initial, given element, we could present the grammar

$$S \rightarrow \sigma_1; \dots; S \rightarrow \sigma_n; \quad (6)$$

which would, in this trivial case, be nothing but a sentence dictionary. More interestingly, consider the case of a grammar containing the two rules

$$S \rightarrow aS; S \rightarrow a. \quad (7)$$

This pair of rules can generate recursively any of the sentences $a, aa, aaa, aaaa, \dots$.² (Obviously the sentences can be put in one-to-one correspondence with the integers so that the language will be denumerably infinite.) To generate aaa , for example, We proceed as follows:

$$\begin{aligned} S & \text{ (the given, initial symbol),} \\ aS & \text{ (applying the first rewriting rule),} \\ aaS & \text{ (reapplying the first rewriting rule),} \\ aaa & \text{ (applying the second rewriting rule).} \end{aligned} \quad (8)$$

Below we shall study systems of rules of this and of somewhat richer forms.

To recapitulate, then, *the (finite) set of rules specifying a particular language constitutes the grammar of that language.* (This definition is made more precise in later sections.) An acceptable grammar must give a precise

² More precisely, we should say that we are now considering the case of rules of the form $\mu_1\phi_1\mu_2 \rightarrow \mu_1\phi_2\mu_2$, in which μ_1 and μ_2 are variables ranging over arbitrary, possibly null, strings, and ϕ_1 and ϕ_2 are constants. The variables can, obviously, be suppressed in the actual statement of the rules as long as we restrict ourselves to rules of this form.

specification of the (in general, infinite) list of sentences (strings of symbols) that are sentences of this language. As a matter of principle, a grammar must be finite. If we permit ourselves grammars with an unspecifiable set of rules, the entire problem of grammar construction disappears; we can simply adopt an infinite sentence dictionary. But that would be a completely meaningless proposal. Clearly, a grammar must have the status of a theory about those recurrent regularities that we call the syntactic structure of the language. To the extent that a grammar is formalized, it constitutes a mathematical theory of the syntactic structure of a particular natural language.

It should be obvious, however, that a grammar must do more than merely enumerate the sentences of a language (though, in actual fact, even this goal has never been approached). We require as well that the grammar assign to each sentence it generates a *structural description* that specifies the elements of which the sentence is constructed, their order, arrangement, and interrelations and whatever other grammatical information is needed to determine how the sentence is used and understood. A theory of grammar must, therefore, provide a mechanical way of determining, given a grammar G and a sentence s generated by G , what structural description is assigned to s by G . If we regard a grammar as a finitely specified function that enumerates a language as its range, we could regard linguistic theory as specifying a functional that associates with any pair (G, s) , in which G is a grammar and s a sentence, a structural description of s with respect to G ; and one of the primary tasks of linguistic theory, of course, would be to give a clear account of the notion of structural description.

This conception of grammar is recent and may be unfamiliar. Some artificial examples can clarify what is intended. Consider the following three artificial languages described by Chomsky (1956):

Language L_1 . L_1 contains the sentences $ab, aabb, aaabbb, \text{etc.}$; all sentences contain n occurrences of a , followed by n occurrences of b , and only these.

Language L_2 . L_2 contains $aa, bb, abba, baab, aabbaa, \text{etc.}$; all mirror image sentences consisting of a string, followed by the same string in reverse, and only these.

Language L_3 . L_3 contains $aa, bb, abab, baba, aabaab, \text{etc.}$; all sentences consisting of a string followed again by that identical string, and only these.

A grammar G_1 for L_1 may take the following form:

Given: S ,

$$\begin{aligned} F1: S &\rightarrow ab, \\ F2: S &\rightarrow aSb, \end{aligned} \tag{9}$$

where S is comparable to an axiom and $F1, 2$ are rules of formation by which admissible strings of symbols can be derived from the axiom. Derivations would proceed after the manner of Example 8. A derivation terminates whenever the grammar contains no rules for rewriting any of the symbols in the string.

In the same vein a grammar G_2 for L_2 might be the following:

Given: S ,

$$\begin{aligned} F1: S &\rightarrow aa, \\ F2: S &\rightarrow bb, \\ F3: S &\rightarrow aSa, \\ F4: S &\rightarrow bSb. \end{aligned} \tag{10}$$

An interesting and important feature of both L_1 and L_2 is that new constructions can be embedded inside of old ones. In $aabbaa$, for example, there is in L_2 a relation of dependency between the first and sixth elements; nested inside it is another dependency between the second and fifth elements; inside that, in turn, is a relation between the third and fourth elements. As G_1 and G_2 are stated, of course, there is no upper limit to the number of embeddings that are possible in an admissible string.

There can be little doubt that natural languages permit this kind of parenthetical embedding and that their grammars must be able to generate such sequences. For example, the English sentence (*the rat (the cat (the dog chased) killed) ate the malt*) is surely confusing and improbable but it is perfectly grammatical and has a clear and unambiguous meaning. To illustrate more fully the complexities that must in principle be accounted for by a real grammar of a natural language, consider this English sentence:

Anyone₁ who feels that if₂ so-many₃ more₄ students₅ whom we₆ haven't₆ actually admitted are₅ sitting in on the course than₄ ones we have that₃ the room had to be changed, then₂ (11) probably auditors will have to be excluded, is₁ likely to agree that the curriculum needs revision.

There are dependencies in Example 11 between words with the same subscript; the result is a system of nested dependencies, as in L_2 . Furthermore, to complicate the picture further, there are dependencies that cross those indicated, for example, between *students* and *ones*, between *haven't . . . admitted* and *have* 10 words later (with an understood occurrence of *admitted* deleted). Note, incidentally, that we can have nested dependencies, as in Example 11, in which a variety of constructions are involved; in the special case in which the *same* nested construction occurs more than once, we speak of *self-embedding*.

Of course, we can safely predict that Example 11 will never be produced, except as an example, just as we can, with equal security, predict that such perfectly well-formed sentences as *birds eat*, *black crows are black*, *black crows are white*, *Tuesday follows Monday*, etc., will never occur in normal adult discourse. Like other sentences that are too obviously true, too obviously false, too complex, too inelegant, or that fail in innumerable other ways to be of any use for ordinary human affairs, they are not used. Nevertheless, Example 11 is a perfectly well-formed sentence with a clear and unambiguous meaning, and a grammar of English must be able to account for it if the grammar is to have any psychological relevance.

A grammar that will generate the language L_3 must be quite a bit more complex than that for L_2 , if we restrict ourselves to rules of the form $\phi \rightarrow \psi$, where ϕ and ψ are strings, as proposed (see Chapter 12, Sec. 3, for further discussion). We can, however, construct a simple grammar for this language if we allow more powerful rules. Let us establish the convention that the symbol x stands for any string consisting of just the symbols a , b and let us add the symbol $\#$ as a boundary symbol marking the beginning and the end of a sentence. Then we can propose the following grammar for L_3 :

Given: $\#S\#$,

$$F1: S \rightarrow aS,$$

$$F2: S \rightarrow bS,$$

(12)

$$F3: \#xS\# \rightarrow \#xx\#.$$

Rules $F1$ and $F2$ permit the generation of an arbitrary string of a 's and b 's: F_3 repeats any such string. Clearly, the language generated is exactly L_3 (with boundary symbols on each sentence).

It is important to note, however, that $F3$ has a different character from the other rules, since it necessitates an analysis of the total string to which it applies; this analysis goes well beyond what is allowed by the restricted form of rule we considered first.

If we adopt richer and more powerful rules such as $F3$, then we can often effect a great simplification in the statement of a grammar; that is to say, we can make use of generalizations regarding linguistic structure that would otherwise be simply unstatable. There can be no objection to permitting such rules, and we shall give some attention to their formulation and general features in Sec. 5. Since a grammar is a theory of a language and simplicity and generality are primary goals of any theory construction, we shall naturally try to formulate the theory of linguistic structure to accommodate rules that permit the formulation of deeper generalizations. Nevertheless, the question whether it is possible in principle to generate natural languages with rules of a restricted form—such as $F1$ to $F4$ in

Example 10—retains a certain interest. An answer to this question (either positive or negative) would reveal certain general structural properties of natural language systems that might be of considerable interest.

The dependency system illustrated in L_3 is quite different from that of L_2 . Thus in the string *baabaa* of L_3 the dependencies are not nested, as they are in the string *aabbaa* of L_2 ; instead the fourth symbol depends on the first, the fifth on the second, and the sixth on the third. Dependency systems of this sort are also to be found in natural language (see Chapter 12, Sec. 4.2, for some examples) and thus must also be accommodated by an adequate theory of grammar. The artificial languages L_2 and L_3 , then, illustrate real features of natural language, and we shall see later that the illustrated features are critical for determining the adequacy of various types of models for grammar.

In order to illustrate briefly how these considerations apply to a natural language, consider the following small fragment of English grammar:

Given: # S #,

$$\begin{aligned}
 F1: S &\rightarrow AB, \\
 F2: A &\rightarrow CD, \\
 F3: B &\rightarrow EA, \\
 F4: C &\rightarrow a, \textit{the}, \textit{another}, \dots, \\
 F5: D &\rightarrow \textit{ball}, \textit{boy}, \textit{girl}, \dots, \\
 F6: E &\rightarrow \textit{hit}, \textit{struck}, \dots
 \end{aligned}
 \tag{13}$$

$F4$ to 6 are groups of rules, in effect, since they offer several alternative ways of rewriting C , D , and E . (Ordinarily, we refer to A as a noun phrase and to B as a verb phrase, etc., but these familiar names are not essential to the formal structure of the grammar, although they may play an important role in general linguistic theory.) In any real grammar, of course, there must also be phonological rules that code the terminal strings into phonetic representations. For simplicity, however, we shall postpone any consideration of the phonological component of grammar until Sec. 6.

With this bit of grammar, we can generate such terminal strings as # *the boy hit the girl* #. In this simple case all of the terminal strings have the same phrase structure, which we can indicate by bracketing with labeled parentheses,

$$\# (S_{(A(Cthe)_C (Dboy)_D)_A (B(Ehit)_E (A(Cthe)_C (Dgirl)_D)_B)_S} \#,$$

or, equivalently, with a labeled tree graph of the kind shown in Fig. 3. We assume that such a tree graph must be a part of the structural description of any sentence; we refer to it as a *phrase-marker* (P -marker). A grammar must, for adequacy, provide a P -marker for each sentence.

Each *P*-marker contains, as the successive labels of its final nodes, the record of the vocabulary elements (e.g., words) of which the sentence is actually composed. Two *P*-markers are the same only if they have exactly the same branching structure and exactly the same labels on corresponding nodes. Note that the tree graph of a *P*-marker, unlike the coding trees of Sec. 2, must have a specified ordering of its branches from left to right, corresponding to the order of the elements in the string.

The function of *P*-markers is well illustrated by the sentences displayed in Examples 1*a* and 1*b*. A grammar that merely generated the given string of words would not be able to characterize the grammatical differences between those two sentences.

Linguists generally refer to any word (morpheme) or sequence that functions as a unit in some larger construction as a *constituent*. In the sentence whose *P*-marker is shown in Fig. 3, *girl*, *the girl*, and *hit the girl* are all constituents, but *hit the* is not a constituent. Constituents can be traced back to nodes in the tree; if the node is labeled *A*, then the constituent is said to be of type *A*. The *immediate constituents* of any construction are the constituents of which that construction is directly formed. For example, *the boy* and *hit the girl* are immediate constituents of the sentence in Fig. 3; *hit* and *the girl* are immediate constituents of the verb phrase *B*; etc. We will not be satisfied with any formal characterization of grammar that does not provide at least a structural description in terms of immediate constituents.

A grammar, then, must provide a *P*-marker for each of an infinite class of sentences, in such a way that each *P*-marker can be represented graphically as a labeled tree with labeled lines (i.e., the nodes are labeled and lines

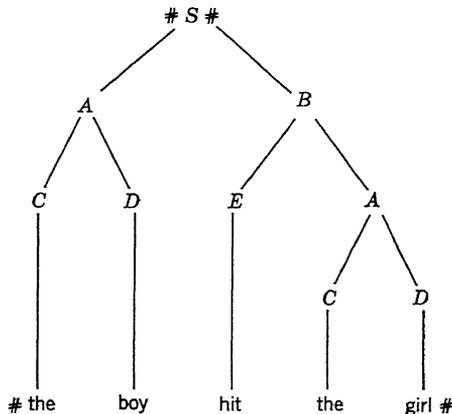


Fig. 3. A graphical representation (*P*-marker) of the derivation of a grammatical sentence.

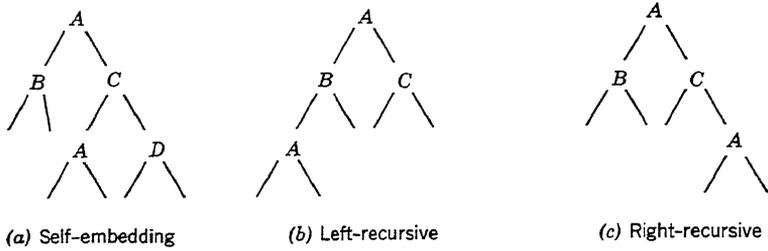


Fig. 4. Illustrating types of recursive elements.

branching from a single node are distinguished with respect to order). By a *branch* of a tree we mean a sequence of lines, each connected to the preceding one. For example, one of the branches in Fig. 3 is the sequence $((S-B), (B-E), (E-hit))$; another is $((A-D), (D-girl))$, etc. Since the tree graph represents proper parenthesization, its branches do not cross. (To make these informal remarks precise in the intended and obvious sense, we would have to distinguish occurrences of the same label.) The symbols that label the nodes of the tree are those that appear in the grammatical rules. Since the rules are finite and there are in any interesting case an infinite number of P -markers, there must be some symbols of the vocabulary of the grammar that can occur indefinitely often in P -markers. Furthermore, there will be branches that contain some of these symbols more than n times for any fixed n . Given a set of P -markers, we say that a symbol of the vocabulary is a *recursive element* if for every n there is a P -marker with a branch in which this symbol occurs more than n times as a label of a node.

We distinguish three different kinds of recursive elements that are of particular importance in later developments. We say that a recursive element A is *self-embedding* if it occurs in a configuration such as that of Fig. 4a; *left-recursive* if it occurs in a configuration such as that of Fig. 4b; *right-recursive* if it occurs in a configuration such as that of Fig. 4c.

Thus, if A is a left-recursive element, it dominates (i.e., is a node from which can be derived) a tree configuration that contains A somewhere on its leftmost branch; if it is a right recursive element it dominates a tree configuration that contains A somewhere on its rightmost branch; if it is a self-embedding element, it dominates a tree configuration that contains A somewhere on an inner branch. It is not difficult to make these definitions perfectly precise, but we will do so only for particular cases of special interest.

The study of recursive generative processes (such as, in particular, generative grammars) has grown out of investigations of the foundations of mathematics and the theory of proof. For a recent survey of this field,

see Davis (1958); for a shorter and more informal introduction see, for example, Rogers (1959) or Trachtenbrot (1962). We return to more general considerations involving recursive generation and its properties in Chapter 12.

In this section we have mentioned a few basic properties of grammars and have given some examples of generative devices that might be regarded as grammars. In the rest of this chapter we try to formulate the characteristics of grammars in more detail. In Sec. 4 we consider more carefully grammars meeting the condition that, in each rule of the form (5), $n = 1$ and each structure is a string, as in Examples 9, 10, and 13. (In Chapter 12 we study some of the properties of these systems.) In Sec. 5 we turn our attention to a richer class of grammars, akin to that suggested for L_3 in Example 12, that do not impose the restrictive conditions that in each rule of the form of Example 5 the structures are limited to strings or that $n = 1$. Finally, in Sec. 6 we indicate briefly some of the properties of the phonological component of a grammar that converts the output of the recursive generative procedure into a sequence of sounds.

We have described a generative grammar G as a device that enumerates a certain subset L of the set Σ of strings in a fixed vocabulary V and that assigns structural descriptions to the members of the set $L(G)$, which we call the *language generated by G* . From the point of view of the intended application of the models that we are considering, it would be more realistic to regard G as a device that assigns a structural description to each string of Σ , where the structural description of a string x indicates, in particular, the respects in which x deviates, if at all, from well-formedness, as defined by G . Instead of partitioning Σ into the two subsets $L(G)$ (well-formed sentences) and $\overline{L(G)}$ (nongrammatical strings), G would now distinguish in Σ a class L_1 of perfectly well-formed sentences and would partially order all strings in Σ in terms of *degree of grammaticalness*. We could say, then, that L_1 is the language generated by G ; but we could attempt to account for the fact that strings not in L_1 can still often be understood, even understood unambiguously by native speakers, in terms of the structural descriptions assigned to these strings. A string of $\Sigma \cap \overline{L_1}$ can be understood by imposing on it an interpretation, guided by its analogies and similarities to sentences of L_1 . We could attempt to relate ease and uniformity of interpretation of a sentence to degree of grammaticalness, given a precise definition of this notion. Deviation from grammatical regularities is a common literary or quasi-literary device, and it need not produce unintelligibility—in fact, as has often been remarked, it can provide a certain richness and compression.

The problems of defining degree of grammaticalness, relating it to interpretation of deviant utterances, and constructing grammars that

assign degrees of grammaticalness other than zero and one to utterances are all interesting and important. Various aspects of these questions are discussed in Chomsky (1955, 1961b), in Ziff (1960a, 1960b, 1961) and in Katz (1963). We consider this matter further in Chapter 13. Here, and in Chapter 12, however, we limit our attention to the special case of grammars that partition all strings into the two categories grammatical and ungrammatical and that do not go on to establish a hierarchy of grammaticalness.

4. A SIMPLE CLASS OF GENERATIVE GRAMMARS

We consider here a simple class of grammars that can be called *constituent-structure grammars* and introduce some of the more important notations that are used in this and the next two chapters. In this section we regard a grammar G ,

$$G = [V, \overset{\frown}{\smile}, \rightarrow, V_T, S, \#],$$

as a system of concatenation meeting the following conditions:

1. V is a finite set of symbols called the *vocabulary*. The strings of symbols of this vocabulary are formed by concatenation; $\overset{\frown}{\smile}$ is an associative and noncommutative binary operation on strings formed on the vocabulary V . We suppress $\overset{\frown}{\smile}$ where no confusion can result.

2. $V_T \subset V$. V_T we call the *terminal vocabulary*. The relative complement of V_T with respect to V we call the *nonterminal* or *auxiliary* vocabulary and designate it by V_N .

3. \rightarrow is a finite, two-place, irreflexive and asymmetric relation defined on certain strings on V and read "is rewritten as." The pairs (ϕ, ψ) such that $\phi \rightarrow \psi$ are called the (*grammatical*) *rules* of G .

4. Where $A \in V$, $A \in V_N$ if and only if there are strings ϕ, ψ, ω such that $\phi A \psi \rightarrow \phi \omega \psi$. $\# \in V_T$; $S \in V_N$; $e \in V_T$; where $\#$ is the *boundary symbol*, S is the *initial symbol* that can be read *sentence*, and e is the identity element with the property that for each string ϕ , $e\phi = \phi = \phi e$.

We define the following additional notions:

5. A sequence of strings $D = (\phi_1, \dots, \phi_n)$ ($n \geq 1$) is a ϕ -*derivation* of ψ if and only if

(a) $\phi = \phi_1$; $\psi = \phi_n$

(b) for each $i < n$ there are strings $\psi_1, \psi_2, \chi, \omega$ such that $\chi \rightarrow \omega$, $\phi_i = \psi_1 \chi \psi_2$, and $\phi_{i+1} = \psi_1 \omega \psi_2$.

The set of derivations is thus completely determined by the finitely specified relation \rightarrow , that is, by the finite set of grammatical rules. Where there is a

ϕ -derivation of ψ , we say that ϕ dominates ψ and write $\phi \Rightarrow \psi$; \Rightarrow is thus a reflexive and transitive relation.

6. A ϕ -derivation D of ψ is *terminated* if ψ is a string on V_T and D is not the proper initial subsequence of any derivation (note that these conditions are independent).

7. ψ is a *terminal string* of G if there is a terminated $\#S\#$ -derivation of $\#\psi\#$; that is, a terminal string is the last line of a terminated derivation beginning with the initial string $\#S\#$.

8. The *terminal language generated by G* is the set of terminal strings of G .

9. Two grammars G and G^* are (*weakly*) *equivalent* if they generate the same terminal language. A stronger equivalence is discussed later.

We want the boundary symbol $\#$ to meet the condition that if (ϕ_1, \dots, ϕ_n) is a $\#S\#$ -derivation then, for each i , $\phi_i = \#\psi_i\#$, where ψ_i does not contain $\#$. To guarantee that this will be the case, we impose on the rules of the grammar (on the relation \rightarrow) the following additional condition:

10. If $\alpha_1 \dots \alpha_m \rightarrow \beta_1 \dots \beta_n$ is a rule of a grammar (where $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_n \in V - \{e\}$), then

- (a) $\beta_i \neq \#$ for $1 < i < n$;
- (b) $\beta_1 = \#$ if and only if $\alpha_1 = \#$;
- (c) $\beta_n = \#$ if and only if $\alpha_m = \#$.

The conditions so far laid down do not show how the grammar may provide a P -marker for every terminal string. This is a further requirement that we will have to keep in mind as we consider additional, more restrictive conditions.

See Čulík (1962) for a critique of an earlier formulation of these notions in Chomsky (1959).

Recalling now our discussion of recursive elements in Sec. 3, we can observe the following, where $A \in V_N$:

1. If there are nonnull strings ϕ, ψ such that $A \Rightarrow \phi A \psi$, then A is a self-embedding element.
2. If there is a nonnull ϕ such that $A \Rightarrow A \phi$, then A is a left-recursive element.
3. If there is a nonnull ϕ such that $A \Rightarrow \phi A$, then A is a right-recursive element.
4. If A is a nonrecursive element, then there are no strings ϕ, ψ such that $A \Rightarrow \phi A \psi$.

The converses of these statements are not necessarily true for grammars of the form we have so far described, although they will be true in the case that we now discuss. Thus suppose the grammar G contains the rules $S \rightarrow BAC, BA \rightarrow BBAC$, but no rule $A \rightarrow \chi$ for any χ . Then there are no strings ϕ, ψ such that $A \Rightarrow \phi A \psi$, but A is recursive (in fact, self-embedding).

Our discussion of these systems requires us to distinguish terminal from nonterminal elements and atomic elements from strings formed by concatenation. It will help to keep these distinctions clear if we adopt the following important convention:

| Type | Single atomic elements | Strings of elements |
|-------------|--------------------------------|---------------------------|
| Nonterminal | A, B, C, \dots | X, Y, Z, \dots |
| Terminal | a, b, c, \dots | x, y, z, \dots |
| Arbitrary | $\alpha, \beta, \gamma, \dots$ | ϕ, χ, ψ, \dots |

This convention has in fact already been followed; it is followed without special comment throughout this and the next two chapters.

We would now like to add conditions to guarantee that a P -marker for a terminal string can be recovered uniquely from its derivation. There is no natural way to do this for the case of grammars of the sort so far discussed. For example, if we have a grammar with the rules

$$S \rightarrow AB; \quad AB \rightarrow cde, \quad (15)$$

there is no way to determine whether in the string cde the segment cd is a phrase of the type A (dominated by A in the phrase marker) or whether de is a phrase of the type B (dominated by B in the phrase marker). We can achieve the desired result most naturally by requiring that in each rule of the grammar only a single symbol can be rewritten. Thus grammars can contain rules of either of the forms in Rule 16a or 16b:

$$A \rightarrow \omega \quad (16a)$$

$$\phi A \psi \rightarrow \phi \omega \psi \text{ (equivalently: } A \rightarrow \omega \text{ in the context } \phi\text{--}\psi\text{)}. \quad (16b)$$

A grammar containing only rules of the type in Rule 16b is called a *context-sensitive* grammar. A grammar containing only rules of the type in Rule 16a is called a *context-free grammar*, and the language it generates, a *context-free language*. (Note, incidentally, that it is the *rules* that are sensitive to or free of their context, not the elements in the terminal string.) In either case, if ϕ is a line of a derivation and ψ is the line immediately succeeding it, then there are unique³ strings $\phi_1, \phi_2, \alpha, \omega$ such that $\phi = \phi_1 \alpha \phi_2$ and $\psi = \phi_1 \omega \phi_2$; and we say that ω is a string of type α (i.e., it is ³Actually, to guarantee uniqueness certain additional conditions must be satisfied by the set of rules; in particular, we must exclude the possibility of such a sequence of lines as AB, ACB , and so on. We will assume without further comment that such conditions are met. They can, in fact, always be met without restricting further the class of languages that can be generated, although, of course, they affect in some measure the class of systems of P -markers that can be generated. In the case of context-sensitive grammars, this further condition is by no means innocuous, as we shall see in Chapter 12, Sec. 3.

dominated by a node labeled α in the tree representing the associated P -marker). In the case of context-free grammars, the converse of each assertion of Observation 14 is true, and we have precise definitions of the various kinds of recursiveness in terms of \Rightarrow . In fact, we shall study the various types of recursive elements only in the case of context-free grammars.

The principal function of rules of the type in Rule 16*b* is to permit the statement of selectional restrictions on the choice of elements. Thus among subject-verb-object sentences we find, for example, *The fact that the case was dismissed doesn't surprise me, Congress enacted a new law, The men consider John a dictator, John felt remorse*, but we do not find the sequences formed by interchange of subject and object: *I don't surprise the fact that the case was dismissed, A new law enacted Congress, John considers the men a dictator, Remorse felt John*. Native speakers of English recognize that the first four are perfectly natural, but the second four, if intelligible at all, require that some interpretation be imposed on them by analogy to well-formed sentences. They are, in the sense described at the close of Sec. 3, of lower, if not zero, degree of grammaticalness. A grammar that did not make this distinction would clearly be deficient; it can be made most naturally and economically by introducing context-sensitive rules to provide specific selectional restrictions on the choice of subject, verb, and object.

A theory of grammar must, ideally, contain a specification of the class of possible grammars, the class of possible sentences, and the class of possible structural descriptions; and it must provide a general method for assigning one or more structural descriptions to each sentence, given a grammar (that is, it must be sufficiently explicit to determine what each grammar states about each of the possible sentences—cf. Chomsky, 1961a, for discussion). To establish the theory of *constituent-structure grammar* finally, we fix the vocabularies V_N and V_T as given disjoint finite sets, meeting the conditions stated. A grammar, then, is simply a finite relation on strings in $V = V_N \cup V_T$ meeting these conditions. Note that the problem of fixing V_T , in the case of natural language, is essentially the problem of constructing a universal phonetic theory (including, in particular, a universal phonetic alphabet and laws determining universal constraints on distribution of its segments). For more on this topic, see Sec. 6 and the references cited there. In addition, we must set a bound on morpheme length (so that V_T is finite) and establish a set of “grammatical morphemes” (e.g., tenses, aspects, and numbers). This latter problem, along with that of giving a substantive interpretation to the members of the set V_N , is the classical problem of “universal grammar,” namely, giving a language-independent characterization of the set of “grammatical

categories" that can function in some natural language, a characterization that will no doubt ultimately involve both formal considerations involving grammatical structure and considerations of an absolute semantic nature. This is a problem that has not been popular in recent decades, but there is no reason to regard it as beyond the bounds of serious study. On the contrary, it remains a significant and basic question for general linguistics.

We return to the study of the various kinds of constituent-structure grammars in Chapter 12.

5. TRANSFORMATIONAL GRAMMARS

We stipulated in Sec. 3 that rules of grammar are each of the form

$$\phi_1, \dots, \phi_n \rightarrow \psi, \quad (17)$$

where each of $\phi_1, \dots, \phi_n, \psi$ is a structure of some sort, and the symbol \rightarrow indicates that, if ϕ_1, \dots, ϕ_n have been generated in the course of a derivation, then ψ can also be generated as an additional step. In Sec. 4 we considered a simple case of grammars, called *constituent-structure grammars*, with rules representing a very special case of (17), namely, the case in which $n = 1$ and in which, in addition, each of the structures ϕ_1 and ψ is simply a string of symbols. (We also required that the grammar meet the additional condition stated in Rule 16.) In this section we consider a class of grammars containing two syntactic subcomponents: a *constituent-structure component* consisting of rules meeting the restrictive conditions discussed in Sec. 4, and several others, and a *transformational component* containing rules of the form of Rule 17, in which n may be greater than 1 and in which each of the structures ϕ_1, \dots, ϕ_n , and ψ is not a string but rather a phrase-marker (cf. p. 288). Such grammars we call *transformational grammars*.

The plausibility of this generalization to transformational grammars is suggested by the obvious psychological fact that some pairs of sentences seem to be grammatically closely related. Why, for example, is the sentence *John threw the ball* felt to be such a close relative of the sentence *The ball was thrown by John*? It cannot be because they mean the same thing, for a similar kind of closeness can be felt between *John threw the ball* and *Who threw the ball* or *Did John throw the ball*, which are not synonymous. Nor can it be a matter of some simple formal relation (in linguistic parlance, a "co-occurrence relation") between the n -tuples of words that fill corresponding positions in the paired sentences, as can be seen, for example by the fact that *The old man met the young woman* and *The old woman met the young man* are obviously not related in the same way in which active and passive are structurally related, although the distinction between this

case and the active-passive case cannot be given in any general "distributional" terms (for discussion, see Chomsky, 1962b). In a constituent-structure grammar all of these sentences would have to be generated more or less independently and thus would bear no obvious relation to one another. But in a transformational grammar these sentences can be related directly by simple rules of transformation.

5.1 Some Shortcomings of Constituent-Structure Grammars

The basic reasons for rejecting constituent-structure grammars in favor of transformational grammars in the theory of linguistic structure have to do with the impossibility of stating many significant generalizations and simplifications of the rules of sentence formation within the narrower framework. These considerations go beyond the bounds of the present survey. For discussion, see Chomsky (1955, Chapters 7 to 9; 1957, Chapters 6 to 7; 1962b), Lees (1957; 1960), and Postal (1963); also Chapter 12, Sec. 4.2. However, some of the respects in which constituent-structure grammars are formally defective are relatively easy to describe and involve some important general considerations that are often overlooked in considering the adequacy of grammars.

A grammar must generate a language regarded as an infinite set of sentences. It must also associate with each of these sentences a structural description; it must, in other words, generate an infinite set of structural descriptions, each of which uniquely determines a particular sentence (though not conversely). Hence there are two kinds of equivalence that we can consider when evaluating the generative capacity of grammars and of classes of grammars. Two grammars will be called *weakly equivalent* if they generate the same language; they will be called *strongly equivalent* if they generate the same set of structural descriptions. In this chapter, and again in Chapter 12, we consider mainly weak equivalence because it is more accessible to study and has been investigated in more detail, but strong equivalence is, ultimately, by far the more interesting notion. Similarly, we have no interest, ultimately, in grammars that generate a natural language correctly but fail to generate the correct set of structural descriptions.

The question whether natural languages, regarded as sets of sentences, can be generated by one or another type of constituent-structure grammar is an interesting and important one. However, there is no doubt that the set of structural descriptions associated with, let us say, English cannot in principle be generated by a constituent-structure grammar, however complex its rules may be. The problem is that a constituent-structure grammar

necessarily imposes too rich an analysis on sentences because of features inherent in the way in which *P*-markers are defined for such grammars. The germ of the problem can be seen in the case of such sentences as

Why has John always been such an easy man to please? (18)

The whole is a sentence; the last several words constitute a noun phrase; the words can be assigned to syntactic categories. But there is no reason to assign any phrase structure beyond that. To assign just this amount of phrase structure, a constituent-structure grammar would have to contain these rules:

$$\begin{aligned}
 S &\rightarrow \text{why has } NP \text{ always been } NP \\
 NP &\rightarrow \text{John} \\
 NP &\rightarrow \text{such an easy } N \text{ to } V_{\text{tr}(\text{ansitive})} \\
 N &\rightarrow \text{man} \\
 V_{\text{tr}} &\rightarrow \text{please,}
 \end{aligned}
 \tag{19}$$

or something of the sort. It is obvious that a collection of rules such as this is quite absurd and leaves unstated all sorts of structural regularities. (Furthermore, the associated *P*-marker is defective in that it does not indicate that *man* is grammatically related to *please* as it is, for example, in *it pleases the man*; that is to say, that the verb-object relation holds for this pair. Cf. Sec. 2.2 of Chapter 13.) In particular, much of the point of constituent-structure grammar is lost if we have to give rules that analyze certain phrases into six immediate constituents, as in Example 19.

This difficulty changes from a serious complication to an inadequacy in principle when we consider the case of true coordination, as, for example, in such sentences as

The man was old, tired, tall, . . . , and friendly. (20)

In order to generate such strings, a constituent-structure grammar must either impose some arbitrary structure (e.g., using a right-recursive rule), in which case an incorrect structural description is generated, or it must contain an infinite number of rules. Clearly, in the case of true coordination, by the very meaning of this term, no internal structure should be assigned at all within the sequence of coordinate items.

We might try to meet this problem by extending the notion of constituent-structure grammar to permit such rule schemata as

$$\text{Predicate} \rightarrow \text{Adj}^n \text{ and Adj} \quad (n \geq 1). \tag{21}$$

Aside from the many difficulties involved in formulating this notion so that descriptive adequacy may be maintained, it is, of course, beside the point in the kind of difficulty that arises in Example 19. In general, for each

particular kind of difficulty that arises in constituent-structure grammars, it is possible to devise some *ad hoc* adjustment that might circumvent it. Much to be preferred, obviously, would be a conceptual revision that would succeed in avoiding the mass of these difficulties in a uniform way, while allowing the simple constituent-structure grammar to operate without essential alteration for the class of cases for which it is adequate and which initially motivated its development. As far as we know, the theory of transformational grammar is unique in holding out any hope that this end can be achieved.

In a transformational grammar the set of rules meets the following conditions. We have, first, a *constituent-structure component* consisting of a sequence of rules of the form $\phi A \psi \rightarrow \phi \omega \psi$, where A is a single symbol, ω is nonnull, and ϕ , ψ are possibly null strings. This constituent-structure component of the grammar will generate a finite number of *C-terminal strings*, to each of which we can associate, as before, a labeled tree—a *P-marker*—representing its constituent structure. We now add to the grammar a set of operations called *grammatical transformations*, each of which maps an n -tuple of *P*-markers ($n \geq 1$) into a new *P*-marker. The recursive property of the grammar can be attributed entirely to these transformations. Among these transformations, some are obligatory—they must be applied in every derivation (furthermore, some transformations are obligatory relative to others, i.e., they must be applied if the others are applied). A string derived by the application of all obligatory and some optional transformations is called a *T-terminal string*. We can regard a *T-terminal string* as being essentially a sequence of morphemes. A *T-terminal string* derived by the use of only obligatory transformations can be called a *kernel string*. If a language contains kernel strings at all, they will represent only the simplest sentences. The idea of using grammatical transformations to overcome the inadequacies of other types of generative grammars derives from Harris' investigation of the use of such operations to "normalize" texts (Harris, 1952a, 1952b). The description we give subsequently is basically that of Chomsky (1955); the exposition follows Chomsky (1961a). A rather different development of the underlying idea was given by Harris (1957).

As we have previously remarked, the reason for adding transformational rules to a grammar is simple. There are some sentences (simple declarative active sentences with no complex noun or verb phrases) that can be generated quite naturally by a constituent-structure grammar—more precisely, this is true only of the terminal strings underlying them. There are others (passives, questions, and sentences with discontinuous phrases and complex phrases that embed sentence transforms, etc.) that cannot be generated in a natural and economical way by a constituent-structure

grammar but that are, nevertheless, related systematically to sentences of simpler structure. Transformations express these relations. When used to generate more complex sentences (and their structural descriptions) from already generated simpler ones, transformations can account for aspects of grammatical structure that cannot be expressed by constituent-structure grammar.

The problem, therefore, is to construct a general and abstract notion of grammatical transformation, one that will incorporate and facilitate the expression of just those formal relations between sentences that have a significant function in language.

5.2 The Specification of Grammatical Transformations

A transformation cannot be simply an operation defined on terminal strings, irrespective of their constituent structure. If it could, then the passive transformation could be applied equally well in Examples 22 and 23:

The man saw the boy \rightarrow The boy was seen by the man (22)

The man saw the boy leave \nrightarrow $\left\{ \begin{array}{l} \text{The boy was seen by the man leave} \\ \text{The boy leave was seen by the man} \end{array} \right.$ (23)

In order to apply a transformation to some particular string, we must know the constituent structure of that string. For example, a transformation that would turn a declarative sentence into a question might prepose a certain element of the main verbal phrase of the declarative sentence. Applied to *The man who was here was old*, it would yield *Was the man who was here old?* The second *was* in the original sentence has been preposed to form a question. If the first *was* is preposed, however, we get the ungrammatical *Was the man who here was old?* Somehow, therefore, we must know that the question transformation can be applied to the second *was* because it is part of the main verbal phrase, but it cannot be applied to the first *was*. In short, we must know the constituent structure of the original sentence.

It would defeat the purpose of transformational analysis to regard transformations as higher level rewriting rules that apply to undeveloped phrase designations. In the case of the passive transformation, for example, we cannot treat it merely as a rewriting rule of the form

$$NP_1 + Auxiliary + Verb + NP_2 \rightarrow NP_2 + Auxiliary \\ + be + Verb + en + by + NP_1, \quad (24)$$

or something of the sort. Such a rule would be of the type required for a constituent-structure grammar, defined in Sec. 4, except that it would not

meet the condition imposed on constituent-structure rules in 16 (that is, the condition that only a single symbol be rewritten), which provides for the possibility of constructing a *P*-marker. A sufficient argument against introducing passives by such a rule as Example 24 is that transformations, so formulated, would not provide a method to simplify the grammar when selectional restrictions on choice of elements appear, as in the examples cited at the end of Sec. 4. In the passives corresponding to those examples the same selectional relations are obviously preserved, but they appear in a different arrangement. Now, the point is that, if the passive transformation were to apply as a rewriting rule at a stage of derivation preceding the application of the selectional rules for subject-verb-object, an entirely independent set of context-sensitive rules would have to be given in order to determine the corresponding agent-verb-subject selection in the passive. One of the virtues of a transformational grammar is that it provides a way to avoid this pointless duplication of selectional rules, with its consequent loss of generality, but that advantage is lost if we can apply the transformation before the selection of particular elements.

It seems evident, therefore, that a transformational rule must apply to a fully developed *P*-marker, and, since transformational rules must reapply to transforms, it follows that the result of applying a transformation must again be a *P*-marker, the *derived P*-marker of the terminal string resulting from the transformation. A grammatical transformation, then, is a mapping of *P*-markers into *P*-markers.

We can formulate this notion of grammatical transformation in the following way. Suppose that *Q* is a *P*-marker of the terminal string *t* and that *t* can be subdivided into successive segments t_1, \dots, t_n in such a way that each t_i is traceable, in *Q*, to a node labeled A_i . We say, in such a case, that *t* is *analyzable* as $(t_1, \dots, t_n; A_1, \dots, A_n)$ with respect to *Q*.

In the simplest case a transformation *T* will be specified in part by a sequence of symbols (A_1, \dots, A_n) that defines its domain by the following rule:

A string t with P-marker Q is in the domain of T if t is analyzable as $(t_1, \dots, t_n; A_1, \dots, A_n)$ with respect to Q. Then (t_1, \dots, t_n) is a proper analysis of t with respect to Q, T, and (A_1, \dots, A_n) is the structure index of T.

To complete the specification of the transformation *T*, we must describe the effect that *T* has on the terms of the proper analysis of any string to which it applies. For instance, *T* may have the effect of deleting or permuting certain terms, of substituting one for another, or adding a constant string in a fixed place, and so on. Suppose that we associate with a transformation *T* an underlying *elementary transformation* T_{e_i} such that

$T_{ei}(i; t_1, \dots, t_n) = \sigma_i$, where (t_1, \dots, t_n) is the proper analysis of t with respect to Q, T . Then the string resulting from application of the transformation T to the string t with P -marker Q is

$$T(t, Q) = \sigma_1 \dots \sigma_n.$$

Obviously, we do not want any arbitrary mapping of the sort just described to qualify as a grammatical transformation. We would not want, for example, to permit in a grammar a transformation that associates such pairs as *John saw the boy* → *I'll leave tomorrow*; *John saw the man* → *why don't you try again*; *John saw the girl* → *China is industrializing rapidly*. Only rules that express genuine structural relations between sentence forms—active-passive, declarative-interrogative, declarative-nominalized sentence, and so on—should be permitted in the grammar. We can avoid an arbitrary pairing off of sentences if we impose an additional, but quite natural, requirement on the elementary transformations. The restriction can be formulated as follows:⁴

If T_{ei} is an elementary transformation, then for all integers i and n and all strings $x_1, \dots, x_n, y_1, \dots, y_n$ it must be the case that $T_{ei}(i; x_1, \dots, x_n)$ is formed from $T_{ei}(i; y_1, \dots, y_n)$ by replacing y_j in the latter by x_j , for each $j \leq n$.

In other words, the effect of an elementary transformation is independent of the particular choice of strings to which it applies. This requirement has the effect of ruling out the possibility of applying transformations to particular strings of actually occurring words (or morphemes). Thus no single elementary transformation meeting this restriction can have both the effect of replacing *John will try* by *will John try?* and the effect of replacing *John tried* by *did John try?*, although this also is clearly the effect of the simple question-transformation. The elementary transformation that we need in this case is that which converts $x_1x_2x_3$ to $x_2x_1x_3$. That is to say, the transformation T_{ei} is defined as follows, for arbitrary strings x_1, x_2, x_3 :

$$T_{ei}(1; x_1, x_2, x_3) = x_2,$$

$$T_{ei}(2; x_1, x_2, x_3) = x_1,$$

$$T_{ei}(3; x_1, x_2, x_3) = x_3.$$

But if this is to yield *did John try?*, it will be necessary to apply it not to the sentence *John tried* but rather to a hypothetical string having the form *John + past + try* (a terminal string that is parallel in structure to the sentence *John will try*) that underlies the sentence *John tried*. In general, we cannot require that terminal strings be related in any simple way to

⁴ A more precise formulation would have to distinguish occurrences of the same string (Chomsky 1955).

actual sentences. The obligatory mappings (both transformational and phonological) that specify the physical shape may reorder elements, add or delete elements, and so on.

For empirical adequacy, the notion of transformation just described must be generalized in several directions. First, we must admit transformations that apply to pairs of *P*-markers. (Transformations such as those previously discussed that apply to a single *P*-marker we shall henceforth call *singular transformations*.) Thus the terminal string underlying the sentence *His owning property surprised me* is constructed transformationally from the already formed strings underlying *it surprised me* and *he owns property* (along with their respective *P*-markers). We might provide for this possibility, in the simplest cases, by allowing all strings $\#S\#\#S\# \dots \#S\#$ to head derivations in the underlying constituent-structure grammar instead of just $\#S\#$. We would then allow structure indices of the form $(\#, NP, V, NP, \#, \#, NP, V, NP, \#)$, thus providing for *His owning property surprised me* and similar cases. (For a more thorough and adequate discussion of this problem, see Chomsky, 1955.)

We must also extend the manner in which the domain of a transformation and the proper analysis of the transformed string is specified. First, there is no need to require that the terms of a structure index be single symbols. Second, we can allow the specification of a transformation to be given by a finite set of structure indices. More generally, we can specify the domain of a transformation simply by a structural condition based on the predicate *Analyzable*, defined above. In terms of this notion, we can define identity of terminal strings and can allow terms of the structure index to remain unspecified. With these and several other extensions, it is possible to provide an explicit and precise basis for transformational grammar.

5.3 The Constituent Structure of Transformed Strings

A grammatical transformation is determined by a structural condition stated in terms of the predicate *Analyzable* and by an elementary transformation. It has been remarked, however, that a transformation must produce not merely strings but derived *P*-markers. We must, therefore, show how constituent structure is assigned to the terminal string formed by a transformation. The best way to assign derived *P*-markers appears to be by a set of rules that would form part of general linguistic theory rather than by an additional clause appended to the specification of each transformation. Precise statement of those rules would require an analysis of fundamental notions going well beyond the informal account we have

sketched. (In this connection, see Chomsky, 1955; Matthews, 1962; Postal, 1962.) Nevertheless, certain features of a general solution to this problem seem fairly clear. We can, first of all, assign each transformation to one of a small number of classes, depending on the underlying elementary transformation on which it is based. For each class we can state a general rule that assigns to the transform a derived *P*-marker, the form of which depends, in a fixed way, on the *P*-markers of the underlying terminal strings. A few examples will illustrate the kinds of principles that seem necessary.

The basic recursive devices in the grammar are the generalized transformations that produce a string from a pair of underlying strings. (Apparently there is a bound on the number of singular transformations that can apply in sequence.) Most generalized transformations are based on elementary transformations that substitute a transformed version of the second of the pair of underlying terminal strings for some term of the proper analysis of the first of this pair. [In the terminology suggested by Lees (1960) these are the *constituent string* and the *matrix string*, respectively.] In such a case a single general principle seems to be sufficient to determine the derived constituent structure of the transform. Suppose that the transformation replaces the symbol α of σ_1 (the matrix sentence) by σ_2 (the constituent sentence). The *P*-marker of the result is simply the former *P*-marker of σ_1 with α replaced by the *P*-marker of σ_2 .

All other generalized transformations are attachment transformations that take a term α of the proper analysis, with the term β of the structure index that most remotely dominates it (and all intermediate parts of the *P*-marker that are dominated by β and that dominate α), and attaches it (with, perhaps, a constant string) to some other term of the proper analysis. In this way we form, for example, *John is old and sad* with the *P*-marker (Fig. 5) from *John is old*, *John is sad* by a transformation with the structure index (*NP*, *is*, *A*, *##*, *NP*, *is*, *A*).

Singular transformations are often just permutations of terms of the proper analysis. For example, one transformation converts Fig. 6*a* into Fig. 6*b*. The general principle of derived constituent structure in this case is simply that the minimal change is made in the *P*-marker of the underlying string, consistent with the requirement that the resulting *P*-marker again be representable in tree form. The transformation that gives *Turn some of the lights out* is based on an elementary transformation that permutes the second and third terms of a three-termed proper analysis; it has the structure index (*V*, *Prt*, *NP*) (this is, of course, a special case of a more general rule).

Figure 6 illustrates a characteristic effect of permutations, namely, that they tend to reduce the amount of structure associated with the terminal string to which they apply. Thus, although Fig. 6*a* represents the kind of

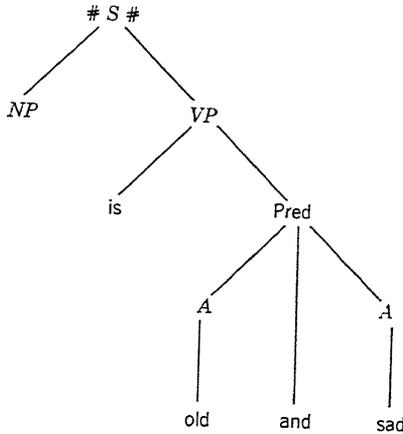


Fig. 5. *P*-marker resulting from the interpretation of *and* by an attachment transformation

purely binary structure regarded as paradigmatic in most linguistic theories, in Fig. 6*b* there is one less binary split and one new ternary division; and *Prt* is no longer dominated by *Verb*. Although binary divisions are, by and large, characteristic of the simple structural descriptions generated by the constituent-structure grammar, they are rarely found in *P*-markers associated with actual sentences. A transformational approach to syntactic description thus allows us to express the element of truth contained in the familiar theories of immediate constituent analysis, with their emphasis on binary splitting, without at the same time committing us to an arbitrary assignment of superfluous structure. Furthermore, by continued use of attachment and permutation transformations in the

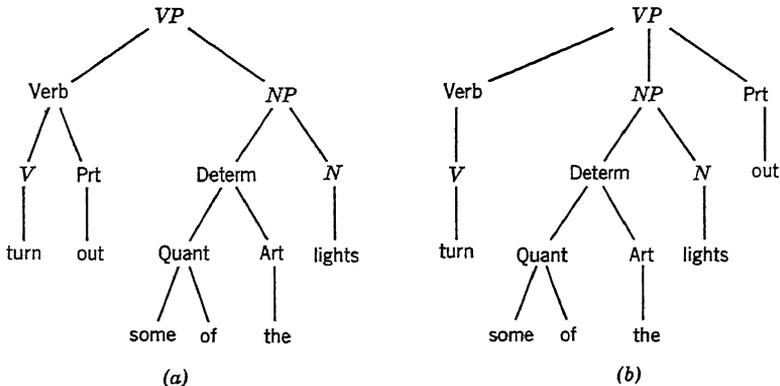


Fig. 6. The singular transformation that carries (a) into (b) is a permutation; its effect is to reduce slightly the amount of structure associated with the sentence.

manner illustrated it is possible to generate classes of P -markers that cannot in principle be generated by constituent-structure grammars (in particular, those associated with the coordinate constructions such as Example 20). Similarly, it is not difficult to see how a transformational approach can deal with the problem noted in connection with Example 18 and others like it (cf. Sec. 2.2 of Chapter 13), although the difficulty of actually working out adequate analyses should not be underestimated.

Some singulary transformations simply add constant strings at a designated place in the proper analysis; others delete certain terms of the proper analysis. The first are treated just like attachment transformations. In the case of deletion we delete nodes that dominate no terminal string and leave the P -marker otherwise unchanged. Apparently it is possible to restrict the application of deletion transformations in a rather severe way. Restrictions on the applicability of deletion transformations play a fundamental role in determining the kinds of languages that can be generated by transformational grammars.

In summary, then, a transformational grammar consists of a finite sequence of context-sensitive rewriting rules $\phi \rightarrow \psi$ and a finite number of transformations of the type just described, together with a statement of the restrictions on the order in which those transformations are applied. The result of a transformation is generally available for further transformation, so that an indefinite number of P -markers of quite varied kinds can be generated by repeated application of transformations. At each stage a P -marker representable as a labeled tree is associated with the terminal string so far derived. The full structural description of a sentence will consist not only of its P -marker but also of the P -markers of its underlying C -terminal strings and its transformational history. For further discussion of the role of such a structural description in determining the way a sentence is understood, see Sec. 2.2 of Chapter 13.

6. SOUND STRUCTURE

We regard a grammar as having two fundamental components, a *syntactic component* of the kind we have already described and a *phonological component* to which we now briefly turn our attention.

6.1 The Role of the Phonological Component

The syntactic component of a grammar contains rewriting rules and transformational rules, formulated and organized in the manner described

of the study of mathematical models for linguistic structure, the processes by which phonetic shape is assigned to utterances have a great deal of independent interest. We shall indicate briefly some of their major features. Our description of the phonological component follows closely Halle (1959a, 1959b) and Chomsky (1959, 1962a).

6.2 Phones and Phonemes

The phonological component can be thought of as an input-output device that accepts a terminal string with a labeled bracketing and codes it as a phonetic representation. The phonetic representation is a sequence of symbols of the phonetic alphabet, some of which (e.g., the first three of Example 26) are directly associated with physically defined features, others (e.g., the symbol + in Example 26), with features of transition. Let us call the first kind *phonetic segments* and the second kind *phonetic junctures*. Let us consider more carefully the character of the phonetic segments.

Each symbol of the universal phonetic alphabet is an abbreviation of a certain set of physical features. For example, the symbol [p^h] represents a labial aspirated unvoiced stop. These symbols have no independent status in themselves; they merely serve as notational abbreviations. Consequently a representation such as Example 26, and, in general, any phonetic representation, can be most appropriately regarded as a *phonetic matrix*: the rows represent the physical properties that are considered primitive in the linguistic theory in question and the columns stand for successive segments of the utterance (aside from junctures). The matrix element (*i*, *j*) indicates whether (or to what degree) the *j*th segment has the *i*th property. The phonetic segments thus correspond to columns of a

matrix. In Example 26 the symbol [ə] might be an abbreviation for the column [vocalic, nonconsonantal, grave, compact, unrounded, voiced, lax, tertiary stress, etc.], assuming a universal phonetic theory based on features that have been proposed by Jakobson as constituting a universal phonetic system. Matrices with such entries constitute the output of the phonological component of the grammar.

What is the input to the phonological component? The terminal string Example 25 consists of *lexical morphemes*, such as *Ted, book*; *grammatical morphemes*, such as *past, plural*; and certain *junctural elements*, such as #. The junctural elements are introduced by rules of the syntactic component in order to indicate positions in which morphological and syntactic structures have phonetic effects. They can, in fact, be regarded as grammatical morphemes for our purposes. Each grammatical morpheme is in general, represented by a single terminal symbol, unanalyzed into

features. On the other hand, the lexical morphemes are represented rather by strings of symbols that we call *phonemic segments* or simply *phonemes*.⁵ Aside from the labeled brackets, then, the input to the phonological component is a string consisting of phonemes and special symbols for grammatical morphemes. The representation in Example 25 is essentially accurate, except for the fact that lexical morphemes are given in ordinary orthography instead of in phonemic notation. Thus *Ted, see, the, book*, should be replaced by /ted/, /sī/, /ðī/, /buk/, respectively. We have, of course, given so little detail in Example 26 that phonetic and phonemic segments are scarcely distinguished in this example.

We shall return shortly to the question: What is the relation between phonemic and phonetic segments? Observe for now that there is no requirement so far that they be closely related.

Before going on to consider the status of the phonemic segments more carefully, we should like to warn the reader that there is considerable divergence of usage with regard to the terms phoneme, phonetic representation, etc., in the linguistic literature. Furthermore, that divergence is not merely terminological; it reflects deep-seated differences of opinion, far from resolved today, regarding the real nature of sound structure. This is obviously not the place to review these controversies or to discuss the evidence for one or another position. (For detailed discussion of these questions, see Halle, 1959b, Chomsky, 1962b, and the forthcoming *Sound Pattern of English* by Halle & Chomsky.) In the present discussion our underlying conceptions of sound structure are close to those of the founders of modern phonology but diverge quite sharply from the position that has been more familiar during the last twenty years, particularly in the United States—a position that is often called neo-Bloomfieldian. In particular, our present usage of the term phoneme is much like that of Sapir (e.g., Sapir, 1933), and our notion of a universal phonetic system has its roots in such classical work as Sweet (1877) and de Saussure (1916—the Appendix to the Introduction, which dates, in fact, from 1897). What we, following Sapir, call phonemic representation is generally called morphophonemic today. It is generally assumed that there is a level of representation intermediate between phonetic and morphophonemic, this new intermediate level usually being called phonemic. However, there seems to us good reason to reject the hypothesis that there exists an intermediate level of this sort and to reject, as well, many of the assumptions concerning sound structure that are closely interwoven with this hypothesis in many contemporary formulations of linguistic theory.

⁵ More precisely, we should take the phonemes to be the segments that appear at the stage of a derivation at which all grammatical morphemes have been eliminated by the phonological rules.

Clearly, we should attempt to discover general rules that apply to such large classes of elements as consonants, stops, voiced segments, etc., rather than to individual elements. We should, in short, try to replace a mass of separate observations by simple generalizations. Since the rules will apply to classes of elements, elements must be identified as members of certain classes. Thus each phoneme will belong to several overlapping categories in terms of which the phonological rules are stated. In fact, we can represent each phoneme simply by the set of categories to which it belongs; in other words, we can represent each lexical item by a *classificatory matrix* in which columns stand for phonemes and rows for categories and the entry (i, j) indicates whether or not phoneme j belongs to category i . Each phoneme is now represented as a sequence of categories, which we can call *distinctive features*, using one of the current senses of this term. Like the phonetic symbols, the phonemes have no independent status in themselves. It is an extremely important and by no means obvious fact that the distinctive features of the classificatory phonemic matrix define categories that correspond closely to those determined by the rows of the phonetic matrices. This point was noted by Sapir (1925) and has been elaborated in recent years by Jakobson, Fant, and Halle (1952) and by Jakobson and Halle (1956); it is an insight that has its roots in the classical linguistics that flourished in India more than two millenia ago.

6.3 Invariance and Linearity Conditions

The input to the phonological component thus consists, in part, of distinctive-feature matrices representing lexical items; and the output consists of phonetic matrices (and phonetic junctures). What is to be the relation between the categorial, distinctive-feature matrix that constitutes the input and the corresponding phonetic matrix that results from application of the phonological rules? What is to be the relation, for example, between the input matrix abbreviated as /ted/ (where each of the symbols /t/, /e/, /d/ stands for a column containing a plus in a given row if the symbol in question belongs to the category associated with that row, a minus if the symbol is specified as not belonging to this category, and a blank if the symbol is unspecified with respect to membership in this category) and the output matrix abbreviated as [t^he¹d] (where each of the symbols [t^h], [e¹], [d] stands for a column, the entries of which indicate phonetic properties)?

The strongest requirement that could be imposed would be that the

input classificatory matrix must literally be a submatrix of the output phonetic matrix, differing from it only by the deletion of certain redundant entries. Thus the phonological rules would fill in the blanks of the classificatory matrix to form the corresponding phonetic matrix. This strong condition, for example, is required by Jakobson and, implicitly, by Bloch in their formulations of phonemic theory.⁶ If this condition is met, then phonemic representation will satisfy what we can call the invariance condition and the linearity condition.

By the *linearity condition* we refer to the requirement that each phoneme must have associated with it a particular stretch of sound in the represented utterance and that, if phoneme *A* is to the left of phoneme *B* in the phonemic representation, the stretch associated with *A* precedes the stretch associated with *B* in the physical event. (We are limiting ourselves here to what are called *segmental phonemes*, since we are regarding the so-called supra-segmentals as features of them.)

The *invariance condition* requires that to each phoneme *A* there be associated a certain defining set $\Sigma(A)$ of physical phonetic features, such that each variant (allophone) of *A* has all the features of $\Sigma(A)$, and no phonetic segment which is not a variant (allophone) of *A* has all of the features of $\Sigma(A)$.

If both the invariance and linearity conditions were met, the task of building machines capable of recognizing the various phonemes in normal human speech would be greatly simplified. Correctness of these conditions would also suggest a model of perception based on segmentation and classification and would lend support to the view that the methods of analysis required in linguistics should be limited to segmentation and classification. However, correctness of these requirements is a question of fact, not of decision, and it seems to us that there are strong reasons to doubt that they are correct. Therefore, we shall not assume that for each phoneme there must be some set of phonetic properties that uniquely identifies all of its variants and that these sets literally occur in a temporal sequence corresponding to the linear order of phonemes.

We cannot go into the question in detail, but a single example may illustrate the kind of difficulty that leads us to reject the linearity and invariance conditions. Clearly the English words *write* and *ride* must appear in any reasonable phonemic representation as /rayt/ and /rayd/, respectively—that is, they differ phonemically in voicing of the final consonant. They differ phonetically in the vowel also. Consider, for example,

⁶ They would not regard what they call phonemic representations as the input to the phonological component. However, as previously mentioned, we see no way of maintaining the view that there is an intermediate representation of the type called “phonemic” by these and other phonologists.

a dialect in which *write* is phonetically [rayt] and *ride* is phonetically [ra·yd], with the characteristic automatic lengthening before voiced consonants. To derive the phonetic from the phonemic representation in this case, we apply the phonetic rule,

vowels become lengthened before voiced segments, (27)

which is quite general and can easily be incorporated into our present framework. Consider now the words *writer* and *rider* in such a dialect. Clearly, the syntactic component will indicate that *writer* is simply *write* + agent and *rider* is simply *ride* + agent, where the lexical entries *write* and *ride* are exactly as given; that is, we have the phonemic representations /rayt + r/, /rayd + r/ for *writer*, *rider*, respectively. However, there is a rather general rule that the phonemes /t/ and /d/ merge in an alveolar flap [D] in several contexts, in particular, after main stress as in *writer* and *rider*. Thus the grammar for this dialect may contain the phonetic rule,

[t, d] → D after main stress. (28)

Applying Rules 27 and 28, *in this order*, to the phonemic representations /rayt + r/, /rayd + r/, we derive first [rayt + r], [ra·yd + r], by Rule 27, and eventually [rayDr], [ra·yDr], by Rule 28, as the phonetic representations of the words *writer*, *rider*. Note however, that the phonemic representations of these words differ only in the *fourth* segment (voiced consonant versus unvoiced consonant), whereas the phonetic representations differ only in the *second* segment (longer vowel versus shorter vowel). Consequently, it seems impossible to maintain that a sequence of phonemes $A_1 \dots A_m$ is associated with the sequence of phonetic segments $a_1 \dots a_m$, where a_i contains the set of features that uniquely identify A_i in addition to certain redundant features. This is a typical example that shows the untenability of the linearity and invariance conditions for phonemic representation. It follows that phonemes cannot be derived from phonetic representations by simple procedures of segmentation and classification by criterial attributes, at least as these are ordinarily construed.

Notice, incidentally, that we have nowhere specified that the phonetic features constituting the universal system must be defined in absolute terms. Thus one of the universal features might be the feature "front versus back" or "short versus long." If a phonetic segment A differs from a phonetic segment B only in that A has the feature "short" whereas B has the feature "long," this means that in any particular context X — Y the longer element is identified as B and the shorter as A . It may be that A in one context is actually as long as or longer than B in another context. Many linguists, however, have required that phonetic features must be defined in absolute terms. Instead of the feature "short versus long," they require us to identify the absolute length (to some approximation) of each

segment. If we add this requirement to the invariance condition, we conclude that even partial overlapping of phonemes—that is, assignment of a phone *a* to a phoneme *B* in one context and to the phoneme *C* in a different context, in which the choice is contextually determined—cannot be tolerated. Such, apparently, is the view of Bloch (1948, 1950). This is an extremely restrictive assumption which is invalidated not only by such examples as the one we have just given but by a much wider range of examples of partial overlapping (see Bloch, 1940, for examples). In fact, work in acoustic phonetics (Liberman, Delattre, & Cooper, 1952; Schatz, 1954) has shown that if this condition must be met, where features are defined in auditory and acoustic terms (as proposed in Bloch, 1950), then not even the analysis of the stops /p, t, k/ can be maintained, since they overlap, a consequence that is surely a reduction to absurdity.

The requirements of relative or of absolute invariance both suggest models for speech perception, but the difficulty (or impossibility) of maintaining either of these requirements suggests that these models are incorrect and leads to alternative proposals of a kind to which we shall return.

We return now to the main theme.

6.4 Some Phonological Rules

We have described the input to the phonological component of the grammar as a terminal string consisting of lexical morphemes, grammatical morphemes, and junctures, with the constituent structure marked. This component gives as its output a phonetic matrix in which the columns stand for successive segments and the rows for phonetic features. Obviously, we want the rules of the phonological component to be as few and general as possible. In particular, we prefer rules that apply to large and to natural classes of elements and that have a simple and brief specification of relevant context. We prefer a set of rules in which the same classes of elements figure many times. These and other requirements are met if we define the complexity of the phonological component in terms of the number of features mentioned in the rules, where the form of rules is specified in such a way as to facilitate valid generalizations (Halle, 1961). We then choose simpler (more general) grammars over more complex ones with more feature specifications (more special cases).

The problem of phonemic analysis is to assign to each utterance a phonemic representation, consisting of matrices in which the columns stand for phonemes and the rows for distinctive (classificatory) features, and to discover the simplest set of rules (where simplicity is a well-defined

formal notion) that determine the phonetic matrices corresponding to given phonemic representations. There is no general requirement that the linearity and invariance conditions will be met by phonemic representations. It is therefore an interesting and important observation that these conditions are, in fact, substantially met, although there is an important class of exceptions.

In order to determine a phonetic representation, the phonological rules must utilize other information outside the phonemic representation; in particular, they must utilize information about its constituent structure. Consequently, it is in general impossible for a linguist (or a child learning the language) to discover the correct phonemic representation without an essential use of syntactic information. Similarly, it would be expected that in general the perceiver of speech should utilize syntactic cues in determining the phonemic representation of a presented utterance—he should, in part, base his identification of the utterance on his partial understanding of it, a conclusion that is not at all paradoxical.

The phonological component consists of (1) a sequence of rewriting rules, including, in particular, a subsequence of *morpheme structure rules*, (2) a sequence of *transformational rules*, and (3) a sequence of rewriting rules that we can call *phonetic rules*. They are applied to a terminal string in the order given.

Morpheme structure rules enable us to simplify the matrices that specify the individual lexical morphemes by taking advantage of general properties of the whole set of matrices. In English, for example, if none of the three initial segments of a lexical item is a vowel, the first must be /s/, the second a stop, and the third a liquid or glide. This information need not therefore be specified in the matrices that represent such morphemes as *string* and *square*. Similarly, the glide ending an initial consonant cluster need not be further specified, since it is determined by the following vowel; except after /s/, it is /y/ if followed by /u/, and it is /w/ otherwise. Thus we have *cure* and *queer* but not /kwūr/ or /kyīr/. There are many other rules of this sort. They permit us to reduce the number of features mentioned in the grammar, since one morpheme structure rule may apply to many matrices, and they thus contribute to simplicity, as previously defined. (Incidentally, the morpheme structure rules enable us to make a distinction on a principled and *non ad hoc* basis between permissible and nonpermissible nonsense syllables.)

Transformational phonemic rules determine the phonetic effects of constituent structure. (Recall that the fundamental feature of transformational rules, as they have been defined, is that they apply to a string by virtue of the fact that it has a particular constituent structure.) In English there is a complex interplay of rules of stress assignment and vowel reduction

that leads to a phonetic output with many degrees of stress and an intricate distribution of reduced and unreduced vowels (Chomsky, Halle, & Lukoff, 1956; Halle & Chomsky, forthcoming). These rules involve constituent structure in an essential manner at both the morphological and the syntactic level; consequently, they must be classified as transformational rather than rewriting rules. They are ordered, and apply in a *cycle*, first to the smallest constituents (that is, lexical morphemes), then to the next larger ones, and so on, until the largest domain of phonetic processes is reached. It is a striking fact, in English at least, that essentially the same rules apply both inside and outside the word. Thus we have only a single cycle of transformational rules, which, by repeated application, determines the phonetic form of isolated words as well as of complex phrases. The cyclic ordering of these rules, in effect, determines the phonetic structure of a complex form, whether morphological or syntactic, in terms of the phonetic structure of its underlying elements.

The rules of stress assignment and vowel reduction are the basic elements of the transformational cycle in English. Placement of main stress is determined by constituent type and final affix. As main stress is placed in a certain position, all other stresses in the construction are automatically weakened. Continued reapplication of this rule to successively larger constituents of a string with no original stress indications can thus lead to an output with a many-leveled stress contour. A vowel is reduced to [ɪ] in certain phonemic positions if it has never received main stress at an earlier stage of the derivation or if successive cycles have weakened its original main stress to tertiary (or, in certain positions, to secondary). The rule of vowel reduction applies only once in the transformational cycle, namely, when we reach the level of the word.

A detailed discussion of these rules is not feasible within the limits of this chapter, but a few comments may indicate how they operate. Consider in particular, the following four rules,⁷ which apply in the order given:

A *substantive* rule that assigns stress in initial position in nouns (also stems) under very general circumstances (29a)

A *nuclear stress* rule that makes the last main stress dominant, thus weakening all other stresses in the construction. (29b)

The *vowel reduction* rule. (29c)

A rule of *stress adjustment* that weakens all nonmain stresses in a word by one. (29d)

⁷ These differ somewhat from the rules that would appear in a more detailed and general grammar. See Halle & Chomsky (forthcoming) for details.

From the verbs *permit*, *tórmént*, etc., we derive the nouns *pérmit*, *tórment* in the next transformational cycle by the substantive rule, the stress on the second syllable being automatically weakened to secondary. The rule of stress adjustment gives primary-tertiary as the stress sequence in these cases. The second syllable does not reduce to [ɨ], since it is protected by secondary stress at the stage at which the rule of vowel-reduction applies.

Thus for *pérmit*, *tórment* we have the following derivations:

| | |
|---|--------------------------------|
| 1. $[_N[_V \text{ per} + \text{ mit}]_V]_N$ | $[_N[_V \text{ tórment}]_V]_N$ |
| 2. $[_N[_V \text{ per} + \text{ mit}]_V]_N$ | $[_N[_V \text{ tórment}]_V]_N$ |
| 3. $[_N \text{ per} + \text{ mit}]_N$ | $[_N \text{ tórment}]_N$ |
| 4. $[_N \text{ per} + \text{ mit}]_N$ | $[_N \text{ tórment}]_N$ |
| 5. $\text{per} + \text{mit}$ | tórment |
| 6. $\text{per} + \text{mit}$ | tórment |
| 7. $p^h\text{írmit}$ | $t^h\text{órment}$ |

Line 1 is the phonemic, line 7 the phonetic representation (details omitted). Line 2 is derived by a general rule (that we have not given) for *tórmént* and by Rule 29*b* for *permit* (since the heaviest stress in this case is zero). Line 3 terminates the first transformational cycle by erasing innermost brackets. Line 4 results from Rule 29*a*. Line 5 terminates the second transformational cycle, erasing innermost brackets. Line 6 results from Rule 29*d* (29*c* being inapplicable because of secondary stress on the second vowel), and line 7 results from other phonetic rules.

Consider, in contrast, the word *torrent*. This, like *tórmént*, has phonemic /e/ as its second vowel (cf. *torrential*), but it is not, like *tórmént*, derived from a verb *torrént*. Consequently, the second vowel does not receive main stress on the first cycle; it will therefore reduce by Rule 29*c* to [ɨ]. Thus we have reduced and unreduced vowels contrasting in *tórmént-tórrént* as a result of a difference in syntactic analysis. Initial stress in *tórrént* is again a result of Rule 29*a*.

The same rule that forms *pérmit* and *tórment* from *permit* and *tórmént* changes the secondary stress of the final syllable of the verb *advocate* to tertiary, so that it is reduced to [ɨ] by the rule of vowel reduction 29*c*. Thus we have reduced and unreduced vowels contrasting in the noun *advocate* and the verb *advocate* and generally with the suffix *-ate*. Exactly the same rules give the contrast between reduced and unreduced vowels

3. small boys' school
 (applying Rule 29*b* and erasing brackets)

III.1. [_N[_{NP} small boys]_{NP} school]_N

2. [_N small boys school]_N
 (applying Rule 29*b* to the innermost constituent and erasing brackets)

3. small boys school
 (applying Rule 29*a* and erasing brackets)

4. small boys school
 (by a rule of wide applicability that we have not given).

In short, a phonetic output that has an appearance of great complexity and disorder can be generated by systematic cyclic application of a small number of simple transformational rules, where the order of application is determined by what we know, on independent grounds, to be the syntactic structure of the utterance. It seems reasonable, therefore, to assume that rules of this kind underlie both the production and perception of actual speech. On this assumption we have a plausible explanation for the fact that native speakers uniformly and consistently produce and identify new sentences with these intricate physical characteristics (without, of course, any conscious awareness of the underlying processes or their phonetic effects). This suggests a somewhat novel theory of speech perception—that identifying an observed acoustic event as such-and-such a particular phonetic sequence is, in part, a matter of determining its syntactic structure (to this extent, understanding it). A more usual view is that we determine the phonetic and phonemic constitution of an utterance by detecting in the sound wave a sequence of physical properties, each of which is the defining property of some particular phoneme; we have already given some indication why this view (based on the linearity and invariance conditions for phonemic representation) is untenable.

We might imagine a sentence-recognizing device (that is to say, a perceptual model) that incorporates both the generative rules of the grammar and a heuristic component that samples an input to extract from it certain cues relating to the rules used to generate it, selecting among alternative possibilities by a process of successive approximation. With this approach, there is no reason to assume that each segmental unit has a particular defining property or, for that matter, that speech segments literally occur in sequence at all. Moreover, it avoids the implausible assumption that there is one kind of grammar for the talker and another kind for the listener.

Such an approach to perceptual processes has occasionally been suggested recently (MacKay, 1951; Bruner, 1958; Halle & Stevens, 1959, 1962; Stevens, 1960—with regard to speech perception, this view was in fact proposed quite clearly by Wilhelm von Humboldt, 1836). Recognition and understanding of speech is an obvious topic to study in developing this idea. On the basis of the sampled cues, a hypothesis can be formed about the spoken input; from this hypothesis an internal representation can be generated; by comparison of the input and the internally generated representation the hypothesis can be tested; as a result of the test, the hypothesis can be accepted or revised (cf. Sec. 2 of Chapter 13). Although speech perception is extremely complex, it is natural to normal adult human beings, and it is unique among complex perceptual processes in that we have, in this case at least, the beginnings of a plausible and precise generative theory of the organizing principles underlying the input stimuli.

References

- Berge, C. *Théorie des graphes et ses applications*. Paris: Dunod, 1958.
- Bloch, B. A set of postulates for phonemic analysis. *Language*, 1948, **24**, 3–46.
- Bloch, B. Phonemic overlapping. *Am. Speech*, 1940, **16**, 278–84. Reprinted in M. Joos (Ed.), *Readings in linguistics*. Washington: Am. Council Learned Socs., 1957. Pp. 93–96.
- Bloch, B. Studies in colloquial Japanese IV: Phonemics. *Language*, 1950, **26**, 86–125. Reprinted in M. Joos (Ed.), *Readings in linguistics*. Washington: Am. Council Learned Socs., 1957. Pp. 329–348.
- Bruner, J. S. Neural mechanisms in perception. In H. C. Solomon, S. Cobb, & W. Penfield (Eds.), *The brain and human behavior*. Baltimore: Williams and Wilkins, 1958. Pp. 118–143.
- Chomsky, N. *Logical structure of linguistic theory*. Microfilm, Mass. Inst. Tech. Library, 1955.
- Chomsky, N. Three models for the description of language. *IRE Trans. on Inform. Theory*, 1956, **IT-2**, 113–124.
- Chomsky, N. *Syntactic structures*. The Hague: Mouton, 1957.
- Chomsky, N. The transformational basis of syntax. In A. A. Hill (Ed.), *IVth Univer. of Texas Symp. on English and Syntax*, 1959 (unpublished).
- Chomsky, N. On the notion “Rule of grammar.” In R. Jakobson (Ed.), *Structure of language and its mathematical aspects*, *Proc. 12th Symp. in App. Math.* Providence, R. I.: American Mathematical Society, 1961. Pp. 6–24. (a) Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Chomsky, N. Some methodological remarks on generative grammar. *Word*, 1961, **17**, 219–239. (b).
- Chomsky, N. Explanatory models in linguistics. In E. Nagel, P. Suppes, & A. Tarski, (Eds.), *Logic, Methodology, & Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford: Stanford Univer. Press, 1962. Pp. 528–550. (a).

- Chomsky, N. The logical basis of linguistic theory. *Proc. Ninth Int. Cong. of Linguists*, 1962. Preprints, Cambridge, Mass., 1962. Pp. 509-574. (b). Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Chomsky, N., Halle, M., & Lukoff, F. On accent and juncture in English. In *For Roman Jakobson*. The Hague: Mouton, 1956.
- Čulík, K. *On some axiomatic systems for formal grammars and languages*. Mimeographed, 1962.
- Davis, M. *Computability and unsolvability*. New York: McGraw-Hill, 1958.
- Halle, M. *Sound pattern of Russian*, The Hague: Mouton, 1959. (a).
- Halle, M. Questions of linguistics. *Nuovo Cimento*, 1959, 13, 494-517. (b). Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Halle, M. On the role of simplicity in linguistic descriptions. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Symp. in App. Math.* Providence, R. I.; Amer. Mathematical Society, 1961. Pp. 89-94.
- Halle, M., & Chomsky, N. *Sound pattern of English*. (In prep.)
- Halle, M., & Stevens, K. N. Analysis by synthesis. *Proc. Seminar on Speech Compression and Production*, AFCRC-TR-59-198, 1959. Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Halle, M., & Stevens, K. N. Speech recognition: a model and a program for research. *IRE Trans. on Inform. Theory*, 1962, IT-8, 155-159.
- Harris, Z. S. Discourse analysis. *Language*, 1952, 28, 1-30. (a). Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Harris, Z. S. Discourse analysis: a sample text. *Language*, 1952, 28, 474-494. (b).
- Harris, Z. S. Co-occurrence and transformation in linguistic structure. *Language*, 1957, 33, 283-340. Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Humboldt, W. von. *Über die Verschiedenheit des menschlichen Sprachbaues*. Berlin, 1836. Facsimile edition: Bonn, 1960.
- Jakobson, R., Fant, C. G. M., & Halle, M. *Preliminaries to speech analysis*. Tech. Rept. 13, Acoustics Laboratory, Mass. Inst. Tech., Cambridge, Mass., 1952.
- Jakobson, R., & Halle, M. *Fundamentals of language*. The Hague: Mouton, 1956.
- Katz, J. Semi-sentences. In J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Kraft, L. G. *A device for quantifying, grouping, and coding amplitude modulated pulses*. MS thesis, Dept. Elec. Eng., Mass. Inst. Tech., 1949.
- Lees, R. B. Review of Chomsky, *Syntactic Structures*. *Language*, 1957, 33, 375-408.
- Lees, R. B. *A grammar of English nominalizations*. Supplement to *International J. Amer. Linguistics*. Baltimore, 1960.
- Liberman, A. M., Delattre, P., & Cooper, F. S. The role of selected stimulus variables in the perception of unvoiced stop consonants. *Amer. J. Psychol.*, 1952, 65, 497-516.
- MacKay, D. M. Mindlike behavior in artefacts. *Brit. J. Philos. Science*, 1951, 2, 105-121.
- Mandelbrot, B. On recurrent noise limiting coding. In *Proc. Symposium on Information Networks*, Polytechnic Institute of Brooklyn, 1954. Pp. 205-221.
- Matthews, G. H. Hidatsa syntax. Mimeographed, M.I.T., 1962.
- McMillan, B. Two inequalities implied by unique decipherability. *IRE Trans. on Inform. Theory*. December 1956, IT-2, 115-116.

- Miller, G. A. Speech and communication. *J. acoust. Soc. Amer.*, 1958, 30, 397-398.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana, Ill.: Univer. of Illinois Press, 1957.
- Postal, P. *Some syntactic rules in Mohawk*. PhD dissertation, Dept. of Anthropology, Yale University, 1962.
- Postal, P. Constituent analysis. Supplement to *International J. Amer. Linguistics*. Baltimore. (In press.)
- Rogers, H. The present theory of Turing machine computability. *J. soc. indust. appl. math.*, 1959, 7, 114-130.
- Sapir, E. Sound patterns in language. *Language*, 1925, 1, 37-51. Reprinted in D. G. Mandelbaum (Ed.), *Selected writings of Edward Sapir*. Berkeley: Univer. of California Press, 1949. Pp. 33-45.
- Sapir, E. La réalité psychologique des phonèmes. *J. de psychologie normale et pathologique*, 1933, 247-265. Reprinted in D. G. Mandelbaum (Ed.), *Selected writings of Edward Sapir*. Berkeley: Univer. of California Press, 1949. Pp. 46-60.
- Saussure, F. de. *Cours de linguistique générale*. Paris: 1916. Translation by W. Baskin, *Course in general linguistics*, New York: Philosophical Library, 1959.
- Schatz, C. D. The role of context in the perception of stops. *Language*, 1954, 30, 47.
- Schützenberger, M. P. On an application of semi-group methods to some problems in coding. *IRE Trans. on Inform. Theory*, 1956, IT-2, 47-60.
- Shannon, C. E. Communication in the presence of noise. *Proc. IRE*, 1949, 37, 10-21.
- Stevens, K. N. Toward a model for speech recognition. *J. acoust. Soc. Amer.*, 1960, 34, 47-55.
- Sweet, H. *A handbook of phonetics*. Oxford: Clarendon Press, 1877.
- Trakhtenbrot, B. A. *Algorithms and automatic computing machines*. Boston: Heath, 1963. Translated by J. Khristian, J. D. McCawley, & S. A. Schmitt from 2nd edition *Algoritmy i mashinnoe reshenie zadach*, 1960.
- Wallace, A. F. C. On being just complicated enough. *Proc. Nat. Acad. Sci.*, 1961, 47, 458-464.
- Ziff, P. *Semantic Analysis*. Ithaca, New York: Cornell Univer. Press, 1960. (a).
- Ziff, P. On understanding "Understanding utterances." Mimeographed, 1960. (b). Reprinted in J. Katz and J. Fodor (Eds.), *Readings in philosophy of language*. New York: Prentice-Hall, 1963.
- Ziff, P. About ungrammaticalness. Mimeographed, University of Pennsylvania, 1961.

I 2

*Formal Properties of Grammars*¹

Noam Chomsky

Massachusetts Institute of Technology

¹ *The preparation of this chapter was supported in part by the U.S. Army, the Air Force Office of Scientific Research, and the Office of Naval Research; and in part by the National Science Foundation (Grant No. NSF G-13903).*

Contents

| | |
|--|-----|
| 1. Abstract Automata | 326 |
| 1.1. Representation of linguistic competence, | 326 |
| 1.2. Strictly finite automata, | 331 |
| 1.3. Linear-bounded automata, | 338 |
| 1.4. Pushdown storage, | 339 |
| 1.5. Finite transducers, | 346 |
| 1.6. Transduction and pushdown storage, | 348 |
| 1.7. Other kinds of restricted-infinite automata, | 352 |
| 1.8. Turing machines, | 352 |
| 1.9. Algorithms and decidability, | 354 |
| 2. Unrestricted Rewriting Systems | 357 |
| 3. Context-Sensitive Grammars | 360 |
| 4. Context-Free Grammars | 366 |
| 4.1. Special classes of context-free grammars, | 368 |
| 4.2. Context-free grammars and restricted-infinite automata, | 371 |
| 4.3. Closure properties, | 380 |
| 4.4. Undecidable properties of context-free grammars, | 382 |
| 4.5. Structural ambiguity, | 387 |
| 4.6. Context-free grammars and finite automata, | 390 |
| 4.7. Definability of languages by systems of equations, | 401 |
| 4.8. Programming languages, | 409 |
| 5. Categorical Grammars | 410 |
| References | 415 |

Formal Properties of Grammars

A proposed theory of linguistic structure, in the sense of Chapter 11, must specify precisely the class of possible sentences, the class of possible grammars, and the class of possible structural descriptions and must provide a fixed and uniform method for assigning one or more structural descriptions to each sentence generated by an arbitrarily selected grammar of the specified form. In Chapter 11 we developed two conceptions of linguistic structure—the theory of constituent-structure grammar and the theory of transformational grammar—that meet these minimal conditions. We observed that the empirical inadequacies of the theory of constituent-structure grammar are rather obvious; for this reason there has been no sustained attempt to apply it to a wide range of linguistic data. In contrast, there is a fairly substantial and growing body of evidence that the theory of transformational grammar may provide an accurate picture of grammatical structure (Chomsky, 1962*b*, and references cited there).

On the other hand, there are very good reasons why the *formal* investigation of the theory of constituent-structure grammars should be intensively pursued. As we observed in Chapter 11, it does succeed in expressing certain important aspects of grammatical structure and is thus by no means without empirical motivation. Furthermore, it is the only theory of grammar with any linguistic motivation that is sufficiently simple to permit serious abstract study. It appears that a deeper understanding of generative systems of this sort, and the languages that they are capable of describing, is a necessary prerequisite to any attempt to raise serious questions concerning the formal properties of the richer and much more complex systems that do offer some hope of empirical adequacy on a broad scale. For the present this seems to be the area in which the study of mathematical models is most likely to provide significant insight into linguistic structure and the capacities of the language user.

In accordance with the terminology of Chapter 11, we may distinguish the *weak generative capacity* of a theory of linguistic structure (i.e., the set of languages that can be enumerated by grammars of the form permitted by this theory) from its *strong generative capacity* (the set of systems of structural descriptions that can be enumerated by the permitted grammars). This survey is largely restricted to weak generative capacity of constituent-structure grammars for the simple reason that, with a few exceptions, this is the only area in which substantial results of a mathematical character

have been achieved. Ultimately, of course, we are interested in studying strong generative capacity of empirically validated theories rather than weak generative capacity of theories which are at best suggestive. It is important not to allow the technical feasibility for mathematical study to blur the issue of linguistic significance and empirical justification. We want to narrow the gap between the models that are accessible to mathematical investigation and those that are validated by confrontation with empirical data, but it is crucial to be aware of the existence and character of the gap that still exists. Thus, in particular, it would be a gross error to suppose that the richness and complexity of the devices available in a particular theory of generative grammar can be measured by the weak generative capacity of this theory. In fact, it may well be true that the correct theory of generative grammar will permit generation of a very wide class of languages but only a very narrow class of systems of structural descriptions, that is to say, that it will have a broad weak generative capacity but a narrow strong generative capacity. Thus the hierarchy of theories that we establish in this chapter (in terms of weak generative capacity) must not be interpreted as providing any serious measure of the richness and complexity of theories of generative grammar that may be proposed.

1. ABSTRACT AUTOMATA

1.1 Representation of Linguistic Competence

At the outset of Chapter 11 we raised the problem of constructing (a) models to represent certain aspects of the competence achieved by the mature speaker of a language and (b) models to represent certain aspects of his behavior as he puts this competence to use. The second task is concerned with the actual *performance* of a speaker or hearer who has mastered a language; the first involves rather his *knowledge* of that language. Psychologists have long realized that a description of what an organism does and a description of what it knows can be very different things (cf. Lashley 1929, p. 553; Tolman, 1932, p. 364). A generative grammar that assigns structural descriptions to an infinite class of sentences can be regarded as a partial theory of what the mature speaker of the language knows. It in no sense purports to be a description of his actual performance, either as a speaker or as a listener. However, one can scarcely hope to develop a sensible theory of the actual use of language except on the basis of a serious and far-reaching account of what a language-user knows.

The generative grammar represents the information concerning sentence

structure that is available, in principle, to one who has acquired the language. It indicates how, ideally—leaving out any limitations of memory, distractions, etc.—he would understand a sentence (to the extent that the processes associated with “understanding” can be interpreted syntactically). In fact, such sentences as Example 11 in Chapter 11 are quite incomprehensible on first hearing, but this has no bearing on the question whether those sentences are generated by the grammar that has been acquired, just as the inability of a person to multiply 18,674 times 26,521 in his head is no indication that he has failed to grasp the rules of multiplication. In either case an artificial increase in memory aids, time, attention, etc., will probably lead the subject to the unique correct answer. In both cases there are problems that so exceed the user’s memory and attention spans that the correct answer will never be approached, and in both cases there is no reasonable alternative to the conclusion that recursive rules specifying the correct solution are represented somehow in the brain, despite the fact that (for quite extraneous reasons) this solution cannot be achieved in actual performance.

In a work that inaugurated the modern era of language study Ferdinand de Saussure (1916) drew a fundamental distinction between what he called *langue* and *parole*. The first is the grammatical and semantic system represented in the brain of the speaker; the second is the actual acoustic output from his vocal organs and input to his ears. Saussure drew an analogy between *langue* and a symphony, between *parole* and a particular performance of it; and he observed that errors or idiosyncracies of a particular performance may indicate nothing about the underlying reality. *Langue*, the system represented in the brain, is the basic object of psychological and linguistic study, although we can determine its nature and properties only by study of *parole*—just as a speaker can construct this system for himself only on the basis of actual observation of specimens of *parole*. It is the child’s innate *faculté de langage* that enables him to register and develop a linguistic system (*langue*) on the basis of scattered observations of actual linguistic behavior (*parole*). Other aspects of the study of language can be seriously undertaken only on the basis of an adequate account of the speaker’s linguistic intuition, that is, on the basis of a description of his *langue*.

This is the general point of view underlying the work with which we are here concerned. It has sometimes been criticized—even rejected wholesale—as “mentalistic”. However, the arguments that have been offered in support of this negative evaluation of the basic Saussurian orientation do not seem impressive. This is not the place to attempt to deal with them specifically, but it appears that the “antimentalistic” arguments that have been characteristically proposed would, were they correct, apply as well

against any attempt to construct explanatory theories. They would, in other words, simply eliminate science as an intellectually significant enterprise. Particular "mentalistic" theories may be useless or uninformative (as also "behavioral" or "mechanistic" theories), but this is not because they deal with "mentalistic" concepts that are associated with no necessary and sufficient operational or "behavioral" criterion. Observations of behavior (e.g., specimens of *parole*, particular arithmetical computations) may constitute the evidence for determining the correctness of a theory of the individual's underlying intellectual capacities (e.g., his *langue*, his innate *faculté de langage*, his knowledge of arithmetic), just as observations of color changes in litmus paper may constitute the evidence that justifies an assumption about chemical structure or as meter readings may constitute the evidence that leads us to accept or reject some physical theory. In none of these cases is the subject matter of the theory (e.g., innate or mature linguistic competence, ability to learn arithmetic or knowledge of arithmetic, the nature of the physical world) to be confused with the evidence that is adduced for or against it. As a general designation for psychology, "behavioral science" is about as apt as "meter-reading science" would be for physics (cf. Köhler, 1938, pp. 152-169).

Our discussion departs from a strict Saussurian conception in two ways. First, we say nothing about the semantic side of *langue*. The few coherent remarks that might be made concerning this subject lie outside the scope of the present survey. Second, our conception of *langue* differs from Saussure's in one fundamental respect; namely, *langue* must be represented as a generative process based on recursive rules. It seems that Saussure regarded *langue* as essentially a storehouse of signs (e.g., words, fixed phrases) and their grammatical properties, including, perhaps, certain "phrase types." Consequently, he was unable to deal with questions of sentence structure in any serious way and was forced to the conclusion that formation of sentences is basically a matter of *parole* rather than *langue*, that is, a matter of free and voluntary creation rather than of systematic rule. This bizarre consequence can be avoided only through the realization that infinite sets with certain types of internal structure (such as, in particular, sentences of a natural language with their structural descriptions) can be characterized by a finite recursive generative process. This insight was not generally available at the time of Saussure's lectures. Once we reformulate the notion of *langue* in these terms, we can hope to incorporate into its description a full account of syntactic structure. Furthermore, even the essentially finite parts of linguistic theory—phonology, for example—must now receive a rather different formulation, as we observed briefly in Chapter 11, Sec. 6. New and basic questions of a semantic nature can also be raised. Thus we can ask how a speaker uses the

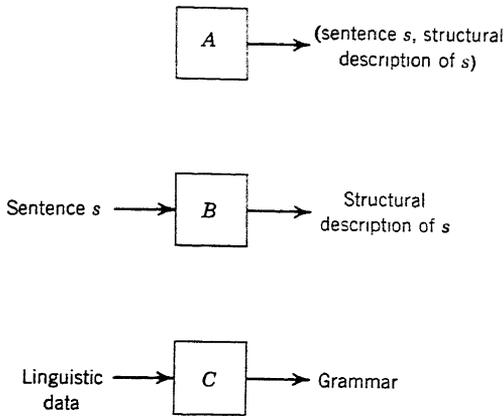


Fig. 1. Three types of psycholinguistic models suggested by the Saussurian conception of language.

recursive devices that specify sentences and their structural descriptions, on all levels, to interpret presented sentences, to produce intended ones, to utilize deviations from normal grammatical structure for expressive and literary purposes, etc. (cf. Katz & Fodor, 1962). It is impossible to maintain seriously the widespread view that our knowledge of the language involves familiarity with a fixed number of grammatical patterns, each with a certain meaning, and a set of meaningful items that can be inserted into them, and that the meaning of a new sentence is basically a kind of compound of these component elements.

With this modification, the Saussurian conception suggests for investigation three kinds of models, which are represented graphically in Fig. 1.

The device *A* is a grammar that generates sentences with structural descriptions; that is to say, *A* represents the speaker's linguistic intuition, his knowledge of his language, his *langue*. If we want to think of *A* as an input-output device, the inputs can be regarded as integers, and *A* can be regarded as a device that enumerates (in some order that is of no immediate interest) an infinite class of sentences with structural descriptions. Alternatively, we can think of the device *A* as being a theory of the language.

The device *B* in Fig. 1 represents the perceptual processes involved in determining sentence structure. Given a sensory input *s*, the hearer represented by *B* constructs an internal representation—a percept—which we call the structural description of *s*. The device *B*, then, would constitute a proposed account of the process of coming to understand a sentence, to the (by no means trivial) extent that this is a matter of determining its grammatical structure.

The device *C* represents the *faculté de langage*, the innate abilities that

make it possible for an organism to construct for itself a device of type *A* on the basis of experience with a finite corpus of utterances and, no doubt, information of other kinds.

The converse of *B* might be thought of as a model of the speaker, and, in fact, Saussure did propose a kind of account of the speaker as a device with a sequence of concepts as inputs and a physical event as output. But this doctrine cannot survive critical analysis. In the present state of our understanding, the problem of constructing an input-output model for the speaker cannot even be formulated coherently.

Of the three tasks of model construction just mentioned, the first is logically prior. A device of type *A* is the output of *C*—it is, in other words, one major result of the learning process. It also seems that one of the most hopeful ways to approach the problem of characterizing *C* is through an investigation of linguistic universals, the structural features common to all generative grammars. For acquisition of language to be possible at all there must be some sort of initial delimitation of the class of possible systems to which observed samples may conceivably pertain; the organism must, necessarily, be preset to search for and identify certain kinds of structural regularities. Universal features of grammar offer some suggestions regarding the form this initial delimitation might take. Furthermore, it seems clear that any interesting realization of *B* that is not completely *ad hoc* will incorporate *A* as a fundamental component; that is to say, an account of perception will naturally have to base itself on the perceiver's knowledge of the structure of the collection of items from which the preceived objects are drawn. These, then, are the reasons for our primary concern with the nature of grammars—with devices of the type *A*—in this chapter. It should be noted again that the logical priority of *langue* (i.e., the device *A*) is a basic Saussurian point of view.

The primary goal of theoretical linguistics is to determine the general features of those devices of types *A* to *C* that can be justified as empirically adequate—that can qualify as explanatory theories in particular cases. *B* and *C*, which represent actual performance, must necessarily be strictly finite. *A*, however, which is a model of the speaker's knowledge of his language, may generate a set so complex that no finite device could identify or produce all of its members. In other words, we cannot conclude, on the basis of the fact that the rules of the grammar represented in the brain are finite, that the set of grammatical structures generated must be of the special type that can be handled by a strictly finite device. In fact, it is clear that when *A* is the grammar of a natural language *L* there is no finite device of type *B* that will always give a correct structural description as output when and only when a sentence of *L* is given as input. There is nothing surprising or paradoxical about this; it is not a necessary

consequence of the fact that L is infinite but rather a consequence of certain structural properties of the generating device A .

Viewed in this way, several rather basic aspects of linguistic theory can be regarded, in principle at least, as belonging to the general theory of (abstract) automata. This theory has been studied fairly extensively (for a recent survey, see McNaughton, 1961), but it has received little attention in the technical literature of psychology and is not readily available to most psychologists. It seems advisable, therefore, to survey some well-known concepts and results (along with some new material) as background for a more specific investigation of sentence-generating devices in Secs. 2 to 5.

1.2 Strictly Finite Automata

The simplest type of automaton is the strictly finite automaton. We can describe it as a device consisting of a *control unit*, a *reading head*, and a *tape*. The control unit contains a finite number of parts that can be arranged in a finite number of distinct ways. Each of these arrangements is called an *internal state* of the automaton. The tape is blocked off into squares; it can be regarded as extending infinitely far both to the left and to the right (i.e., as doubly infinite). The reading head can scan a single tape square at a time and can sense the symbols a_0, \dots, a_D of a finite alphabet A (where a_0 functions as the identity element). We assume that the tape can move in only one direction—say right to left. We designate a particular state of the automaton as its *initial state* and label it S_0 . The states of the automaton are designated S_0, \dots, S_n ($n \geq 0$).

We can describe the operation of the automaton in the following manner. A sequence of symbols $a_{\beta_1}, \dots, a_{\beta_k}$ ($0 \leq \beta_i \leq D$) of the alphabet A is written on consecutive squares of the tape, one symbol to a square. We assume that the symbol $\#$, which is not a member of A , appears in all squares to the left of a_{β_1} and in all squares to the right of a_{β_k} . The control unit is set to state S_0 . The reading head is set to scan the square containing the symbol a_{β_1} . This initial tape-machine configuration is illustrated in Fig. 2.

The control unit is constructed so that when it is in a certain state and the reading head scans a particular symbol it switches into a new state while the tape advances one square to the left. Thus in Fig. 2 the control unit will switch to a new state S_i while the tape moves, so that the reading head is now scanning the symbol a_{β_2} . This is the second tape-machine configuration. The machine continues to compute in this way until it blocks (i.e., reaches a tape-machine configuration for which it has no instruction) or until it makes a first return to its initial state. In the latter

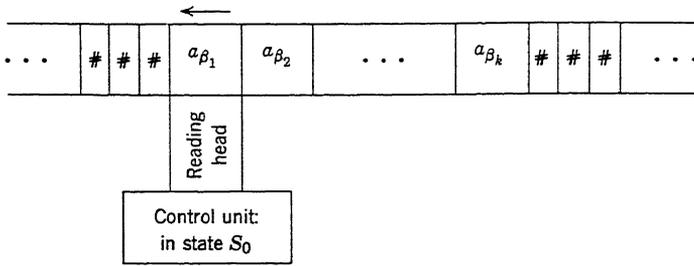


Fig. 2. Initial tape-machine configuration.

case, if the reading head is scanning the square to the right of a_{β_k} (in which case, incidentally, the machine is blocked, since this square contains $\# \notin A$), we say that the automaton has *accepted* (equivalently, *generated*) the string $\# a_{\beta_1} \dots a_{\beta_k} \#$. The set of strings accepted by the automaton is the *language accepted* (*generated*) by the automaton.

The behavior of the automaton is thus described by a finite set of triples (i, j, k) , $0 \leq i \leq D$, and $0 \leq j, k \leq n$, in which the triple (i, j, k) is interpreted as a rule asserting that if the control unit is in the state S_j and the reading head is scanning the symbol a_i then the control unit can shift to state S_k . The total behavior of the automaton can be represented by a *state diagram* consisting of nodes labeled by the names of the states and oriented paths (arrows) connecting the nodes, the paths labeled by the symbols of the vocabulary A . In the graph the node labeled S_j is connected by an arrow labeled a_i to the node labeled S_k just in case (i, j, k) is one of the triples describing the behavior of the automaton. An illustrative graph is shown in Fig. 3. (When we interpret these systems as grammars, the triples play the role of grammatical rules.) A finite automaton, then, is represented by an arbitrary finite directed graph with lines labeled by symbols of A . Tracing through the graph from S_0 to a first return to S_0 by one of the permissible paths, we generate the sentence $\# x \#$, in which x is the string consisting of the successive symbols labeling the arrows traversed in this path.

It is immaterial whether we picture an automaton as a source generating a sentence symbol by symbol as it proceeds from state to state or as a reader switching from state to state as it receives each successive symbol of the sentences it accepts. This is merely a matter of how we choose to interpret the notations. In either case, for the kind of system we have been considering, we have the following definition of a sentence:

Definition 1. *A string x of symbols is a sentence generated by the finite automaton F if and only if there is a sequence of symbols $(a_{\beta_1}, \dots, a_{\beta_r})$ of the alphabet of F and a sequence of states $(S_{\gamma_1}, \dots, S_{\gamma_{r+1}})$ of F such that*

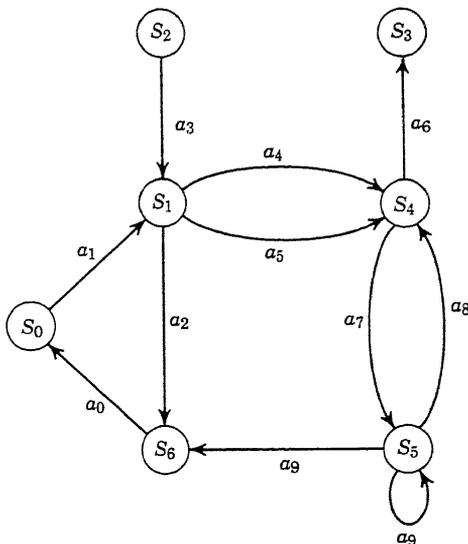


Fig. 3. Graph of finite automaton defined by the triples $(0, 6, 0)$, $(1, 0, 1)$, $(2, 1, 6)$, $(3, 2, 1)$, $(4, 1, 4)$, $(5, 1, 4)$, $(6, 4, 3)$, $(7, 4, 5)$, $(8, 5, 4)$, $(9, 5, 5)$, $(9, 5, 6)$. State S_0 is the initial and terminal state for all sentences. States S_2 and S_3 would normally be omitted, since they can play no role in the generation of sentences.

- (i) $\gamma_1 = \gamma_{r+1} = 0$; (ii) $\gamma_i \neq 0$ for $1 < i < r + 1$; (iii) $(\beta_i, \gamma_i, \gamma_{i+1})$ is a rule of F for each i ($1 \leq i \leq r$); (iv) $x = \# a_{\beta_1} \dots a_{\beta_r} \#$.

Any set of sentences generated by a finite automaton we call a *regular language*. A more familiar term in the literature for such sets is *regular event* (cf. Kleene, 1956; Rabin & Scott, 1959—the equivalence of the alternative formulations is worked out explicitly in Bar-Hillel & Shamir, 1960; Čulík, 1961).

Note that the device M shifts left with each interstate transition, that the identity symbol a_0 can occupy a square of the input tape, and that the instruction (i, j, k) applies to M just in case it is in state S_j scanning a square containing a_i . Equivalently, we can stipulate that a_0 cannot occupy a square of the tape, that the instruction (i, j, k) applies when M is in state S_j and either $i = 0$ or M is scanning a_i , and that the input tape moves left only if an instruction (i, j, k) is applied with $i \neq 0$. In this case we can think of the instruction $(0, j, k)$ as permitting a shift from S_j to S_k independent of the input symbol and with no shift of the input tape. Notice that with this formulation the device M will be blocked only if it is in state

S , scanning a symbol a_i for which it has no instruction (i, j, k) (as in the alternative formulation) and, furthermore, if it has no instruction $(0, j, k)$.

We generally omit the boundary symbol $\#$ in citing sentences generated by these and other sorts of automata.

Two such automata are *equivalent* if they generate the same language. We say that an automaton is *deterministic* if there are no rules (i, j, k) and (i, j, l) , where $k \neq l$, and no rule $(0, j, k)$ for $k \neq 0$ (that is, identity transitions are confined to return to S_0). The state of a deterministic device is uniquely determined (except for possible return to S_0) by the input string that it has read and the state from which it began computation.

A good deal is known about such devices. We state here, for later reference, two theorems without proof.²

Theorem 1. *Given finite automata F_1, F_2 , we can construct finite automata G_1, G_2, G_3 such that G_1 is deterministic and equivalent to F_1 , G_2 accepts just those strings in A that are rejected by F_1 , and G_3 accepts just those strings accepted by F_1 or by F_2 .*

Thus the set of all regular languages in the alphabet A is a Boolean algebra. We could, incidentally, eliminate all identity inputs in a deterministic device if we were to define "acceptance of a string x " in terms of entry into one of a set of designated *final states* instead of in terms of return to the initial state. These alternatives are equivalent as long as we allow any number of final states.

The most important result concerning regular languages is the structural characterization theorem of Kleene (1956). The theorem asserts that all regular languages, and only these, can be constructed from the finite languages by a few simple set-theoretic operations. It thus leads to simple and intuitive ways of representing any regular language (cf. Chomsky & Miller, 1958; McNaughton & Yamada, 1960). We can see this in the following way:

Given an alphabet A , we define a *representing expression* recursively as follows:

Definition 2. (i) *Every finite string in A is a representing expression.*
 (ii) *If X_1 and X_2 are representing expressions, then X_1X_2 is a representing expression.* (iii) *If X_i , $1 \leq i \leq n$, are representing expressions, then $(X_1, \dots, X_n)^*$ is a representing expression.*

This definition gives us (i) some representing expressions corresponding to strings in A and permits us to form more representing expressions either (ii) by juxtaposing them or (iii) by separating them by commas and grouping them inside parentheses marked by a star.

² Chomsky & Miller (1958). For these and many other results concerning finite automata and certain restricted infinite automata, see also Rabin & Scott (1959).

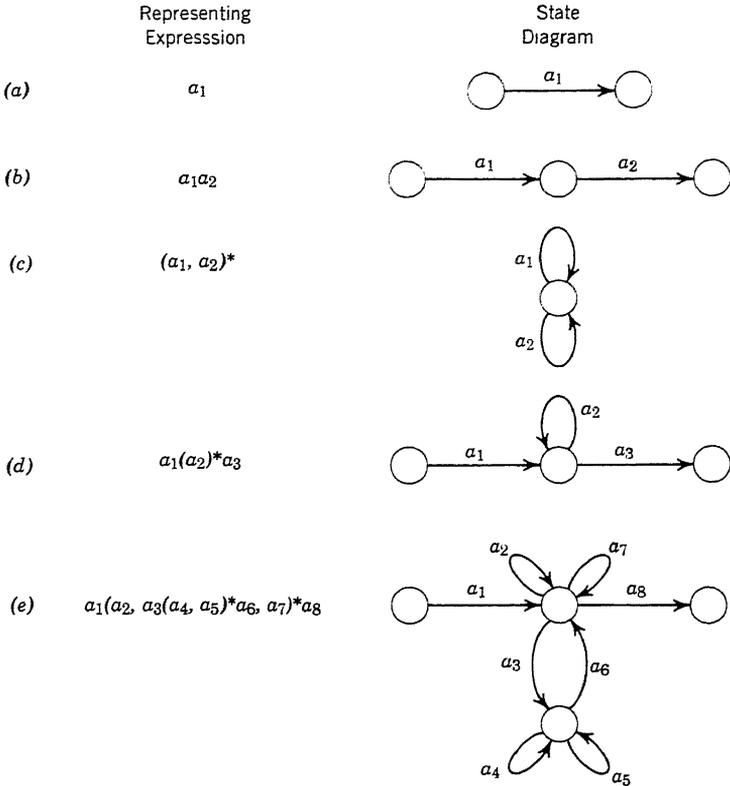


Fig. 4. Illustration of the use of the representing expressions.

Next, we want to say exactly what it is that these expressions are supposed to represent:

Definition 3. (i) *A finite string in A represents itself (better: the unit class containing it).* (ii) *If X_1 represents the set of strings Σ_1 and X_2 represents the set of strings Σ_2 , then X_1X_2 represents the set of all strings VW such that $V \in \Sigma_1$ and $W \in \Sigma_2$.* (iii) *If $X_i, 1 \leq i \leq n$, represents the set of strings Σ_i , then $(X_1, \dots, X_n)^*$ represents the set of all strings $V_1 \dots V_m$ such that for $1 \leq j \leq m$ there is a k ($1 \leq k \leq n$) such that $V_j \in \Sigma_k$.*

Note that according to condition (iii) the representing expression $(X_1, \dots, X_n)^*$ does not specify the order in which the elements V_j must occur but only the n sets from which they are chosen. The simplest way to grasp the import of Def. 3 is in terms of the state diagrams, or parts of state diagrams, that generate the sets of strings being represented. Figure 4b and 4c illustrate the formation of a set product (represented by juxtaposition) and the "star" operation [represented by $(X_1, \dots, X_n)^*$].

Figure 4d illustrates a combination of these operations. Thus it would generate a_1a_3 , $a_1a_2a_3$, $a_1a_2a_2a_3$, Figure 4e illustrates a still more complex element, etc.

We are now prepared to state:

Theorem 2. *L is a regular language if and only if there are representing expressions X_i , where $1 \leq i \leq n$, such that L is the sum of the sets Σ_i represented, respectively, by X_1, \dots, X_n . (Chomsky & Miller, 1958).*

The proof involves a demonstration that for any state diagram of an automaton F there is an equivalent state diagram composed of elements such as those shown in Fig. 4; from this alternative diagram the representing expressions for F can be read off directly. The constructed equivalent automaton is generally nondeterministic, of course.

A special class of finite automata of some interest consists of those with the property that the state of the automaton is determined by the last k symbols of the input sequence that it has accepted, for some fixed k . Such an automaton is called a *k-limited automaton* and the language it generates, a *k-limited language*.

Suppose that M is a k -limited automaton with a vocabulary V of D symbols. We can determine its behavior by a $D^k \times D$ matrix in which each column corresponds to an element $W \in V$ and each row to a string ϕ of length k of elements of V . The corresponding entry will be zero or one as the automaton does or does not accept W , having just accepted ϕ . Each such ϕ defines a state of the automaton.

This notion is familiar in the study of language under the following modification. Suppose that each entry of the defining matrix is a number between zero and one, representing the frequency with which the word corresponding to the given column occurs after the string of k words defining the given row in some sample of a language. Interpret this matrix as a description of a probabilistic k -limited automaton that generates strings in accordance with this set of transitional probabilities; that is, if the automaton is in the state defined by the i th row of the matrix that corresponds to the sequence of symbols $W_1^i \dots W_k^i$, then the entry (i, j) gives the probability that the next word it generates will be W_j . After having generated (accepted) W_j , it switches to the state defined by the string $W_2^i \dots W_k^i W_j$. Where $k \geq 1$, such a device generates what is called a $(k + 1)$ -order approximation to the sample from which the probabilities were derived (cf. Shannon & Weaver, 1949; Miller & Selfridge, 1950). We return to this notion in Sec. 1.2 of Chapter 13.

Clearly not every finite automation is a k -limited automaton. For example, the three-state automaton with the state diagram shown in Fig. 5 is not a k -limited automaton for any k . However, for each regular language L , there is a 1-limited language L^* and a homomorphism f such

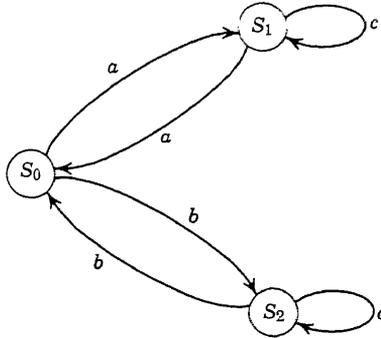


Fig. 5. A finite automaton that is not k -limited for any k .

that $L = f(L^*)$ (Schützenberger, 1961a). In fact, let L be accepted by a deterministic automaton M with no rule $(i, j, 0)$ for $i \neq 0$ (clearly there is such an M). Let M^* have the input alphabet consisting of the symbols (a_i, S_j) and the internal states $[a_i, S_j]$, where a_i is in the alphabet of M and S_j is a state of M , with $[a_0, S_0]$ as initial state. The transitions of M^* are determined by those of M by the following principle: if (i, j, k) is a rule of M , then M^* can move from state $[a_i, S_j]$ (for any l) to state $[a_i, S_k]$ when reading the input symbol (a_i, S_k) . Let L^* be the language accepted by M^* . Let f be the homomorphism that maps (a_i, S_j) into a_i for each i, j . Then $L = f(L^*)$ and L^* is 1-limited.

Suppose that we now relax the requirement that the tape must always shift left with each interstate transition. Instead, allow the direction of shift to be determined, as is the next state, by the present state and the symbol being read. The behavior of the automaton is now determined by a set of quadruples (i, j, k, l) , where i, j, k are, as before, indices of a letter, a state, and a state, respectively, and where l is one of $(+1, 0, -1)$. Following Rabin and Scott (1959), we interpret these quadruples in the following way:

Definition 4. Let (i, j, k, l) be one of the rules defining the automaton M . If the control unit of M is in state S_j and its reading head is scanning a square containing the symbol a_i , then the control unit may shift to state S_k while the tape shifts l squares to the left. A device of this sort we call a two-way automaton.

We regard a shift of -1 square to the left as a shift of one square to the right.

We can again say that such a device *accepts* (generates) a string exactly as a finite automaton does. That is to say, it accepts the string x only under the following condition. Let x be written on consecutive squares

of the tape, which is otherwise filled with $\#$'s. Let the control unit be set to the initial state S_0 , scanning the leftmost square not containing $\#$. Suppose that the device now computes until its first return to S_0 , at which point the control unit is scanning a square containing $\#$. In this case it accepts x .

It might be expected that by thus relaxing the conditions that a finite automaton must meet we would increase the generative capacity of the device. This is not the case, however, and we have (Rabin & Scott, 1959; Shepherdson, 1959) the following theorem:

Theorem 3. *The sets that can be generated by two-way automata are again the regular languages.*

The proof involves showing that the device can spare itself the necessity of returning to look a second time at any given part of the tape if, before it leaves that part, it thinks of all the questions (their number will be finite) it might later come back to ask, answers all of those questions right then, and carries the table of question-answer pairs forward along the tape with it, altering the answers when necessary as it goes along. Thus it is possible to construct an equivalent one-way automaton, although the price is to increase the number of internal states of the control unit.

1.3 Linear-Bounded Automata

Suppose that we were to allow a two-way automaton to write a symbol on the tape as it switches states. The symbols written on the tape belong to an *output alphabet* $A_O = \{a_0, \dots, a_p, \dots, a_q\}$ ($\# \notin A_O$), where $A_I = \{a_0, \dots, a_p\}$ is the *input alphabet*. We now have to specify the behavior of the device by a set of quintuples (i, j, k, l, m) , in which the set of quadruples (i, j, k, l) specifies a two-way automaton and the scanned symbol a_i is replaced by a_m (which may, of course, be identical with a_i) as the device switches from state S_j to S_k . Following Myhill (1960), in essentials, we have Def. 5.

Definition 5. *Let (i, j, k, l, m) be one of the rules defining M . If the control unit of M is in state S_j and its reading head is scanning a square containing a_i , then the control unit may switch to S_k while the tape shifts l squares left and the scanned symbol a_i is replaced by a_m . We call this device a linear-bounded automaton.*

Acceptance of a string is defined as before. In such a device the tape is now used for storage, not just for input. Therefore, when a linear-bounded automaton M is given the input x , it has available to it an amount of memory determined by $c \lambda(x) + q$, where q is the fixed memory of the control unit, c is a constant (determined by the size of the output alphabet),

and $\lambda(x)$ is the length of x . It is thus a simple, potentially infinite automaton, and, as we shall see directly, it can generate languages that are not regular.

It is sometimes convenient, in studying or attempting to visualize the performance of an automaton, to assign to it a somewhat more complex structure. Thus we can regard the device as having separate subparts for carrying out various aspects of its behavior. In particular, we can regard a linear-bounded automaton as having two separate infinite tapes, one solely for input and the other solely for computation, with the second tape having as many squares available for computation as are occupied by alphabetic symbols (i.e., occurring between $\# \dots \#$) on the input tape. We can also regard it as having several independent computation tapes of this kind. These modifications require appropriate changes in the description of the operation of the control unit, but it is not difficult to describe them in such a way as to leave the generative capacity of the class of automata in question unmodified.

1.4 Pushdown Storage

One special class of linear-bounded automata of particular interest is the following. Consider an automaton M with two tapes, one an *input tape*, the other a *storage tape*. The control unit can read from the input tape, and it can read from or write on the storage tape. The input tape can move in only one direction, let us say, right to left. The storage tape can move in either direction. Symbols of the input alphabet A_I can appear on the input tape, and symbols of the output alphabet A_O can be read from or printed on the storage tape, where A_I and A_O are as previously given. We assume that A_O contains a designated symbol $\sigma \notin A_I$ that will be used only to initiate or terminate computation in a way that will be described directly. In Secs. 1.4 to 1.6 we designate the identity element of A_O and A_I as e instead of a_0 . The other symbols of A_O we continue to designate as a_1, \dots, a_q . We continue to regard A_I and A_O as "universal alphabets," independent of the particular machine being considered.

We define a *situation* of the device as a triple (a, S_i, b) , in which a is the scanned symbol of the input tape, S_i is the state of the control unit, and b is the scanned symbol of the storage tape. Each step of a computation depends, in general, on the total situation of the device.

In the *initial tape-machine configuration* the input tape contains the symbols $a_{\beta_1}, \dots, a_{\beta_k}$ (where now $\beta_i \neq 0$) in successive squares, flanked on both sides by $\#$; and the control unit is in state S_0 scanning the leftmost

symbol a_{β_1} of $x = a_{\beta_1} \dots a_{\beta_k}$ (as in Fig. 2). The scanned square of the storage tape contains σ , and every other square contains $\#$. Thus the device is in the situation $(a_{\beta_1}, S_0, \sigma)$ in its initial configuration. The device computes in the manner subsequently described until its first return to S_0 . The input string x is *accepted* by the device if, at this point, $\#$ is being scanned on both the input and the storage tapes, that is, if the device is in the *terminal situation* $(\#, S_0, \#)$.

The special feature of these devices which distinguishes them from general linear-bounded automata is this. When the storage tape moves one square to the right, its previously scanned symbol is "erased." When the storage tape moves k squares to the left, exposing k new squares, k successive symbols of A_0 (all distinct from e) are printed in these squares. When it does not move, nothing is printed on it or erased from it. Thus only the rightmost symbol in storage is available at each stage of the computation. The symbol most recently written in storage is the earliest to be read out of storage. Furthermore, the storage tape will necessarily be completely *blank* (i.e., it will contain only $\#$) when the terminal situation $(\#, S_0, \#)$ is reached.

The device M which behaves in the way just described is called a *push-down storage (PDS) automaton*, following Newell, Shaw, and Simon (1959). This organization of memory has found wide application in programming, and, in particular, its utility for analysis of syntactic structure by computers has been noted by many authors. The reasons for this, as well as the intrinsic limitations, will become clearer when we see that the theory of PDS automata is, in fact, essentially another version of the theory of context-free grammar (see Chapter 11, Sec. 4). Note that a PDS automaton with a possibly nondeterministic control unit is a device that carries out "predictive analysis" in the sense of Rhodes (cf. Oettinger, 1961). Hence this theory, too, is essentially a variant of context-free grammar.

Let us now turn to a more explicit specification of PDS automata. We assume, selecting one of the two equivalent formulations mentioned on p. 333 above, that e cannot occupy a square of the input or storage tape. Thus we extend the definition of "situation" to include triples (e, S_i, b) , (a, S_i, e) , and (e, S_i, e) ; and we assume that when the device is in the situation (a, S_i, b) it is also, automatically, in the situations (e, S_i, b) , (a, S_i, e) , and (e, S_i, e) ; that is, any instruction that applies to the situations (e, S_i, b) , (a, S_i, e) , or (e, S_i, e) may apply when the device is in state S_i reading a on the input tape and b on the storage tape. The input tape will actually shift left only when an instruction involving $a \neq e$ on the input tape is applied.

Let us define a function $\lambda(x)$ (read "length of x ") for certain strings x as follows: $\lambda(\sigma) = -1$; $\lambda(e) = 0$; $\lambda(za_i) = \lambda(z) + 1$, where za_i is a string in $A_0 - \{\sigma\}$, $(1 \leq i \leq q)$.

Each instruction for a PDS automaton can now be given in the standardized form

$$(a, S_i, b) \rightarrow (S_j, x), \tag{1}$$

where $a \in A_I, b \in A_O, x = \sigma$ or x is a string on $A_O - \{\sigma\}$, and $j = 0$ if and only if $b = \sigma = x$. The instruction (1) applies when the device is in the situation (a, S_i, b) and has the following effect: the control unit switches to state S_j ; the input tape is moved $\lambda(a)$ squares to the left; the symbols of x are printed successively on the squares to the right of the previously scanned square of the storage tape—in particular, if $x = a_{\gamma_1} \dots a_{\gamma_m}$, then a_{γ_k} is printed in the k th square to the right of the previously scanned square of the storage tape, replacing the contents of this square—while the storage tape is moved $\lambda(x)$ squares to the left. Thus, if $x \neq \sigma$, the device is now scanning (on the storage tape) the rightmost symbol of x ; if $x = e$, it is still scanning b ; if $x = \sigma$, it is scanning the symbol to the left of b . Furthermore, we can think of each square to the right of the scanned square of the storage tape as being automatically erased (replaced by #). In any event, we define the *contents* of the storage tape as the string to the left of and including the scanned symbol, and we say that the storage tape *contains* this string. More precisely, if #, $a_{\beta_1}, \dots, a_{\beta_n}$ appear in successive squares of the storage tape, where a_{β_n} occupies the scanned square, then the string $a_{\beta_1} \dots a_{\beta_n}$ is the contents of the storage tape. If # is the scanned symbol of the storage tape, we say that this tape contains the string e (its contents is e) or that the storage tape is blank.

Note that when the automaton M applies Instruction 1 the input tape will move one square to the left if $a \neq e$ and will not move if $a = e$. Furthermore, if M is scanning # on the input tape, the Instruction 1 can apply only if $a = e$. The condition that $j = 0$ if and only if $b = \sigma = x$ implies that if M begins to compute from its initial configuration, then on its first return to S_0 it will necessarily be in a situation $(a, S_0, \#)$, for some a . If $a = \#$, then M is in the terminal situation $(\#, S_0, \#)$, scanning # on both input and storage tapes, and it therefore accepts the string that occupied the input tape in the initial configuration. There may, in fact, be further vacuous computation at this point if there is an instruction in the form of (1) with $a = e = b$ and $i = 0$, but this does not affect generative capacity. We can regard the device as blocked when it reaches the terminal situation. Its storage tape will be blank at this point and at no other stage of a computation.

We can give a somewhat simpler characterization of the family of languages accepted by PDS automata without explicit reference to tape manipulations, etc. Given A_I, A_O and σ , let us define a PDS automaton M as a finite set of instructions of the form of Instruction 1. For each i let $a_i \sigma = e$ —that is, σ is a general “right-inverse.” A *configuration*

of M is a triple $K = (x, S_i, y)$, where S_i is a state, x is a string in A_I , and y is a string in A_O . Think of x as being the still unread portion of the input tape (i.e., the string to the right of and including the scanned symbol) and y as the contents of the storage tape, where S_i is the present state. When I is the Instruction 1, we say that configuration K_2 follows from configuration K_1 by I if $K_1 = (ay, S_i, zb)$ and $K_2 = (y, S_j, zbx)$. We say that M accepts w if there is a sequence of configurations K_1, \dots, K_m such that $K_1 = (w, S_0, \sigma)$, $K_m = (e, S_0, e)$, and for each $i < m$ there is an instruction I of M such that K_{i+1} follows from K_i by I . M accepts (generates) the language L just in case L is the set of all strings accepted by M .

The memory of a PDS device can be represented in terms of the set of strings on an internal alphabet, transition from one internal configuration to another corresponding to addition or deletion of letters at the right-hand end of the strings associated with internal configurations. Thus from the "state" represented by the string $\phi\alpha$ transition is permitted only to states represented ϕ or $\phi\alpha\beta$. It may be instructive to compare a PDS device, which has, in this interpretation, an infinite set of potential states, with a k -limited automaton. As it was defined above, the memory of a k -limited automaton can also be represented in terms of the set of strings on an internal alphabet (in this case identical to the input alphabet). Transition in a k -limited automaton corresponds to the addition of a letter to the right-hand end of the string representing a state and simultaneous deletion of a letter from the left-hand end of that string. Thus the total set of potential states is finite.

A device with PDS is a special type of linear-bounded automaton. It can, of course, easily perform many tasks that a finite automaton cannot, although it makes only a "single pass" through the input data (i.e., the input tape moves in only one direction). Consider, for example, the task of generating (accepting) the language L_2' consisting of all strings $\#xx^*\#$, where x is a nonnull string of a 's and b 's and x^* is the mirror-image of x , that is, x read from right to left (cf. language L_2 in Chapter 11, Sec. 3, p. 285). This task is clearly beyond the range of a finite automaton, since the number of available states must increase exponentially as the device accepts successive symbols of the first half of the input string. Consider, however, the PDS automaton M with the input alphabet $\{a, b, c\}$, the internal states S_0, S_1 , and S_2 , and the following rules, where α ranges over $\{a, b\}$:

- | | | |
|-------|---|-----|
| (i) | $(\alpha, S_0, e) \rightarrow (S_1, \alpha)$ | |
| (ii) | $(\alpha, S_1, e) \rightarrow (S_1, \alpha)$ | |
| (iii) | $(c, S_1, e) \rightarrow (S_2, e)$ | (2) |
| (iv) | $(\alpha, S_2, \alpha) \rightarrow (S_2, \sigma)$ | |
| (v) | $(e, S_2, \sigma) \rightarrow (S_0, \sigma)$ | |

The control unit has the state-diagram shown in Fig. 6, in which the triple (r, s, t) is on the arrow leading from state S_i to state S_j just in case the device has the rule $(r, S_i, s) \rightarrow (S_j, t)$. Clearly, this device will accept a string if and only if it is in L_2' . For example, the successive steps in accepting $\#abcba\#$ are given in Fig. 7.

Evidently pushdown storage is an appropriate device for accepting (generating) languages such as L_2' , which have, in the obvious sense, nesting of units (phrases) within other units, that is, the kind of recursive property that in Chapter 11, Sec. 3, we called *self-embedding*. We shall see in Secs. 4.2 and 4.6 that the essential properties of context-free grammars (cf. Chapter 11, Sec. 4) distinguishing them from finite automata are that they permit self-embedding and symmetries in the generated strings. Consequently, we would expect that pushdown storage would be useful in dealing with languages with grammars of this type. This class obviously includes many familiar artificial languages (e.g., sentential calculus and probably many programming languages—cf. Sec. 4.8). It is, in fact, a straightforward matter to construct a PDS automaton that will recognize or generate the sentences of such systems. Oettinger (1961) has pointed out that if we equip a PDS device with an output tape and adjust its instructions to permit it to map an input string into a corresponding output (using its PDS in the computation) we can instruct it to translate between ordinary and Polish notation, for example. To some approximation, context-free constituent-structure grammars are partially adequate for natural languages; that is, nesting of phrases (self-embedding) and symmetry are basic properties of natural languages. Consequently, such devices as PDS will no doubt be useful for actual handling of natural-language texts by computers for one or another purpose.

The device (2) is deterministic. In the case of finite automata we have observed (cf. Theorem 1) that, given any finite automaton, there is an equivalent deterministic one. This observation is not true, however, for the class of PDS automata. There is, for example, no deterministic PDS automaton that will accept the language $L_2 = \{xx^* \mid x^* \text{ the mirror-image of } x\}$ (thus L_2 consists of the strings formed by deleting the midpoint element c from the strings of L_2'), since the device will have no way of knowing when it has reached the middle of the input string; but L_2 is accepted by the nondeterministic PDS device derived from the device (2) for L_2' by replacing Rule iii by

$$(\alpha, S_1, e) \rightarrow (S_2, \alpha). \tag{3}$$

This amounts to dropping the arrow labeled (c, e, e) in Fig. 6 and connecting S_1 to S_2 with two arrows, one labeled (a, e, a) and the other (b, e, b) . The device uses Instruction (3) when it “guesses” that it has

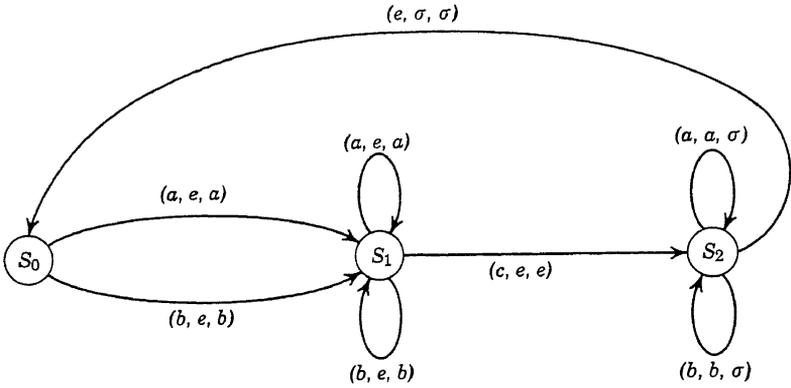


Fig. 6. State diagram for M .

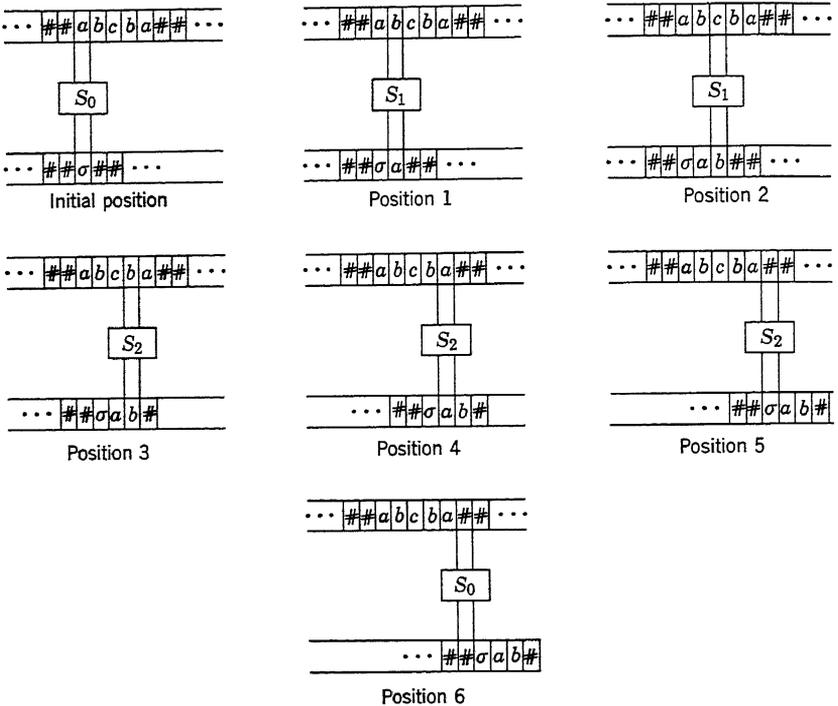


Fig. 7. Generation of $\#abcba\#$ with pushdown storage.

reached the middle of the string. If the guess is wrong, its computation will not terminate with acceptance (just as when the input is not a string of L_2); if the guess is right and the input is in L_2 , the computation will terminate with acceptance.

In this discussion we have assumed that the next act of the device is determined in part by the symbol being scanned on the storage tape. It is interesting to inquire to what extent control from the storage tape is essential for PDS devices. Consider the two subclasses of PDS devices defined by the following conditions: M is a *PDS automaton without control* if each rule is of the form $(a, S_i, e) \rightarrow (S_j, x)$; M is a *PDS automaton with restricted control* if each rule is of one of the two forms $(a, S_i, e) \rightarrow (S_j, x)$, $x \neq \sigma$, or $(a, S_i, b) \rightarrow (S_j, \sigma)$. In other words, in the case of a PDS device with restricted control the symbol being scanned on the storage tape plays a role in determining only the computations that "erase" from storage. Thus the device of Fig. 6 has restricted control. In the case of a PDS automaton without control the storage tape is acting only as a counter. We can, without loss of generality, assume that only one symbol can be written on it.

Concerning these families of automata we observe, in particular, the following:

Theorem 4. (i) *The family of PDS automata without control is essentially richer in generative capacity than the family of finite automata but essentially poorer than the full family of PDS devices.* (ii) *For each PDS device there is an equivalent PDS device with restricted control.*

As far as Part i is concerned, it is obvious that a PDS automaton without control can accept the language $L_1 = \{a^n b^n\}$ (cf. Chapter 11, Sec. 3) but not the language L_2 or L_2' . In fact, these languages are beyond the range of a device with any finite number of infinite counters that shift position independently in a fixed manner with each interstate transition (e.g., a counter could register the number of times the device has passed through a particular state or produced a particular symbol), where the decision whether to accept an input string depends on an elementary property of the counters (i.e., are they equal; do they read zero, as in the case of PDS automata; etc.). Although a device with q counters and k states has a potentially infinite memory, after p transitions at most kp^q configurations of a state and a counter reading can have been reached, and obviously 2^p different configurations must be available after p transitions for generating sentences of L_2 of length $2p$ (the availability of identity transitions does not affect this observation). See Schützenberger (1959) for a discussion of counter systems in which this description is made precise.

Part ii of Theorem 4 follows as a corollary of several results that we shall establish (cf. Theorem 6, Sec. 1.6).

1.5 Finite Transducers

Suppose that we have a PDS device M meeting the additional restriction that the storage tape never moves to the right, that is, each rule of M is of the form $(a, S_i, b) \rightarrow (S_j, x)$, where $x \neq \sigma$. Beginning in its initial configuration with the successive symbols $\#, a_{\beta_1}, \dots, a_{\beta_n}, \#$ on the input tape, the device will compute in accordance with its instructions, moving its storage tape left whenever it prints a string x on that tape. Suppose that the device continues to compute until it reaches the situation $(\#, S_i, a_j)$, for some i, j ; that is, it does not block before reading in the entire input tape. At this point the storage tape contains some string $y = wa_j$, and we say that the device M maps the string $a_{\beta_1} \dots a_{\beta_n}$ into the string y . We call M a *transducer*, which maps input strings into output strings and, correspondingly, input languages into output languages. We designate by $M(L)$ the set of strings y such that, for some $x \in L$, M maps x into y . Note that a transducer can never reach a configuration in which it is scanning $\#$ on the storage tape. Consequently, it can never accept its input string in the sense of "acceptance" as previously defined. In the case of a transducer we can regard the storage tape as an *output tape*.

In the case of a transducer the restrictions on the form of instructions for PDS automata that involve return to S_0 are clearly inoperative. In fact, we can allow an instruction I of a transducer to be of the form $(a, S_i, b) \rightarrow (S_j, x)$ [as in (1)], where $a \in A_I$, $b \in A_O$ and x is a string in $A_O - \{\sigma\}$, dropping the other restrictions.

Where M is a transducer, it is clear that the storage tape is playing no essential role in determining the course of the computation; and, in fact, we can construct a device, T , that effects the same mapping as M , while meeting the additional restriction that the next state is determined only by the input symbol and the present state. We designate the states of T in the form (S_i, a) , where S_i is a state of M and $a \neq e$ is a symbol of its output alphabet. The initial state of T is (S_0, σ) . Where M has the rule $(a, S_i, b) \rightarrow (S_j, x)$, T will have the rule

$$[a, (S_i, b), b] \rightarrow [(S_j, c), x], \quad (4)$$

where either $x = yc$ or $x = e$ and $c = b$. Clearly the behavior of T is in no way different from that of M , but in the case of T the next step in the computation depends only on the input symbol and the present state. It is thus a PDS device without control. Eliminating the redundant specification of the scanned symbol of the storage tape, we can give all the rules of T in the form

$$(a, \Sigma_i) \rightarrow (\Sigma_j, x), \quad (5)$$

indicating that when T is in state Σ_i and is scanning a (if $a \neq e$) or is scanning any symbol (if $a = e$) on the input tape it may switch to state Σ_j , move the input tape $\lambda(a)$ squares left, and the storage tape $\lambda(x)$ squares left, printing x on the newly exposed squares (if any) of the storage tape. Each transducer can thus be fully represented by a state-diagram, in which nodes represent states, and an arrow labeled (a, x) leads from Σ_i to Σ_j just in case Rule 5 is an instruction of the device.

Suppose that in the state-diagram representing the transducer M there is no possibility of traversing a closed path, beginning and ending with a given node, following only arrows labeled (e, x) , for some x . More formally, there is no sequence of states $(S_{x_1}, \dots, S_{x_k})$ of M such that $\alpha_1 = \alpha_k$, and, for each $i < k$, there is an x_i such that $(e, S_{x_i}) \rightarrow (S_{x_{i+1}}, x_i)$ is an instruction of M . If this condition is met, the number of outputs that can be given with a single input is bounded, and we call M a *bounded transducer*.

The mapping effected by a transducer we call a (*finite*) *transduction*. A transduction is a mapping of strings into strings (hence languages into languages) of a kind that can be performed by a strictly finite device.

Given a bounded transducer T , we can obviously eliminate as many of the instructions of the form $(e, S_i) \rightarrow (S_j, x)$ as we like without affecting the transduction performed by simply allowing the device to print out longer strings on interstate transitions. Alternatively, by adding a sufficient number of otherwise unused states and enough rules of the form of Rule 5 where $a = e$ we can construct, corresponding to each transducer T , a transducer T' which performs the same transduction as T but has rules only of the form $(a, S_i) \rightarrow (S_j, b)$, where $b \in A_O$.

Note that, given such a T' , we can construct immediately the "inverse" transducer T^* that maps the string y into the string x just in case T maps x into y by simply interchanging input and output symbols in the instructions of T' ; for example, in Rule 5, where $x \in A_O$, interchanging a and x . This amounts to replacing each label (a, b) on an arrow of the state-diagram by the label (b, a) . Hence, given any transducer T , we can construct the inverse transducer T^* that maps y into x just in case T maps x into y and that maps the language L onto the set of all strings x such that T maps x into $y \in L$. If T is bounded, its inverse T^* may still be unbounded. If the inverse T^* of T is bounded, then T is called *information lossless*. For general discussion of various kinds of transduction, see Schützenberger (1961a) and Schützenberger & Chomsky (1962).

We shall study the effects of transduction on context-free languages in Sec. 4.5. Note that if T is a transducer mapping L onto L' , where L is a regular language, then L' is also a regular language. Note also that for each regular language L there is a transducer T_L that will map L onto

the particular language U (alternatively U onto L), where U is the set of all strings in the output alphabet (alternatively, the input alphabet—if the input alphabet contains only e , the transducer will, by definition, be unbounded; otherwise, it can always be bounded). These and many related facts are fairly obvious from inspection of state-diagrams.

1.6 Transduction and Pushdown Storage

We have described a transducer as a PDS device which never moves its storage tape to the right—that is, it never erases—on any computation step. It maps an input string x into an output string y . A general PDS device, on the other hand, uses its storage tape to determine its later steps, in particular, its ultimate acceptance of the input string x . It terminates its computation with acceptance of x only if, on termination, the contents of the storage tape is simply e , that is, the storage tape is blank. We could, therefore, think of a general PDS device as defining a mapping of the strings it accepts into the empty string e , which is the contents of the storage tape when the computation terminates with acceptance of the input. (The device would essentially represent the characteristic function of a certain set of strings.) We now go on to show how we can associate with each PDS device M a transducer T constructed so that when and only when M accepts x (i.e., maps it into e) T maps x into a string y which, in a sense that we shall define, reduces to e .³

Suppose that M is a PDS device with input alphabet A_I and output alphabet $A_O = \{e, a_1, \dots, a_q\}$. We will construct a new device M' with the input alphabet A_I and the output alphabet $A_{O'}$ with $2q + 1$ symbols, where $A_{O'} = A_O \cup \{a'_1, \dots, a'_q\}$. We will treat each element a'_i as essentially the “right inverse” of a_i . More formally, let us say that the string x reduces to y just in case there is a sequence of strings z_1, \dots, z_m ($m \geq 1$) such that $z_1 = x$, $z_m = y$, and for each $i < m$ there are strings w_i, \bar{w}_i and $a_{\beta_i} \in A_O$ such that $z_i = w_i a_{\beta_i} a'_{\beta_i} \bar{w}_i$ and $z_{i+1} = w_i \bar{w}_i$. In other words, x reduces to y if $x = y$ or if y can be formed from x by successive deletions of substrings $a_i a'_i$.

We say that the string x is blocked if $x = y a z a'_i w$, where z reduces to e and either ya reduces to e or $a \in A_O - \{e, a_i\}$. If x is blocked, then, for all v , xv is blocked and does not reduce to e . We say that the storage tape is blocked if the string that it contains is blocked.

The new device M' will be a PDS automaton which never moves its

³ The results in this section and Sec. 4.2 are the product of work done jointly with M. P. Schützenberger. For a concise summary, see Chomsky (1962a). See Schützenberger (1962a,b,d) for generalizations and related results.

storage tape to the right. It will be constructed in such a way that if M does not accept x then, with x as input, M' will terminate its computation before reading through x or with the storage tape blocked; and, if M does accept x , M' will be able to compute in such a way that when it has read through all of x the storage tape will not be blocked—in fact, its contents will reduce to e .

The states of M' will be designated by the same symbols as those of M , and S_0 will again be the initial state.

Suppose that K and K' are tape-machine configurations of M and M' , respectively, meeting the following conditions. K is attainable from the initial configuration of M . The string w contained on the storage tape of M' in K' reduces to the string y contained on the storage tape of M in K . Furthermore, if $y \neq e$, then $w = za_k$ for some k (i.e., it has an unprimed symbol to its extreme right). M and M' are scanning the same square of identical input tapes and are in the same internal state. In this case we say that K and K' *match*. Note that when K and K' match then either M has terminated with the storage tape blank (in which case the contents of the storage tape of M' is za_k which reduces to e) or M and M' are in the same situation.

The instructions of M' are determined by those of M by the following rule. Let

$$(b, S_i, a_k) \rightarrow (S_j, x) \tag{6}$$

be an instruction of M . If $x \neq \sigma$, then M' will also have Instruction 6. Suppose that $x = \sigma$. Then, if $a_k = \sigma$ (in which case $j = 0$), M' will have Instruction 7, and, if $a_k \neq \sigma$, then for each r ($1 \leq r \leq q$) M' will have Instruction 8

$$(b, S_i, \sigma) \rightarrow (S_0, \sigma') \tag{7}$$

$$(b, S_i, a_k) \rightarrow (S_j, a_k' a_r' a_r). \tag{8}$$

Suppose now that K_1 and K_2 are configurations of M , K_1' is a configuration of M' that matches K_1 , K_1 is not terminal, and Instruction 6 of M carries it from K_1 to K_2 . Clearly, if $x \neq \sigma$ in Instruction 6, then the instruction of M' corresponding to 6 will carry M' from K_1' to a configuration K_2' that matches K_2 .

Suppose then that $x = \sigma$ in Instruction 6. Since K_1 is by assumption not a terminal configuration, M in K_1 must contain on its storage tape a string ya_k for some k . Either $y = e$ or $y = za_r$ for some r .

Suppose $y = e$. It must be that $a_k = \sigma$ and that $j = 0$ in Instruction 6. Hence M' has the corresponding Instruction 7, which carries it from K_1' to a configuration K_2' . But K_1' matches K_1 , and thus the contents of the storage tape of M' in K_1' must be $t\sigma$, where t reduces to e . Applying Instruction 7, M' moves into K_2' where the contents of the storage tape is $t\sigma\sigma'$, which reduces to e . Thus K_2' matches K_2 .

Suppose that $y = za_r$. Then Instruction 6 carries M into K_2 in which the storage tape contains za_r . Since K_1' matches K_1 , the contents of the storage tape of M' in K_1' must be a string ta_rua_k , where t reduces to z and u to e . By construction, M' has Instruction 8 (corresponding to Instruction 6); it carries M' into K_2' , which is identical to K_2 with respect to input tape and internal state and in which the storage tape contains $ta_rua_k a_k' a_r' a_r$, which reduces to $za_r = y$, and K_2' matches K_2 .

In each case we see that if Instruction 6 carries M from K_1 to K_2 then M' has an instruction to carry it from K_1' (which matches K_1) to K_2' (which matches K_2).

Suppose that K_1 is again a nonterminal configuration of M and K_1' is a matching configuration of M' , that an instruction I' of M' carries M' into the configuration K_2' , and that there is no instruction of M to carry it into a configuration K_2 that matches K_2' . It is clear that I' was not derived (by the construction previously given) from an instruction of M of the form of Instruction 6, where $x \neq \sigma$. Thus we can assume that I' is either Instruction 7 or 8 and that I' was derived by the construction presented from Instruction I : $(b, S_i, a_k) \rightarrow (S_j, \sigma)$. In any case, since K_1 and K_1' match and neither is a terminal configuration, the storage tape of M in K_1 must contain a string ya_k , where $y = e$ or $y = za_s$ for some s , and the storage tape of M' in K_1' must contain a string va_k , where v reduces to y .

Suppose that I' is Instruction 7. Then $a_k = \sigma$, and the contents of the storage tape of M' in K_2' is $v\sigma\sigma'$. M' terminates in the state S_0 with the storage tape containing a string that reduces to y ; but since $\sigma (= a_k)$ cannot be printed on the storage tape in any step of M and since the contents of the storage tape of M in K_1 is $y\sigma$, it must be that $y = e$. Thus I carries M from K_1 to a configuration K_2 which matches K_2' , contrary to assumption.

Thus it must be that I' is Instruction 8. Then the contents of the storage tape of M' in K_2' is $va_k a_k' a_r' a_r$, which reduces to $ya_k a_k' a_r' a_r$, and, in turn, to $ya_r' a_r$. If $y = e$, then the storage tape of M' is blocked in K_2' . Suppose that $y = za_s$. Then I carries M into a configuration K_2 in which the storage tape contains za_s . Suppose that $s = r$. Then the contents of the storage tape of M' in K_2' , which reduces to $ya_r' a_r = za_s a_r' a_r$, reduces further to $za_r = za_s$. But in this case K_2 and K_2' match, contrary to assumption. Therefore $r \neq s$. But in this case, the storage tape of M' in K_2' again is blocked. In any case, then, it is blocked if I' is Instruction 8.

As we have observed, once the storage tape is blocked it will remain blocked for the rest of the computation. Hence, once I' is applied, the device M' cannot reach a configuration in which the storage tape reduces to e .

Briefly, then, M' makes the guess that, after "erasing" $a_k \neq \sigma$ and entering configuration K_2 , M will be scanning a_r on the storage tape. It thus writes $a'_k a'_r a_r$ on its storage tape, so that it, too, is now scanning a_r , having "erased" both a_k (by a'_k) and a_r (by a'_r). If the guess was right, the new configuration of M' matches K_2 ; if it was wrong, the storage tape of M' is blocked, and the computation cannot terminate with the storage tape containing a string that reduces to e .

But M and M' have the identical initial configuration when they have identical input tapes. Hence, if M accepts x , M' will be able to compute from its initial configuration with input x until it terminates in the situation $(\#, S_0, \sigma')$ with a string y on the storage tape which reduces to e ; and, if M does not accept x , then no computation of M' from its initial configuration with x on the input tape can terminate in the situation $(\#, S_j, a)$, for some a , with a string that reduces to e as the contents of the storage tape. (Note that the contents of the storage tape of M' can reduce to e if and only if M' has just printed σ' and returned to S_0 by an instruction such as 7.)

Note, also, that M' is a PDS automaton which never moves the storage tape to the right (never "erases"). We have already shown in Sec. 1.5 how, corresponding to each device of this sort, we can construct an equivalent transducer which operates independently of the contents of the storage tape. Let T be the transducer that is constructed from M' in the manner described in Sec. 1.5. Then M accepts x if and only if T maps x into a string y that reduces to e .

Thus we have the following general result.

Theorem 5. *Given a PDS automaton M , we can construct a transducer T such that M accepts x if and only if T maps x into a string y that reduces to e .*

Suppose, now, that $L(M)$ is the language accepted by the PDS automaton M , T is the corresponding transducer guaranteed by Theorem 5, K is the set of strings in the output alphabet of T that reduce to e , U_I is the set of all strings in the input alphabet of T , and T' is the inverse transducer that maps x onto y just in case T maps y onto x . Then $L(M) = T'(K \cap T(U_I))$. But U_I is a regular language and, as we have noted in Sec. 1.5, it follows that $T(U_I)$ is a regular language. It is also easy to show that K is a context-free language (cf. Sec. 4, Chapter 11).

We shall see in Sec. 4.6 that the intersection of a context-free language and a regular language is a context-free language and that a finite transducer maps a context-free language into another context-free language. It follows, then, that $L(M)$ is a context-free language.

We shall also see in Theorem 17, Sec. 4.2, that corresponding to each context-free language there is a PDS automaton with restricted control

(cf. Sec. 1.4) that accepts it. From these results, then, we can conclude the following:

Theorem 6. *The following are equivalent:*

- (i) *L is accepted by a PDS automaton;*
- (ii) *L is accepted by a PDS automaton with restricted control [cf. Theorem 4(ii)];*
- (iii) *L is a context-free language.*

1.7 Other Kinds of Restricted-Infinite Automata

The field of restricted-infinite automata is of great potential interest not only for the theory of computability but also, one would expect, for psychology, since psychologically relevant models representing the knowledge and competence of an organism will presumably be neither strictly finite nor unmanageably infinite in the sense of the devices to which we shall turn our attention in Sec. 1.8. However, the investigation of this topic has only recently been undertaken and only a few initial results are available. Ritchie (1960) has investigated a hierarchy of devices with the general property that at each level of the hierarchy the memory available to a device is a function of the length of the input, where this function can be computed by a device of the next lowest level, the lowest level being that of the finite automata. Yamada (1960) has studied the case of "real time" automata, which are subject to the condition that the number of computing steps allowed is determined by the length of the input (see McNaughton, 1961, for a survey of some of his results). Schützenberger (1961b, 1962e) has developed the theory of finite automata equipped with sets of counters that can change state in accordance with simple arithmetical conditions with each step of computation. Schützenberger (1961c) has also related some of these results to a general theory of context-free grammars (cf. Sec. 4.7). The devices studied in Secs. 2 and 3 are restricted-infinite automata, and it seems reasonable to predict that it is in this setting that the mathematical study of grammar will ultimately find its natural home.

1.8 Turing Machines

Each device M of the kind we have considered is characterized by the property that the amount of time it will take and the amount of space it will use in solving a particular problem (i.e., in accepting or rejecting a certain input) is, in some sense, predictable in advance. In fact, the deepest

and most far-reaching mathematical investigations in the theory of automata concern devices that do not share this property. Such devices, called Turing machines, we now consider briefly.

We obtain a Turing machine by taking a linear-bounded automaton, as defined in Def. 5, and adding to its rules quintuples $(\#, j, k, l, m)$, where $j \neq 0$, having just the properties defined in Def. 5. These rules have the effect of allowing the device to use previously inaccessible portions of tape on the left or right in the process of computation, since $\#$'s can now be rewritten as symbols of the alphabet. It is also customary to require that these devices be deterministic in the sense that for a given tape-machine configuration no more than one move is possible; if (i, j, k, l, m) and (i, j, k', l', m') are rules, then $k = k'$, $l = l'$, and $m = m'$. We can say that a Turing machine accepts (generates) a string under essentially the conditions previously given. Specifically, we write on the tape the symbols of the string ϕ in successive squares, the rest of the infinite tape being filled by $\#$'s. We set the control unit to S_0 with the reading head scanning the leftmost symbol $\alpha \neq \#$. If the device computes until a first return to S_0 , we say that the machine has accepted (generated) ϕ . Furthermore, the sequence of symbols which now appears on the tape between $\#$'s spells a string ψ that we can call the output of the machine. We can, in other words, regard it as a partial function that maps ϕ into ψ under the conditions just stated.

The automata obtained in this way are totally different in their behavior from those considered in Secs. 1.2 to 1.7. There is, for example, no general way to determine whether, for a given input, the device will run into a block or an infinite loop. There is, furthermore, no way to determine, from systematic inspection of the instructions for the device, how long it will compute or how much tape it will use before it gives an answer, if it does accept the string. There is no uniform and systematic way to determine from a study of the rules for a Turing machine whether it will ever give an output at all or whether its output or the set it accepts will be finite or infinite; nor is it in general possible to determine by some mechanical procedure whether two such devices will ever give the same output or will accept the same set. Nevertheless, it is important to observe that a Turing machine is specified by a finite number of rules and at any point in a computation it will be using only a finite amount of tape (i.e., only a finite number of squares will appear between the bounding strings of $\#$). Furthermore, if it is going to accept a given input, this fact will be known after a finite number of operations. However, if it does not accept the input, this fact may never be known. (If, after a certain number of steps, it has still not returned to S_0 , we do not know whether this is because it has not computed long enough or because it never will reach this terminal state,

and we may never know.) The study of Turing machines constitutes the basis for a rapidly developing branch of mathematics (recursive function theory). For surveys of this field, see Davis (1958) and Rogers (1961).

1.9 Algorithms and Decidability

It is interesting to observe that there are Turing machines that are *universal* in the sense that they can mimic the behavior of any arbitrary Turing machine. Suppose, in fact, that among the strings formed from an alphabet A , which we can assume to be the common alphabet of all Turing machines, we select an infinite number to represent the integers; for example, let us take $a_1 = 1$ and regard the string $1 \dots 1$ consisting of n successive 1's as representing the number n . (We shall, henceforth use the notation x^n for the string consisting of n successive occurrences of the string x .) Suppose now that we have an enumeration M_1, M_2, \dots of all of the infinitely many Turing machines. This enumeration can be given in a perfectly straightforward and definite way. Then there is a *universal* Turing machine M_u with the following property: M_u will accept the input $1^n \alpha x$ and give, with this input, the output y , just in case M_n accepts x and gives y as the corresponding output (where α is some otherwise unused symbol). We can think of the input tape to M_u as containing the stored program 1^n , which instructs M_u to act in the manner of the n th Turing machine when any input x is written on its tape. Each Turing machine can thus be regarded as one of the programs for a universal machine M_u . An ordinary digital computer is, in effect, a universal Turing machine such as M_u if we make the idealized assumption that memory (e.g., new tape units) can be added whenever needed, without limit, in the course of a particular computation. The program stored in the computer instructs it as to which Turing machine it should mimic in its computations.

Given a set Σ of strings, we are often interested in determining whether a particular string x is or is not a member of Σ . Furthermore, we are often interested in determining whether there is a mechanical (effective) procedure by which, given an arbitrary string x , we can tell after a finite amount of time whether or not x is a member of Σ . If such a procedure exists, we say that the set Σ is *decidable* or *computable*, that there is an *algorithm for determining membership in Σ* , or that *the decision problem for Σ is (recursively) solvable* (these all being equivalent locutions). An algorithm for determining whether an arbitrary element x is a member of the set Σ can be regarded as a computer program with the following property. A computer storing this program, given x as input, is guaranteed to terminate its computation with the answer *yes* (when $x \in \Sigma$) or *no* (when

$x \notin \Sigma$). We must assume here that the computer memory is unbounded; that is to say, we are dealing with an idealized digital computer—a universal Turing machine.

Suppose that we now revise our characterization of Turing machines just to the following extent: we add to the control unit a designated state S^* and we say that, given the input x , the device *accepts* x if it returns to S_0 , as before, and that it *rejects* x if it reaches S^* . Call such a device a two-output Turing machine. Given Turing machines M_1 and M_2 , which accept the disjoint sets Σ_1 and Σ_2 , respectively, it is always possible to construct a two-output machine M_3 that will accept just Σ_1 and reject just Σ_2 .

With this revision, consider again the question of decidability. A set Σ is decidable if there is a computer program that is guaranteed to determine, of an arbitrary input x , whether x is a member of Σ , after a finite number $t(x)$ of steps. We can now reformulate the notion of decidability as follows: a set is decidable if there is a two-output Turing machine that will accept all its members and reject all its nonmembers.

A set is called *recursively enumerable* just in case there is a Turing machine that accepts all strings of this set and no others. This machine is then said to *recursively enumerate* (*generate*) the set. A set is *recursive* if both it and its complement are recursively enumerable. It is clear that a set is recursive just in case it is decidable in the sense just defined, for, if Σ is recursive, then there is a Turing machine M_1 that accepts Σ and a Turing machine M_2 that accepts its complement. Consequently, as previously observed, we can construct a two-output machine M_3 that will accept the set Σ enumerated by M_1 and reject the set $\bar{\Sigma}$ (= complement of Σ) enumerated by M_2 . To determine whether a string x is in Σ , we can write x on the input tape and set M_3 to computing. This amounts to setting both M_1 and M_2 to computing synchronously with input x . After some finite time, one or the other machine will have come to a stop and have accepted x ; therefore, after some finite time, M_3 will either have accepted or rejected x , and we shall know whether x is in Σ or its complement. On the other hand, if Σ is decidable, then it is recursive, since the two-output device, which accepts all its members and rejects all nonmembers, can easily be separated into two Turing machines, one of which accepts all members and the other, all nonmembers.

A classic result in Turing-machine theory is that there are recursively enumerable sets that are not recursive. In fact, some rather familiar sets have this property. Thus the set of valid schemata of elementary logic (i.e., the theory of *and*, *or*, *not*, *some*, *all*—called first-order predicate calculus or quantification theory) is recursive if we restrict ourselves to one-place predicates but nonrecursive (though recursively enumerable)

if we allow two-place predicates, that is, relations. Or consider a formalized version of ordinary elementary number theory. If it meets the usual conditions of adequacy for axiomatic systems, then it is known that although the set of theorems is recursively enumerable it is not recursive. In fact, elementary number theory has the further property that there is a mechanical procedure f such that if M_i is any two-output Turing machine that accepts all of the theorems deducible from some consistent set of axioms for this theory and rejects the negations of all of these theorems, then $f(i)$ is a formula (in fact, a true formula) of elementary number theory that is neither accepted nor rejected by M_i . There does not exist a two-output machine which accepts a set Σ and rejects its complement $\bar{\Sigma}$ if Σ contains all of the theorems of this system and none of their negations. There are, furthermore, perfectly reasonable sets that are not recursively enumerable; for example, the set of true statements in elementary number theory or the set of all satisfiable schemata of quantification theory.

The notions of decidability and existence of algorithms can be immediately extended beyond the particular questions just discussed. Let us define a *problem* as a class of questions, each of which receives a yes-or-no answer. The problem is called *recursively solvable* or *decidable* if there is a mechanical procedure, which, applied to any question from this class, will, after a finite time, give either *yes* or *no* as its answer. Decidability of problems can, in general, be formulated in the manner indicated previously. To consider a case to which we will return, suppose that we are given a set of generative grammars G_1, G_2, \dots of a certain sort (note that to be *given* a set is, in this context, to be given a device that recursively enumerates it), and that we are interested in determining whether the *equivalence problem* is recursively solvable. This is the problem of determining, of an arbitrary pair of grammars G_i, G_j , whether or not they generate the same language. We can formulate the question as follows: is there a two-output Turing machine that accepts the string $1^i a 1^j$ if G_i and G_j generate the same language and rejects the string $1^i a 1^j$ if G_i and G_j generate different languages? The equivalence problem is recursively solvable for the set G_1, G_2, \dots just in case there is such a device. Or consider the problem of determining, of an arbitrary G_i , whether it generates a finite set. This problem is recursively solvable just in case there is a two-output Turing machine which accepts the string 1^i in case G_i generates a finite language and rejects it in case G_i generates an infinite language. Similarly, other decision problems can be formulated in this way, for example, the problems of equivalence and finite generation that we have observed to be unsolvable for Turing machines.

Turing machines (and computers) are usually regarded as devices for the computation of functions rather than for the generation of sets of

strings, as we have described them. The two points of view are completely equivalent, however. Thus, as previously indicated, we can regard a Turing machine as representing the (partial) function that maps x into y whenever with input x it computes until it accepts x (i.e., returns to S_0 for the first time), at which point its tape contains exactly y (bounded by #'s). Equivalently, we can add to the alphabet a new symbol σ and can regard a Turing machine as representing the relation that holds between x and y just in case it accepts the string $x\sigma y$, and as representing the function that maps x into y , just in case this relation is a function. Either way, the same functions are representable. A two-output Turing machine, for example, could be regarded as a computable partial function that maps the set Σ that it accepts into the integer 1 and the set $\Sigma^* \subset \Sigma$ that it rejects into the integer 2. The specific terms in which we have described Turing machines and other automata, though equivalent to the usual ones, are adapted specifically to the purposes of this chapter.

2. UNRESTRICTED REWRITING SYSTEMS

We are now ready to investigate the interconnections of the theory of grammar (Chapter 11, Secs. 3 to 6) and the theory of automata (Sec. 1) and to study the capacity and formal properties of various kinds of grammars. Unfortunately, this survey must be restricted to a discussion of constituent-structure grammars rather than transformational grammars and (with the exception of Sec. 4.6) to the question of the enumeration of sets of sentences rather than sets of structural descriptions. As noted previously, the reason is, simply, that little of any significance is known concerning those more interesting but much more difficult questions.

In Chapter 11, Sec. 3, we introduced a class of grammars based on rewriting rules $\phi \rightarrow \psi$, where ϕ and ψ are strings of symbols of a finite vocabulary V . In this section we shall consider only these systems, presupposing the notions defined in Chapter 11, Sec. 4, and following the notational conventions established there. In particular, we assume to be given a universal terminal vocabulary V_T and a universal nonterminal vocabulary V_N disjoint from V_T , where $V = V_T \cup V_N$.

When no further constraints are placed on the kinds of rewriting rules that are available, the grammars defined in Chapter 11, Sec. 4, can be called *unrestricted rewriting systems*. The problem of relating these unrestricted systems to the theory of automata, as outlined in Sec. 1, is quite simple. In fact, any Turing machine can be represented directly as an unrestricted rewriting system and conversely. We can state this fact as follows:

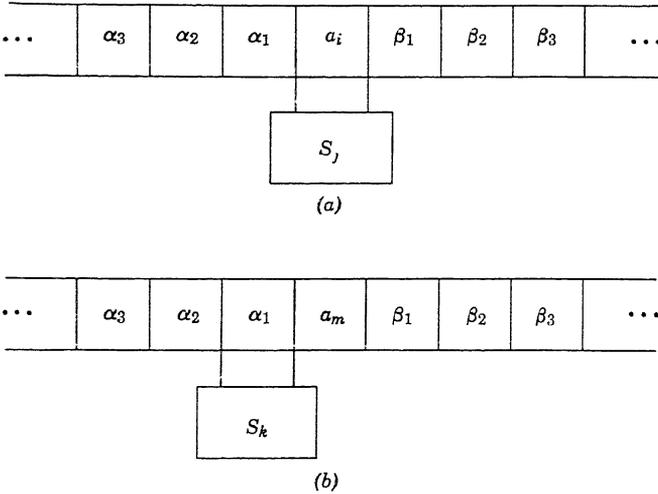


Fig. 8. Successive tape-machine configurations of a Turing machine.

Theorem 7. *L* is a terminal language generated by an unrestricted rewriting system if and only if *L* is a recursively enumerable set of strings $\{\#x_1\#, \#x_2\#, \dots\}$, where x_i contains no occurrences of #.

For a proof of this result, due to Post, see Davis (1958, Chapter 6, Sec. 2). The basic idea of the proof is that successive tape-machine configurations of a Turing machine can be determined by rewriting rules $\phi \rightarrow \psi$. Suppose, for example, that a Turing machine *M* has the rule $(i, j, k, -1, m)$, indicating that when it is in state S_j scanning the symbol a_i it replaces a_i by a_m and switches to state S_k as the tape moves right one space. At one moment, then, the machine will be as indicated in Fig. 8a and at the next moment as indicated in Fig. 8b, where $\alpha_1, \alpha_2, \dots, \beta_1, \beta_2, \dots$ are symbols on the tape. Consider now a vocabulary *V* consisting of the alphabet *A* of the Turing machine *M*, the symbol #, the symbol *S*, and symbols designating the states of *M*. In the present example the configuration of Fig. 8a can be represented in this vocabulary by the string of symbols (9) and the configuration of Fig. 8b by the string (10), in which the state symbol appears to the left of the symbol that is currently being scanned by the machine:

$$\dots \alpha_3 \alpha_2 \alpha_1 S_j a_i \beta_1 \beta_2 \beta_3 \dots \tag{9}$$

$$\dots \alpha_3 \alpha_2 S_k \alpha_1 a_m \beta_1 \beta_2 \beta_3 \dots \tag{10}$$

The transition of the Turing machine *M* from String 9 to String 10 can be accomplished by a rewriting rule:

$$\alpha_1 S_j a_i \rightarrow S_k \alpha_1 a_m. \tag{11}$$

It is possible to demonstrate that with a finite number of such rewriting rules the entire behavior of M can be unambiguously represented.

In Sec. 1.8 we described the behavior of a Turing machine in the following way: on its tape appears a finite string ϕ of symbols of its alphabet A (not containing $\#$) with an infinite string of $\#$'s to the left and to the right of ϕ . It begins operation in state S_0 , scanning the leftmost symbol of ϕ , and continues until it blocks or returns to S_0 . In the latter case we say that it accepts (generates) the string $\#\phi\#$ (it may, of course, continue to compute indefinitely). Suppose that we now modify the description in the following way. When a Turing machine M is about to return to state S_0 , it moves the tape instead until the reading head is scanning the rightmost square not containing $\#$. In this square it prints $\#$ and moves the tape right, again printing $\#$ and moving the tape right, etc., until it reaches the last square not containing $\#$, where it prints the symbol S , at which point it blocks in a designated final state. This terminal routine is easily arranged. With this modification, the set of generated languages and the character of the rules is left unchanged. The machine M accepts (generates) $\#\phi\#$, just in case, given the input ϕ , it computes until it blocks with the tape containing all $\#$'s except for a single occurrence of S . Furthermore, at each step of the computation there is on the tape a string ϕ containing no occurrences of $\#$, flanked on both sides by infinite strings of $\#$.

In the foregoing manner we can completely describe the behavior of such a Turing machine M by a particular set Σ of rewriting rules that will convert a string $\#S_0\phi\#$ to $\#S\#$ just in case M accepts $\#\phi\#$. Because of the deterministic character of M , the set Σ is *monogenic*; that is to say, given a string χ , there is at most one string ψ that can result from application of Σ to χ . Now consider the set of rules Σ' containing $\psi \rightarrow \chi$ just in case $\chi \rightarrow \psi$ is in Σ , and containing also a final rule $\#S_0 \rightarrow \#$. This set Σ' of rules is an unrestricted rewriting system. A $\#S\#$ -derivation of this system will terminate in $\#\phi\#$ just in case M accepts $\#\phi\#$.

From Theorem 7 we see that unrestricted rewriting systems are universal. If a language can be generated at all by what in the intuitive sense is a finitely storable, well-defined procedure, it can be generated by a grammar of this type. However, such systems are of little interest to us in the present context. In particular, there is no natural and uniform method to associate with each terminated derivation a P -marker of the desired kind for its terminal string. It is in this sense that an arbitrary Turing machine, or an unrestricted rewriting system, is too unstructured to serve as a grammar. By imposing further conditions on the grammatical rules, we arrive at systems that have more linguistic interest but less generative power. As remarked in Sec. 1.8, a particular Turing machine can be regarded as nothing more or less than a program of a perfectly arbitrary kind for a

digital computer with potentially infinite memory. Obviously, a computer program that succeeded in generating the sentences of a language would be, in itself, of no scientific interest unless it also shed some light on the kinds of structural features that distinguish languages from arbitrary, recursively enumerable sets. If all we can say about a grammar of a natural language is that it is an unrestricted rewriting system, we have said nothing of any interest. (See Chomsky, 1961, 1962b, for further discussion).

The most restrictive condition that we shall state will limit grammars to devices with the generative capacity of strictly finite automata. We shall see that these devices cannot, in principle, serve as grammars for natural languages. Consequently we are interested primarily in devices with more generative capacity than finite automata but that are more structured (and, presumably, have less generative capacity) than arbitrary Turing machines. In other words, we shall be concerned with devices that fall into the general area of restricted-infinite automata.

3. CONTEXT-SENSITIVE GRAMMARS

Suppose we take a system G that meets all the requirements defining an unrestricted rewriting system and impose on it the following further condition:

Condition 1. *If $\phi \rightarrow \psi$ is a rule of G , then there are nonnull symbols $a_1, \dots, a_m, b_1, \dots, b_n$, where $m \leq n$, such that $\phi = a_1 \dots a_m$ and $\psi = b_1 \dots b_n$.*

In brief, Condition 1 requires that if $\phi \rightarrow \psi$ is a rule of the grammar then ψ is not shorter than ϕ . A grammar meeting *Condition 1* we call a *type 1 grammar*.

Henceforth, for each Condition i that we establish we shall call the grammars meeting it *type i grammars*; a language generated by a type i grammar we call a *type i language*. An unrestricted rewriting system we call a *type 0 grammar*. The conditions that we shall consider are increasingly strong; that is to say, for each i a type $i + 1$ grammar will also satisfy the defining condition for a type i grammar, but some type i grammars will not qualify as type $i + 1$ grammars.

Condition 1 imposes an essential limitation on generative capacity. Since, in derivations of type 1 grammars, each line must be at least as long as the preceding line, the following theorem is obvious:

Theorem 8. *Every type 1 language is recursive.*

In fact, given a type 1 grammar G and a string x , only a finite number of derivations (those whose final lines are not longer than x) need be tested to determine whether G generates x . Not all recursive languages are

generated by type 1 grammars, however, as can be shown by a straightforward diagonal argument.

Although type 1 grammars generate only recursive sets, there is a certain sense in which they come close to generating arbitrary recursively enumerable sets. In order to simplify the discussion of this matter, we restrict ourselves to sets of positive integers (without loss of generality, since any set of finite strings can be effectively coded into a set of integers).

Recall once again the characterization of a Turing machine given in Sec. 1.8. Here we consider Turing machines with the alphabet $\{1, e\}$, where e is the identity element (i.e., a square containing e is regarded as blank and replacement of 1 by e constitutes erasure). We can assume that each particular machine M operates in the following manner: given the *input* sequence 1^i (i.e., i successive occurrences of 1 flanked by infinite strings of #), where $i \geq 1$, M begins to compute in its initial state S_0 while scanning the leftmost occurrence of 1, and continues until it terminates in a designated final state S_F . At this point the tape will contain the string ϕ flanked by #'s, where ϕ contains j occurrences of the symbol 1 and k occurrences of e . We can construct M so that it will not enter state S_F unless $j \geq 1$, and we can assume without loss of generality that, at termination in S_F , M is scanning the leftmost symbol distinct from # and that all occurrences of 1 precede all occurrences of e (that is to say, we can easily add to each M a component that will automatically convert it to this final configuration when it terminates). Thus the *output* of M , if M ever reaches state S_F when computing with the input 1^i , will be the string $1^j e^k$ ($j \geq 1$, $k \geq 0$). We have already observed (in Sec. 1.2) that M may never terminate for certain (or all) inputs and that there is no algorithm for determining from the rules of M whether it will terminate with a particular input or even whether there is some input for which it will terminate. If M does terminate with the output $1^j e^k$, given input 1^i , we say that M maps the integer i into the integer j . This description is quite general, and any Turing machine that represents a partial function (i.e., a function that may not be defined for certain elements of its domain) mapping positive integers into positive integers can be described in this way. It is well known that the range of a Turing machine, so described, is a recursively enumerable set and that each recursively enumerable set is the range of some such Turing machine.

Observe that such a Turing machine never writes the symbol #, although it can read and erase # (thus extending the portion of the tape available for computation). Consequently, if M terminates with the output $1^j e^k$ given the input 1^i , then $j + k \geq i$. Furthermore, we see immediately that if, in the manner indicated in Sec. 2, we construct a set of rewriting rules that mirror the behavior of M exactly, then this set of rules will constitute a monogenic type 1 grammar G . Condition 1 is satisfied because

the amount of tape actually being used never decreases. If M computes the output $1^j e^k$ from the input 1^i , then G will produce a $\#S_0 1^i \#$ -derivation terminating in the string $\#S_F 1^j e^k \#$ and conversely. Although M may enumerate an arbitrary recursively enumerable set (as the range of the function it represents), the set of outputs that it generates is recursive.

Suppose now that we select a Turing machine M , and associate with it a type 1 grammar G in the manner just described. Suppose further that we form G^* , which consists of the rules of G along with the following four rules for generating appropriate "initial strings" for G :

$$\begin{aligned} S &\rightarrow S_0 1; & S &\rightarrow S_x 1; \\ S_x &\rightarrow S_0 1; & S_x &\rightarrow S_x 1. \end{aligned} \tag{12}$$

These rules provide a terminated $\#S\#$ -derivation for each string $\#S_0 1^i \#$, where $i \geq 1$, and only for these strings. Consequently, the complete grammar G^* will provide a terminated $\#S\#$ -derivation for a string $\#\phi\#$ just in case $\phi = S_F 1^j e^k$ ($j \geq 1$), and, for some i , M terminates with the output $1^j e^k$, given the input 1^i . Thus, given an arbitrary Turing machine enumerating the recursively enumerable set Σ as its range, we can construct a type 1 grammar that will generate all and only the strings $\#S_F \phi \psi \#$, where $\phi \in \Sigma$, and ψ is a string of e 's (of length computable from ϕ , in fact, where $\phi \in \Sigma$). The problem of determining whether the range of an arbitrary Turing machine is null, finite, or infinite is known to be recursively unsolvable. Consequently, the corresponding problems for type 1 grammars are also recursively unsolvable.

Theorem 9. *There is no algorithm for determining whether an arbitrary type 1 grammar G generates the null set of strings, a finite set of strings, or an infinite set of strings.* (Scheinberg, 1960a.)

For just the same reason there is no algorithm for determining whether some particular string appears as a proper subpart of a line of an $\#S\#$ -derivation of G . Furthermore, since a grammar gives no terminated $\#S\#$ -derivations just in case it is equivalent to the grammar G_{null} with the single rule $S \rightarrow 1S1$, there is, by Theorem 9, no way to determine whether G is equivalent to G_{null} , and, in general:

Theorem 10. *There is no algorithm for determining whether two type 1 grammars are equivalent.* (Scheinberg, 1960a.)

In fact, for quite a range of problems unsolvable for Turing machines we can find an analogous problem unsolvable for type 1 grammars. There is, in other words, little that one can tell about the deviations of language generated by a type 1 grammar (other than that the language is necessarily recursive) by systematic investigation of its rules.

Condition 1 is not sufficiently strong to permit the uniform assignment

of P -markers to sentences in the desired way. To guarantee this, as we noted in Chapter 11, Sec. 4, we must require that only one symbol be rewritten at a time (as well as certain other conditions which, as we noted, do not affect generative capacity). This observation leads us to consider a more restrictive condition:

Condition 2. *If $\phi \rightarrow \psi$ is a rule, then there are strings $\chi_1, \chi_2, A, \omega$ (where A is a single symbol and ω is not null) such that $\phi = \chi_1 A \chi_2$ and $\psi = \chi_1 \omega \chi_2$.* In a type 2 grammar each rule $\phi \rightarrow \psi$ (that is, $\chi_1 A \chi_2 \rightarrow \chi_1 \omega \chi_2$) can be regarded as asserting that A can be rewritten ω when in the context $\chi_1 \chi_2$, where χ_1 or χ_2 may, of course, be null. We refer to grammars meeting this condition as *context-sensitive grammars*. Rules of this kind are quite common in actual grammatical descriptions. They can be used to indicate selectional and contextual restrictions on the choice of certain elements or categories, as observed in Chapter 11. In a context-sensitive grammar we can identify the class V_N of nonterminal symbols as the class containing exactly those symbols A such that the grammar contains a rule $\chi_1 A \chi_2 \rightarrow \chi_1 \omega \chi_2$. We shall henceforth assume this convention.

Clearly, then, every type 2 grammar is a type 1 grammar and not conversely. Nevertheless, Condition 2 does not restrict generative capacity. **Theorem 11.** *If G is a type 1 grammar, then there is a type 2 grammar G' that is weakly equivalent to G .*

The proof, which is perfectly straightforward, can be found in Chomsky (1959a).

Since the correspondence given by Theorem 11 is effective, it follows at once that the undecidable problems concerning type 1 grammars remain undecidable when we restrict ourselves to context-sensitive grammars.

Theorem 12. *There is no algorithm for determining, given two context-sensitive grammars, whether these grammars are equivalent, whether either generates a null, finite, or infinite set of strings, or whether an arbitrarily selected string appears as part of a line of a #S#-derivation of either grammar or as part of a sentence of the generated language.*

Here again we find that little of a general nature can be discovered by systematically studying the rules of these systems.

Theorem 12 has an important consequence for the theory of constituent-structure grammar. Any theory of grammar must provide a general method for assigning structural descriptions to sentences, given a grammar; and, in the case of constituent-structure grammars, this can be done in a natural way only if each such grammar meets the Condition C that there are no symbols A, B and no string ω such that $\phi A B \psi, \phi A \omega B \psi$ are successive lines of a derivation of a terminal string (cf. Footnote 3, Chapter 11, Sec. 4). Hence it is reasonable to require of a well-formed constituent-structure

grammar, in addition to the conditions already given, that it meet Condition C. It then follows immediately from Theorem 12 that well-formedness is an undecidable property of context-sensitive grammars. This fact is sufficient to rule out the theory of context-sensitive grammar, in its present form, as a possible theory of grammar. Clearly, a general theory of grammar must provide a recursive class of well-formed grammars as potential candidates for specification of some natural language; that is, there must be an algorithm for determining whether a particular set of rules constitutes a well-formed grammar in accordance with this theory. The theory of context-sensitive grammar, in its present form, does not meet this requirement (though it can be modified in such a way as to meet it. See Footnote 3, Chapter 11, Sec. 4). Note that the theory of transformational grammar is not subject to this difficulty if, as suggested in Chapter 11, Sec. 5, its constituent-structure component generates only a finite set of strings (similarly, if it generates an infinite set of a sufficiently restricted kind, a restriction that is feasible if transformational devices are available to extend generative capacity).

In Chapter 11, Sec. 3, we used as illustrations three artificial languages, L_1 , L_2 , and L_3 , all with the vocabulary $\{a, b\}$, and we demonstrated that L_1 and L_2 can have what we are now calling context-sensitive grammars. (In fact, they have grammars meeting Condition 2 where χ_1 and χ_2 are null). For L_3 , however, we gave a simple grammar that was not an unrestricted rewriting grammar at all. We know, of course, that an unrestricted rewriting grammar for L_3 must exist, since it is obviously a recursively enumerable—in fact, a recursive set. It is interesting to observe, however, that L_3 can indeed be generated by a context-sensitive grammar, although a much more complicated one is required for L_3 than for L_1 or L_2 .

This follows from a general property of context-sensitive grammars, to which we now turn. Consider the following set of rules:

$$\begin{aligned}
 R1: & CDA \rightarrow CE\bar{A}A; CDB \rightarrow CE\bar{B}B \\
 R2: & CE\bar{A} \rightarrow \bar{A}CE; CE\bar{B} \rightarrow \bar{B}CE \\
 R3: & E\alpha\beta \rightarrow \beta E\alpha \\
 R4: & E\alpha\# \rightarrow D\alpha\# \\
 R5: & \alpha D \rightarrow D\alpha \\
 R6: & \bar{A} \rightarrow A; \bar{B} \rightarrow B.
 \end{aligned} \tag{13}$$

In these rules, the variables α and β range over $\{A, B, F\}$. So, for example, Rule R5 is actually to be regarded as the set of three rules: $AD \rightarrow DA$; $BD \rightarrow DB$; $FD \rightarrow DF$.

Given a string $CD\phi F\#$, where ϕ is any string of A 's and B 's, the rules in Example 13 will apply in a unique order (except for some freedom in the case of R6) to produce, finally,

$$\phi CDF\phi\#, \quad (14)$$

at which point none of these rules applies. In short, Rules 13 describe a copying machine. Given such a copying machine, it is not a difficult task to use it as the basis for a grammar that will generate L_3 . Moreover, since all of these rules meet Condition 1, it will be a type 1 grammar. From Theorem 11, therefore, the following theorem is derived.

Theorem 13. *There is a context-sensitive grammar G that generates L_3 .* (Chomsky, 1959a.)

This grammar will be considerably more complex than the grammar (12) proposed in Chapter 11, Sec. 3, since, in particular, it must include a set of rules which has the effect of the copying machine of Rules 13. The grammar (12) in Chapter 11, Sec. 3, can easily be redescribed as a transformational grammar. Here, then, is an elementary, artificial example of the simplification that can often be achieved by extending the scope of grammatical theory to include transformational grammars of the kind described in Chapter 11, Sec. 5.

It is important to observe that the ability of a context-sensitive grammar to generate such languages as L_3 represents a defect rather than a strength. This fact becomes clear when we observe how a context-sensitive copying device actually functions. The basic point is that it is possible to achieve the effect of a permutation $AB \rightarrow BA$ within the limitations of a context-sensitive grammar; but, when, with a sequence of context-sensitive rules, we succeed in rewriting AB as BA , we find that in the associated P -marker the symbol B of BA is of type A (i.e., is traceable back to A in the associated tree) and the symbol A of BA is of type B . For example, if we were to use such rules to convert *John will arrive* to *will John arrive*, we would be forced to assign a P -marker to *will John arrive* that would provide the structural information that *will* in this sentence is a noun phrase (being traceable back to the symbol NP that dominates *John*) and that *John* is a modal auxiliary, contrary to our intention. Were we to attempt to construct a context-sensitive grammar for English, there would be no natural way to avoid this totally unacceptable consequence. (Note that if *will John arrive* is derived from *John will arrive* by a grammatical transformation, in the manner described in Chapter 11, Sec. 5, this counterintuitive consequence does not result).

This observation suggests that it might be important to devise a further condition on context-sensitive grammars that would exclude permutations but would still permit the use of rules to limit the rewriting of certain

symbols to a specific context. A very natural restriction that would have this effect is the following, which has been proposed by Parikh (1961):
 Condition 3. G is a type 2 grammar containing no rule $\chi_1 A \chi_1 \rightarrow \chi_1 \omega \chi_2$, where ω is a single nonterminal symbol (i.e., $\omega \in V_N$).

In a type 3 grammar no symbol can be rewritten as a single nonterminal symbol in any context. With this restriction, it is impossible to construct a sequence of rules with the effect of a simple permutation $AB \rightarrow BA$. Consequently, the copying machine described by Rules 13 cannot be constructed, and the unwanted consequences do not ensue. Presumably, L_3 cannot be generated by a type-3 grammar. Certainly it cannot be by the method previously described.

Condition 3, as it stands, is too strong to be met by actual grammars of natural languages, but it can be revised, without affecting generative capacity, to be perhaps not unreasonable for the construction of grammars of languagelike systems. Suppose that we allow the grammar G to contain a rule $\chi_1 A \chi_2 \rightarrow \chi_1 \alpha \chi_2$ only when α is either terminal (as in Condition 3) or when α dominates only a finite number of strings in the full set of derivations (and P -markers) constructible from G . This essentially amounts to the requirement that if a category is divided into subcategories these subcategories are not phrase types but word or morpheme classes. To the extent that systems of the kind we are now discussing are at all useful for grammatical description, it seems likely that the particular subclass meeting this condition will in fact suffice.

Only one nontrivial property of type 3 grammars is known, namely, that stated in Theorem 14. This class of grammars merits further study, however. It seems that Condition 3 provides a reasonably adequate formalization of the set of linguistic notions involved in the richest varieties of immediate constituent analysis.

4. CONTEXT-FREE GRAMMARS

Consider next the class of grammars meeting the following condition:
 Condition 4. If $\phi \rightarrow \psi$ is a rule, then ϕ is a single (nonterminal) letter and ψ is nonnull.

Thus each rule of the grammar states that a certain nonterminal symbol can be rewritten as a string of symbols, irrespective of the context in which it occurs. A grammar meeting Condition 4 we call a *context-free grammar*. A language generated by a context-free grammar is called a *context-free language*. (Recall that although the rules of a context-free grammar are applicable irrespective of context nevertheless there can be, and usually are, strong contextual constraints among the elements of the terminal string.)

Concerning context-free grammars quite a bit is now known. We shall sketch the major results here, referring occasionally to more detailed presentations elsewhere for proofs and further discussion.

It is immediately clear that if in the statement of Condition 4 we drop the requirement that ψ be nonnull then the generative capacity of the class of context-free grammars is unchanged (except that the "empty" language $\{e\}$ can be generated). See Bar-Hillel, Perles, & Shamir (1960, Sec. 4). We can also, without affecting generative capacity, impose the requirement that there be no rule of the form $A \rightarrow B$ in the grammar (cf. Chomsky, 1959a), so that context-free languages also meet Condition 3.

Although all context-free (type 4) languages are type 3 languages, the converse is not true.

Theorem 14. *The language $\#ac^{nf^2+4^n}d^nb\#$ is a type 3 language that cannot be generated by a context-free grammar.*

The proof is due to Parikh (1961). It is much easier to find examples of type 2 languages (languages generated by the full class of context-sensitive grammars) that cannot be generated by any context-free grammar. In particular, among the languages L_1, L_2, L_3 of the preceding section, although L_1 and L_2 are context-free languages (cf. Chapter 11, Sec. 3), L_3 is clearly not.

Theorem 15. *The language L_3 and the language $\{a^n b^n a^n\}$ are type 2 languages for which there exists no context-free grammar. (Cf. Chomsky, 1959a; Scheinberg, 1960b; Bar-Hillel, Perles, & Shamir, 1960.)*

We can obtain the P -marker (cf. Chapter 11, Sec. 3) of a string generated by a context-free grammar G directly by considering a new context-free grammar G' with the vocabulary of G and the new terminals $] \text{ and } [_A$, where A is any nonterminal of G . Where G has the rule $A \rightarrow \phi$, G' has the rule $A \rightarrow [_A \phi]$. G' will generate a string x containing brackets that indicate the constituent structure of the corresponding string y , formed from x by deleting brackets, that would be generated by G . Under special circumstances, the right bracket $]$ can be deleted without ambiguity, giving a kind of parenthesis-free notation. Clearly the bracketing imposed by G' on a terminal string of G corresponds exactly to that given by the tree-diagrams used in Chapter 11. Thus we can regard the structural description of a string x as a string ϕ in the vocabulary V^* containing V_T and the new symbols $] \text{ and } [_A$, for each $A \in V_N$.

By the same method we can obtain the P -marker of a context-sensitive grammar if it meets Condition C of p. 363. If a context-sensitive grammar meets Condition C, we will say that it is *well formed*. A context-sensitive grammar G is thus well formed if and only if it has the following property: if ϕ, ψ are successive lines of a terminated $\#S\#$ -derivation of G , there exist unique strings $\alpha \in V_N$ and χ_1, χ_2, ω such that $\phi = \chi_1 \alpha \chi_2$ and

$\psi = \chi_1\omega\chi_2$. Extending the vocabulary V to include brackets, as above, let us define $d(\phi)$ (read “debracketization of ϕ ”), for any string ϕ in this extended vocabulary, as the string obtained by deleting all occurrences of brackets (with their subscripts) from ϕ . We can then define a *strong ϕ -derivation* of ψ as a sequence ϕ_1, \dots, ϕ_n such that $\phi_1 = \phi$, $\phi_n = \psi$, and for each $i > n$ there are strings $\omega_1, \dots, \omega_5$ and a symbol α such that $\phi_i = \omega_1\omega_2\alpha\omega_3\omega_4$, $\phi_{i+1} = \omega_1\omega_2[\omega_5]\omega_3\omega_4$, $d(\omega_5) = \omega_5$ and $d(\omega_2\alpha\omega_3) \rightarrow d(\omega_2\omega_5\omega_3)$ is a rule of G . If D is a strong $\#S\#$ -derivation of ψ in G and $d(\psi)$ is a string on V_T , then $d(\psi)$ is a terminal string generated by G and ψ can be taken as the P -marker assigned to it uniquely by the derivation D' , each line of which is the debracketization of the corresponding line of D . Furthermore, for each $\#S\#$ -derivation D of a string x in G there is a corresponding strong $\#S\#$ -derivation that terminates in a string which can be taken as the P -marker uniquely assigned by D to x . Thus for well-formed context-sensitive grammars we have a precise definition of generation of strong P -markers.

As we have noted above, well-formedness, in the sense just defined, is not a decidable property of context-sensitive grammars. We might define a decidable property of well-formedness for such grammars in the following way: G is well formed if it contains no rule of the form $\phi AB\psi \rightarrow \phi A\omega B\psi$. In this case a strong derivation might not be uniquely determined by a weak derivation (as it is if the former condition of well-formedness is met), but it would still be uniquely determined by a weak derivation together with the sequence of rules used to form it (neither of these being dispensable, in general). In fact, as we noted in Chapter 11, Sec. 4, there is no difficulty in imposing effective conditions that eliminate all such indeterminacy, without affecting weak generative capacity (and affecting strong generation only by the imposition of some additional and otherwise irrelevant categorization). These further conditions are, however, rather *ad hoc*.

4.1 Special Classes of Context-Free Grammars

In this section we shall consider various subclasses of the set of context-free grammars that are defined by additional restrictions on the set of rules. Recall that each rule is of the form $A \rightarrow \phi$, where A is a single nonterminal letter and ϕ is a nonnull string. We shall continue to use the notational convention of Chapter 11, Sec. 4. Recall that $\phi \Rightarrow \psi$ just in case there is a ϕ -derivation of ψ . Furthermore, the nonterminal vocabulary V_N consists of exactly those symbols A that appear on the left of a rule $A \rightarrow \phi$ in the grammar.

We call a rule *linear* if it is of the form $A \rightarrow xBy$. It is *right-linear* if it is of the form $A \rightarrow xB$; *left-linear* if it is of the form $A \rightarrow Bx$. A rule is *terminating* if it is of the form $A \rightarrow x$. In terms of these notions, we define several kinds of grammars.

Definition 6. A grammar G is

- (i) *linear* if each nonterminating rule is linear; in particular, if each nonterminating rule is either right-linear or left-linear; (ii) *one-sided linear* if either each nonterminating rule is right-linear or each nonterminating rule is left-linear; (iii) *meta-linear* if all nonterminating rules are linear or of the form $S \rightarrow \phi$ and, furthermore, there is no rule $A \rightarrow \phi S \psi$ for any A, ϕ, ψ ; (iv) *normal* if all nonterminating rules are of the form $A \rightarrow BC$ and all terminating rules are of the form $A \rightarrow a$; (v) *sequential* if its nonterminal vocabulary can be ordered as A_1, \dots, A_n in such a way that for each i, j , if $A_i \Rightarrow \phi A_j \psi$ then $j \geq i$.

In the case of a linear grammar, if ϕ is a line of an $\#S\#$ -derivation, then ϕ contains at most one nonterminal symbol. There is, in other words, only one point at which a derivation can branch at any step. When the first terminating rule applies, the derivation terminates in a terminal string. In a meta-linear grammar there is a bound n on the number of points at which a derivation can branch. This bound is given by the longest rule in which S appears—maximal, that is to say, in the number of nonterminal symbols that appear. When n terminating rules have applied, the derivation terminates in a terminal string.

A one-sided linear grammar is nothing other than a finite automaton, in the sense of Sec. 1.2 (cf. Chomsky, 1956). This is clear in the case of a one-sided linear grammar G with only right-linear (and terminating) rules. We can assume, without loss of generality, that each linear rule of G is of the form $A \rightarrow aB$, where B is not the initial symbol of G , and that each terminating rule is of the form $A \rightarrow a$. Let A_1, \dots, A_n be the nonterminal symbols of G , where A_1 is the initial symbol. We can associate with G the finite automaton F with the same nonterminal vocabulary as G and with states $\bar{A}_1, \dots, \bar{A}_n, \bar{A}_1$ being the initial state. We form the rules of F in the following way. If $A_i \rightarrow aA_j$ is a rule of G , then the triple $(a, \bar{A}_i, \bar{A}_j)$ is a rule of F (interpreted as the instruction to F to switch from state \bar{A}_i to state \bar{A}_j when reading the input symbol a). If $A_i \rightarrow a$ is a rule of G , then the triple $(a, \bar{A}_i, \bar{A}_1)$ is a rule of F . Thus F and G terminate after having generated the same terminal string. Similarly, the rules of a finite automaton immediately give a grammar with only right-linear or terminating rules.

Since if L is a regular language, then L^* consisting of the mirror-images of the strings of L (i.e., containing $a_n \dots a_1$ whenever $a_1 \dots a_n \in L$) is also a regular language, it is clear that each one-sided linear grammar represents

a finite automaton and that each finite automaton can be represented as a one-sided linear grammar.

A *normal grammar* is the kind usually considered in discussions of immediate constituent analysis in linguistics. The terminating rules $A \rightarrow a$ constitute the *lexicon* of the language, which is sharply distinguished from the set of grammatical rules $A \rightarrow BC$, each of which gives binary constituent breaks.

The notion of a *sequential grammar* is motivated by the ease with which the output of such a device can be mechanically computed. Once the rules developing a certain nonterminal symbol A have been applied, thus eliminating all occurrences of A in the last line of the derivation under construction, we can be sure that A will not recur in any later lines of the derivation. A restriction of this sort (but more general, involving also transformational rules) has been suggested and studied in a linguistic application by Matthews (1962).

Definition 7. *Let*

$$\begin{aligned} \lambda &= \{L \mid L \text{ is generated by a linear grammar}\} \\ \lambda_1 &= \{L \mid L \text{ is generated by a one-sided linear grammar}\} \\ \lambda_m &= \{L \mid L \text{ is generated by a meta-linear grammar}\} \\ \nu &= \{L \mid L \text{ is generated by a normal grammar}\} \\ \sigma &= \{L \mid L \text{ is generated by a sequential grammar}\} \\ \gamma &= \{L \mid L \text{ is generated by a context-free grammar}\}. \end{aligned}$$

Thus λ_1 , in particular, is the class of regular languages, as we have just observed. The systems defined in Def. 6 are related to one another in the following way, from the point of view of generative capacity.

Theorem 16. (i) $\lambda_1 \subsetneq \lambda \subsetneq \lambda_m \subsetneq \gamma$ (Chomsky, 1956; Schützenberger & Chomsky, 1962); (ii) $\nu = \gamma$ (Chomsky, 1959a); (iii) $\lambda_1 \subsetneq \sigma \subsetneq \gamma$ (Ginsburg & Rice).

The languages L_1 and L_2 are generated by linear grammars, but, as we observed in Sec. 1.1, cannot be generated by finite automata (one-sided linear grammars). The product $L_1 \cdot L_2 = \{y \mid y = xz, x \in L_1, \text{ and } z \in L_2\}$ of languages L_1 and L_2 of λ is in λ_m but not, in general, in λ . The set of well-formed formulas of sentential calculus in the so-called Polish notation is an example of a language that has no meta-linear grammar but can be generated by the context-free grammar

$$S \rightarrow CSS, \quad S \rightarrow NS, \quad S \rightarrow V, \quad V \rightarrow V', \quad V \rightarrow p, \quad (15)$$

where the sentential letters are p, p', p'', \dots ; C is the sign of the conditional, N of negation.

The languages $L_1 (= \{a^n b^n\})$ and $L_2 (= \{xx^* \mid x \text{ a string on } \{a, b\} \text{ and } x^* \text{ the reflection of } x\})$ are in σ but not λ_1 . An example of a language in γ but not in σ is the language with vocabulary $\{a, b, c, d\}$ and containing the sentence

$$a^{n_{2k-1}} d \dots db^{n_2} da^{n_1} ca^{n_1} db^{n_2} d \dots b^{n_{2k-2}} da^{n_{2k-1}} \quad (16)$$

(which is symmetrical about c) for each sequence $(k, n_1, \dots, n_{2k-1})$ of positive integers. This language is generated by the rules

$$\begin{aligned} S &\rightarrow adAda, & S &\rightarrow aSa, & S &\rightarrow aca, \\ A &\rightarrow bAb, & A &\rightarrow bdSdb \end{aligned} \quad (17)$$

(which are, in fact, linear), but it is not generated by a sequential grammar.

Since normal grammars can generate any context-free language, the common restriction to binary constituent breaks and to a separate lexicon does not limit generative capacity beyond context-free grammars (though, of course, this restriction severely limits the system of structural descriptions that can be generated, i.e., it limits strong generative capacity).

4.2 Context-Free Grammars and Restricted-Infinite Automata

We have observed that context-free and context-sensitive grammars of the various kinds we have been considering are richer in generative capacity than finite automata, though weaker than unrestricted rewriting systems (Turing machines). In particular, we have found languages that are not regular but that can be generated by linear context-free grammars (even with a single nonterminal); we have, on the other hand, noted that even context-sensitive grammars can generate only recursive sets and not all of these. Grammars of the kind we are now considering belong to the category of restricted-infinite automata (cf. Secs. 1.3 to 1.7). In the case of context-free languages we find that each can be accepted by a linear-bounded automaton of the special kind that uses only pushdown storage (PDS) (cf. Secs. 1.4 to 1.6) and, furthermore, that only context-free languages can be accepted by such devices.

To see this, we restrict attention to normal grammars, without loss of generality [cf. Theorem 16(ii)]. We can also clearly assume, with no loss of generality, that if $A \rightarrow BC$ is a rule of a normal grammar then $B \neq C$. Given such a normal grammar G , we can construct a PDS automaton that accepts the language $L(G)$ generated by G in the following way. For each nonterminal A of G , the control unit of M will have two states designated A_l and A_r . For each rule $A \rightarrow a$ of G , M will have the instruction:

$$(a, A_l, e) \rightarrow (A_r, e). \quad (18)$$

For each rule $A \rightarrow BC$ of G , M will have the instructions:

$$\begin{aligned} (e, A_l, e) &\rightarrow (B_l, A), \\ (e, B_r, e) &\rightarrow (C_l, e), \\ (e, C_r, A) &\rightarrow (A_r, \sigma), \end{aligned} \tag{19}$$

where e is the identity element. M is thus, in general, a nondeterministic PDS automaton. Its initial state we designate Σ , where Σ does not appear in G . We assume that the device has the instructions $(e, \Sigma, \sigma) \rightarrow (S_l, e)$ and $(e, S_r, \sigma) \rightarrow (\Sigma, \sigma)$, where S is the initial symbol of G , allowing it to go from Σ to S_l and from S_r to Σ , erasing σ and terminating.

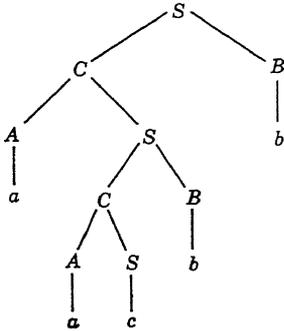
M accepts a string x generated by G by simply tracing systematically through the tree diagram representing the derivation of x by G , from top to bottom and from left to right. To illustrate, consider the grammar G with the rules

$$S \rightarrow CB, \quad C \rightarrow AS, \quad A \rightarrow a, \quad B \rightarrow b, \quad S \rightarrow c \tag{20}$$

generating the language

$$a^n cb^n, \tag{21}$$

Fig. 9. A typical derivation of a sentence in language (21).



with such typical derivations as that shown in Fig. 9 for the string $aacbb$. The corresponding PDS device M will have Instructions 22 corresponding to Instruction 18 and Instructions 23 corresponding to Instructions 19:

$$\begin{aligned} (a, A_l, e) &\rightarrow (A_r, e), \\ (b, B_l, e) &\rightarrow (B_r, e), \\ (c, S_l, e) &\rightarrow (S_r, e), \\ (e, S_l, e) &\rightarrow (C_l, S), \\ (e, C_r, e) &\rightarrow (B_l, e), \\ (e, B_r, S) &\rightarrow (S_r, \sigma), \\ (e, C_l, e) &\rightarrow (A_l, C), \\ (e, A_r, e) &\rightarrow (S_l, e), \\ (e, S_r, C) &\rightarrow (C_r, \sigma). \end{aligned} \tag{22}$$

In accepting the string $aacbb$ with the derivation in Fig. 9, M will compute in the following steps, in which column one represents the input tape, with

the scanned symbol in bold face, column two indicates the state of the control unit, and column three represents the contents of the storage tape.

| Input | Control Unit | PDS |
|-------------------|--------------|---------------|
| 1. aacbb | Σ | σ |
| 2. aacbb | S_l | σ |
| 3. aacbb | C_l | σS |
| 4. aacbb | A_l | σSC |
| 5. aacbb | A_r | σSC |
| 6. aacbb | S_l | σSC |
| 7. aacbb | C_l | σSCS |
| 8. aacbb | A_l | $\sigma SCSC$ |
| 9. aacbb | A_r | $\sigma SCSC$ |
| 10. aacbb | S_l | $\sigma SCSC$ |
| 11. aacbb | S_r | $\sigma SCSC$ |
| 12. aacbb | C_r | σSCS |
| 13. aacbb | B_l | σSCS |
| 14. aacbb | B_r | σSCS |
| 15. aacbb | S_r | σSC |
| 16. aacbb | C_r | σS |
| 17. aacbb | B_l | σS |
| 18. aacbb# | B_r | σS |
| 19. aacbb# | S_r | σ |
| 20. aacbb# | Σ | e |

(24)

Clearly M accepts all and only the strings generated by G , using its storage tape to store as much of the derivation of a generated string x as will be needed in later steps of the computation, as it processes x on its input tape. Furthermore, the same construction can obviously be carried out quite generally for any normal grammar. Observe also that the PDS automaton given by this construction is what in Sec. 1.4 we called a PDS automaton with restricted control. Hence we conclude the following: Theorem 17. *Given a context-free language L , we can construct a PDS automaton with restricted control that accepts L .*

Matthews has shown (Matthews, 1963a, b) that this result can be extended in part to context-sensitive grammars. Given a context-sensitive grammar G , we define a *left-to-right derivation* (a *right-to-left derivation*) as one meeting this condition: if (ϕ, ψ) are successive lines, then $\phi = xA\omega$ and $\psi = x\chi\omega$ (respectively, $\phi = \omega Ax$ and $\psi = \omega\chi x$). Thus only the leftmost (respectively, rightmost) nonterminal may be rewritten. Matthews shows that we can construct a PDS automaton M_{L-R} that will accept a string x if and only if there is a left-to-right derivation of x in G , and there is a

PDS automaton M_{R-L} that will accept a string x if and only if there is a right-to-left derivation of x in G . By Theorem 6, Sec. 1.6, it follows that the languages accepted by M_{L-R} and M_{R-L} are context-free. Consequently their union is context-free (see Theorem 20, Sec. 4.3). Thus, if we consider only the left-to-right and the right-to-left derivations of a context-sensitive grammar G , this grammar will generate a context-free language. Let us say that a context-sensitive grammar is *strictly context-sensitive* if it generates a noncontext-free language. We see, then, that a necessary condition for a grammar to be strictly context-sensitive is that some of its terminal strings have no derivations that are left-to-right or right-to-left.

It is not difficult to show that these observations continue to hold if we define a left-to-right derivation (a right-to-left derivation) as one in which the rewritten symbol of each line is no more than a bounded distance away from the leftmost (respectively, rightmost) nonterminal of this line. See Matthews (forthcoming). Thus a grammar will be strictly context-sensitive only if some of its terminal strings have only derivations which are neither left-to-right nor right-to-left in this extended sense.

Suppose that we say that D is an n -embedded derivation if n is the largest number such that D contains a pair of successive lines $xA\phi B\psi Cy$ and $xA\phi\omega\psi Cy$, where the shorter of ϕ, ψ is of length n . Thus in an n -embedded derivation there is some line in which the rewritten symbol is at a distance of n symbols away from either the leftmost or rightmost nonterminal of this line. One would conjecture that a necessary condition for a grammar to be strictly context-sensitive is that for each n it can generate some terminal strings only by derivations which are m -embedded, where $m > n$. Note, in particular, that in the derivations produced by the "copying device" of Example 13 there are lines in which the rewritten symbol is arbitrarily far from both the rightmost and leftmost nonterminal of these lines. In considering these questions, it is important to bear in mind that the question whether a context-sensitive grammar is strictly context-sensitive is undecidable, as has been observed by Shamir (1963).

We have already shown in Sec. 1.6 that corresponding to each PDS automaton M there is a transducer T with the following property: T maps x into a string y that reduces to \dot{e} , if and only if M accepts x . Consequently, we now see that, given a context-free grammar G , there is a transducer T that maps x into a string y that reduces to e , just in case G generates x . However, we can achieve a somewhat stronger result by carrying out the construction of T from G directly.

Let us define a *modified normal grammar* as a normal grammar that contains no pair of rules $A \rightarrow BC, D \rightarrow CE$ for any nonterminals A, B, C, D, E ; that is, in a modified normal grammar we can tell unambiguously, for each nonterminal, whether it appears on a left branch or a right branch

of a derivation; no nonterminal can appear on both a left and a right branch. Clearly, there is a modified normal grammar equivalent to each normal grammar, hence to each context-free grammar.

Suppose that we now apply a construction much like that of Instructions 18 and 19 to a modified normal grammar G , giving a transducer T . Corresponding to each nonterminal A of G , T will have two states A_l and A_r . In addition, T has the initial state Σ and the instructions $(e, \Sigma) \rightarrow (S_l, e)$ and $(e, S_r) \rightarrow (\Sigma, \sigma')$, where S is the initial symbol of G . The input alphabet of T is the terminal vocabulary V_T of G . Its output alphabet includes, in addition, a symbol a' for each $a \in V_T$, a pair of symbols A, A' for each nonterminal A of G , and σ, σ' . When $A \rightarrow a$ ($a \in V_T$) is a rule of G , T will have the instruction

$$(a, A_l) \rightarrow (A_r, Aaa'A'); \tag{25}$$

when $A \rightarrow BC$ is a rule of G , T will have the instructions

$$\begin{aligned} (e, A_l) &\rightarrow (B_l, A), \\ (e, B_r) &\rightarrow (C_l, e), \\ (e, C_r) &\rightarrow (A_r, A'). \end{aligned} \tag{26}$$

The transducer T so constructed has the essential property of the transducer associated with a PDS device by the construction presented in Sec. 1.6, namely, G generates x if and only if T maps x into a string y that reduces to e by successive deletions of pairs $\alpha\alpha'$. For example, when G is as in Example 20 and T has the input $\#aacbb\#$ (with the derivation of Fig. 9), it will compute in essentially the manner of M in (24), terminating with the string

$$\sigma SCAad'A'SCAad'A'Scc'S'C'Bbb'B'S'C'Bbb'B'S'\sigma', \tag{27}$$

which reduces to e , on its storage tape.

Let us now extend this notion of "reduction" and define K as the class of strings in the output alphabet A_O of T that reduce to e by successive cancellation of substrings $\alpha\alpha'$ or $\alpha'\alpha$ ($\alpha, \alpha' \in A_O$). We are thus essentially regarding α, α' as strict inverses in the free group \mathcal{G} with the generators $\alpha \in A_O$. But since the grammar G from which T was constructed was a modified normal grammar, the output of T can never, in fact, contain a substring $\alpha'x\alpha$, where x reduces to e . Hence this extension of the notion "reduction" is harmless, and under it the transducer T will still retain the property that G generates x if and only if T maps x into a string y that reduces to e .

We have assumed throughout (cf. Chapter 11, Sec. 4), as is natural, that the vocabulary V from which all context-free grammars are constructed is a fixed finite set of symbols, so that K is a particular fixed language in the

vocabulary V' containing V , σ , σ' and a symbol α' for each $\alpha \in V$. Let ϕ be the homomorphism (i.e., the one-state transduction) such that $\phi(\alpha) \rightarrow \alpha$ for $\alpha \in V_T$ and $\phi(x) = e$ for $x \notin V_T$. Let U be the set of all strings on V . Observe now that where G and T are as have been described and $L(G)$ is the language generated by G , we have, in particular, the result that $L(G) = \phi[K \cap T(U)]$.

It is a straightforward matter to construct a PDS automaton that will accept K ; consequently, by Theorem 6, Sec. 1.6, K is a context-free language. As we have observed in Sec. 1.6, $T(U)$ is a regular language. We shall see directly that the intersection of a context-free language with a regular language is a context-free language and that transduction carries context-free languages into context-free languages (Sec. 4.6). Given K and ϕ as in the preceding paragraph, let us define $\psi(L)$ for any language L as $\psi(L) = \phi(K \cap L)$. Summarizing the facts just stated, we have the following general observation:

Theorem 18. *For each regular language R , $\psi(R)$ is a context-free language; for each context-free language L there is a regular language R such that $L = \psi(R)$.*

Thus a context-free language is uniquely determined by the choice of a certain regular language (i.e., finite automaton), and each such choice produces a context-free language, given K , ϕ . This provides a simple algebraic characterization of context-free languages.

Theorem 18 can be extended immediately to the result that each context-free language L is given as the homomorphic image of the intersection of K with some 1-limited language D . (Recall that, as noted in Sec. 1.2, each regular language is the homomorphic image of some 1-limited language.) Furthermore, the various categories of context-free languages that we have defined are easily definable by imposition of simple conditions on D (cf. Schützenberger & Chomsky, 1962, for details). We know from Sec. 1.2 that regular languages consist of strings with a basically periodic structure. From the role of K in characterizing context-free languages, we see that, in a sense, symmetry of structure is the fundamental formal property of the strings of a context-free language (and the substrings of which they are constituted). We might say, rather loosely, that to the extent that the character of some aspect of serially ordered behavior is determined by conditions on contiguous parts (e.g., associative linkage), it is natural to regard the organism carrying out this behavior as essentially a limited automaton; to the extent that this behavior is periodic and rhythmic [e.g., in the case of the examples offered by Lashley (1951) in his critique of the "associative chain" theory], the organism is performing in the manner of a strictly finite automaton; to the extent that such behavior exhibits hierarchic organization and symmetries, the organism is performing in the

manner of a device whose intrinsic competence is expressed by a context-free grammar. Naturally this brief (and loose) classification does not exhaust the possibilities for complex, serially ordered, and integrated acts, and it is to be hoped that, as the theory of richer generative systems (in particular, for the case of language, transformational grammars) develops, deeper and more far-reaching formal properties of such behavior will be revealed and explained.

Note that String 27 is essentially the structural description of the input string $\#aacbb\#$ corresponding to Fig. 9. Specifically, String 27 becomes a structural description of the form described on p. 367 under the homomorphism f defined as follows: for $\alpha \in V_T$, $f(\alpha) = \alpha$ and $f(\alpha') = e$; for $\alpha \in V_N$, $f(\alpha) = [_{\alpha}$ and $f(\alpha') =]$. This amounts to replacing T with Instructions 25 and 26 by the transducer T' identical with T except that it never prints α' (for $\alpha \in V_T$) and that it prints $]$ instead of α' for each $\alpha \in V_N$. As an immediate corollary of Theorem 17, then, we have the following:

Theorem 19. *Given a context-free language L , we can construct a modified normal grammar G generating L and a transducer T with the following property: if G generates x with the structural description ϕ , then T maps x into ϕ ; if T maps x into ϕ and ϕ reduces to e under successive cancellation of substrings, $[_{\alpha} a]$, where $a \in V_T$, then G generates x with the structural description ϕ .*

The transducer T guaranteed by Theorem 19 is thus, in a sense, a "recognition routine" (i.e., a perceptual model) that assigns to arbitrary sentences their structural description with respect to G . It is not, however, a strictly finite recognition routine because of the condition that the output must reduce to e . We shall return (Sec. 4.6) to the problem of constructing an optimal, strictly finite recognition routine for context-free grammars. We know, as a result of this section, that there is a mechanical procedure for constructing a recognition routine with PDS corresponding to each normal context-free grammar, hence to each context-free language in at least one of its grammatical representations (and, one would conjecture, no doubt in all).

To summarize, we have the following results. There is a fixed homomorphism f such that for any regular language R we can find a 1-limited language L such that $R = f(L)$. There are fixed homomorphisms g_1, g_2 , such that given any context-free grammar G' generating $L(G')$ there is a modified normal grammar G generating the language $L(G) = L(G')$ and generating the set of structural descriptions $\Sigma(G)$, and there is a 1-limited language L such that $L(G) = g_1(K \cap L)$ and $\Sigma(G) = g_2(K \cap L)$, where K is the fixed context-free language defined above. Thus the weak generative capacity of any context-free grammar and the strong generative capacity of

any modified normal grammar is specified in this way by the choice of a particular 1-limited language.

We have now noted the following features of the three artificial languages introduced in Chapter 11, Sec. 3. All three are beyond the range of finite automata. L_3 can be generated by a context-sensitive grammar but not by a context-free grammar. L_2 can be generated by a context-free, in fact, linear grammar, but not by a countersystem (cf. Sec. 1.4). L_1 can be generated by a countersystem. Furthermore, a language can be generated by a context-free grammar just in case it is accepted by some PDS automaton.

As we observed in Chapter 11, Sec. 3, the fundamental property of L_2 (namely, that it contains nested dependencies) is a common feature of natural languages. It should be noted that dependency sets of the L_3 type also appear in natural languages. Postal (1962) has found a deep-seated system of this sort in Mohawk, where noun sequences of arbitrary length can be incorporated in verbs, with the order of their elements matched in the incorporated and exterior noun sequence. A language containing such a dependency set is beyond the range of a context-free grammar or a PDS automaton, irrespective of any consideration involving structural descriptions (cf. Chapter 11, Sec. 5.1) and strong generative capacity. Subsystems of this sort are also found in English, though more marginally. Thus Solomonoff (1959, personal communication) and Bar-Hillel and Shamir (1960) note that the word *respectively* gives dependency sets of the L_3 type (e.g., *John and Mary wrote to his and her parents, respectively*). Similarly, alongside the elliptical sentence *John saw the play and so did Bill*, we can have *John saw the play and so did Bill see the play*, but not **John saw the play and so did Bill read the book*, etc.

In the same connection, it should be observed that a language is also beyond the weak generative capacity of context-free grammars or PDS automata if it has the essential formal property of the complement of L_3 , that is, if it contains an infinite set of phrases x_1, x_2, \dots , and sentences of the form $\alpha x_i \beta x_j \gamma$ if and only if i is *distinct* from j (whereas a language is of the L_3 type when it contains such sentences if and only if i is *identical* with j , as in the Mohawk example just cited). But restrictions of this kind are very common (cf., e.g., Harris, 1957, Sec. 3.1). Thus in the comparative construction we can have such sentences as *That one is wider than this one is DEEP* (with heavy stress on *deep*), but not **That one is wider than this one is WIDE*—the latter is replaced obligatorily by *That one is wider than this one is*. Thus in these constructions, characteristically, a repeated element is deleted and a nonrepeated element receives heavy stress. We find an unbounded system of this sort when noun phrases are involved, as in the case of such comparatives as *John is more successful as a painter than*

Bill is as a SCULPTOR, but not **John is more successful as a painter than Bill is as a PAINTER*, which is converted, by an obligatory deletion transformation, to *John is more successful as a painter than Bill is*. As in the case of subsystems of the L_3 type, these constructions show that natural languages are beyond the range of the theory of context-free grammars or PDS automata, irrespective of any consideration involving strong generative capacity.

Considerations of this sort show that, in the attempt to enrich linguistic theory to overcome the deficiencies of constituent-structure grammar (cf. Chapter 11, Sec. 5—note that there only deficiencies in strong generative capacity were considered), it is necessary to develop systems that can deal with infinite sets of strings that are beyond the weak generative capacity of the theory of context-free grammar. In these examples, as in the examples discussed in Chapter 11, it is easy to state the required rules in the form of grammatical transformations and thus to handle linguistic phenomena that are beyond the scope of the theory of constituent-structure grammar.

We have now almost completed the proof of Theorem 6 of Sec. 1.6, which asserts that, with respect to weak generative capacity, context-free grammars correspond exactly to nondeterministic PDS automata. In Sec. 2 we observed that unrestricted rewriting systems correspond exactly to Turing machines, and, in Sec. 4.1, that one-sided linear grammars have exactly the weak generative capacity of finite automata. In the case of finite automata and Turing machines, nondeterminacy does not extend (weak) generative capacity. It has recently been shown that every language accepted by a deterministic linear-bounded automaton is context-sensitive (P. Landweber, 1963). S.-Y. Kuroda has observed that this proof extends to nondeterministic linear-bounded automata, and he has proved that, furthermore, every context-sensitive language is accepted by a nondeterministic linear-bounded automaton. It has also been pointed out by Bar-Hillel, Perles, and Shamir (1961) that two-tape automata, as defined by Rabin and Scott (1959), correspond to linear grammars in the following sense. Suppose that y^* is the reflection of y and that a is a designated symbol of V_T . Then, if T is a two-tape automaton accepting the set of pairs $\{(x_i, y_i)\}$ (where x_i, y_i are strings on $V_T - \{a\}$), there is a linear grammar G generating the language $\{x_i a y_i^*\}$; and, if G is a linear grammar generating the language $\{x_i a y_i\}$ (x_i, y_i strings on $V_T - \{a\}$), there is a two-tape automaton that accepts exactly the set of pairs $\{(x_i, y_i^*)\}$. Summarizing, then, we see that with respect to weak generative capacity there is a close correspondence between the hierarchy of constituent-structure grammars and a certain hierarchy of automata, namely that unrestricted rewriting systems correspond to Turing machines, context-sensitive

grammars to nondeterministic linear-bounded automata, context-free grammars to nondeterministic PDS automata, linear grammars to two-tape automata, and one-sided linear grammars to finite automata.

Kuroda has also shown that the complement (with respect to a fixed vocabulary) of a language accepted by a deterministic linear-bounded automaton is context-sensitive and that every context-free language is accepted by a deterministic linear-bounded automaton. It follows, then, that the complement of a context-free language is context-sensitive. We shall see (in Sec. 4.3) that the complement of a context-free language is not necessarily context-free and (in Sec. 4.4) that there is no algorithm for determining whether or not it is context-free.

4.3 Closure Properties

Regular languages are closed under Boolean operations (i.e., formation of union, intersection, complement with respect to a fixed alphabet), as well as under reflection (i.e., a mapping of each string $a_1 \dots a_n$ into $a_n \dots a_1$), product (i.e., formation of the language $L_1 \cdot L_2 = \{x \mid x = yz, \text{ where } y \in L_1 \text{ and } z \in L_2\}$), and infinite closure (i.e., formation of $\bigcup_n L^n$, where $L^n = L \cdot L \cdot \dots \cdot L$, n times). (Cf. Sec. 1.2.) However, this observation carries over to the case of context-free languages only in part.

Theorem 20. *The set of context-free languages is closed under the operations of reflection, product, infinite closure, and set union.* (Bar-Hillel, Perles, & Shamir, 1960.)

However, the intersection of two context-free languages is not necessarily a context-free language; consequently, the complement of a context-free language with respect to a fixed vocabulary is not necessarily a context-free language.

Theorem 21. *The set of context-free languages is not closed under operations of set intersection or complement (with respect to the fixed vocabulary V).* (Scheinberg, 1960b; Bar-Hillel, Perles, & Shamir, 1960.)

Scheinberg gives as a counterexample the languages $\bar{L}_1 = \{a^n b^n a^m\}$ and $\bar{L}_2 = \{a^m b^n a^n\}$, each of which is context-free but which intersect in the set of strings $\{a^n b^n a^n\}$ which is not context-free (the example in Bar-Hillel, Perles, & Shamir, 1960, is essentially the same). The intersection of two sets can, of course, be represented in terms of complement and union. Thus it follows that the complement of a context-free grammar is not necessarily context-free, since the union of context-free grammars is context-free.

Observe that \bar{L}_1 and \bar{L}_2 of the preceding paragraph are meta-linear,

sequential languages. The union of meta-linear languages is meta-linear; the union of sequential languages is sequential. Consequently, this example shows in fact that the sets λ_m and σ of Def. 7 are not closed under complementation and intersection, just as the full set γ is not closed under these operations; and, furthermore, the intersection of two languages of λ_m or of two languages of σ need not be in γ .

Schützenberger has pointed out (personal communication) that the result can be strengthened to linear grammars (grammars of the class λ of Def. 7). Consider the grammar G_1 with the rules

$$S \rightarrow aaSc, \quad S \rightarrow bSc, \quad S \rightarrow bc, \quad (28)$$

and the grammar G_2 with the rules

$$S \rightarrow aSc, \quad S \rightarrow aSb, \quad S \rightarrow ab. \quad (29)$$

The intersection of the languages generated by G_1 and G_2 is the set of strings $\{a^{2n}b^n a^{2n}\}$, which is not context-free; but G_1 and G_2 are linear. They are, furthermore, grammars of the simplest type above the level of finite automaton, that is, linear with a single nonterminal. We see then that even for this simple case the intersection of the generated languages may not be context-free, and the class λ is also not closed under intersection or complementation (it is closed under union).

The preceding argument does not extend directly to subcategories of the set γ of context-free languages; although it does establish that the complement of a language in λ (or λ_m or σ) is not in λ (or in λ_m or σ , respectively), it does not establish that it is not in γ . It is a reasonable conjecture, however, that the result will extend to these subfamilies. It is, as we shall see directly, an important open question whether the complement of a linear language with a single nonterminal (and with a single terminating rule $S \rightarrow c$, where c appears in no other rule) is context-free.

We know that the class λ_1 of regular languages is closed under complementation and intersection (cf. Theorem 1, Sec. 1.2). Summing up, then: all of the categories of context-free languages defined in Def. 7, Sec. 4.1, are closed under the operation of set union, but only λ_1 is closed under intersection and complementation. Furthermore, the intersection of languages of the categories λ , λ_m , or σ need not be context-free (i.e., in γ) at all.

It is interesting to observe that these properties do not carry over to context-sensitive languages. In particular, the intersection of two context-sensitive languages is context-sensitive (Landweber, forthcoming). The status of the complement of a context-sensitive language remains open, however.

4.4 Undecidable Properties of Context-Free Grammars

We showed in Sec. 3 that a great variety of problems involving context-sensitive grammars are recursively unsolvable. Some, but not all, of these problems are also unsolvable in the case of context-free grammars—in fact, even those of the simplest kind beyond the level of finite automata.

It is, first of all, immediate that:

Theorem 22. *There is an algorithm for determining, given the context-free grammar G , whether the language generated by G is empty, finite, or infinite.* (Bar-Hillel, Perles, & Shamir, 1960.)

Hence in this respect more can be determined about the properties of a context-free grammar by systematic investigation of its rules than about a context-sensitive grammar. However, in a variety of other respects, we see that there are striking limitations on what can be determined in this way. The observations concerning undecidability follow essentially Bar-Hillel, Perles, and Shamir (1960), with a few modifications.

Known undecidable properties of context-free grammars follow by reduction to a problem that was proven to be recursively unsolvable by Post (1946), called the *correspondence problem*. This can be stated as follows. Suppose that we are given a set Σ of n pairs of strings on an alphabet of at least two letters. Thus $\Sigma = \{(x_1, y_1), \dots, (x_n, y_n)\}$. A sequence of integers $I = (i_1, \dots, i_m)$, $1 \leq i_j \leq n$, we call an *index sequence* for Σ . We say that the index sequence I *satisfies* Σ just in case

$$x_{i_1} \dots x_{i_m} = y_{i_1} \dots y_{i_m}. \quad (30)$$

The correspondence problem is the problem of determining, given Σ , whether there is an index sequence I satisfying Σ . Post showed that this problem is recursively unsolvable; that is, there is no algorithm for determining, given an arbitrary sequence Σ of pairs of strings, whether there is an index sequence that satisfies Σ . Note that if I satisfies Σ so does $II = (i_1, \dots, i_m, i_1, \dots, i_m)$. Consequently, either there is no index sequence satisfying Σ or there are infinitely many index sequences satisfying Σ ; and the problem whether there are infinitely many satisfying sequences is also therefore recursively unsolvable. This fact plays an important role in the subsequent discussion.

Let us for the moment restrict ourselves to languages with the alphabet $\{a, b, c\}$. Let us designate by $L(G)$ the language generated by the grammar G ; by \bar{L} , the complement of the language L (with respect to the alphabet $\{a, b, c\}$); by $L_1 \cap L_2$, the intersection of L_1 and L_2 ; by $L_1 \cup L_2$, the union of L_1 and L_2 ; and by x^* , the reflection of the string x .

Suppose that we are given an arbitrary set Σ ,

$$\Sigma = \{(x_1, y_1), \dots, (x_n, y_n)\}, \quad (31)$$

where, for each i , x_i and y_i are strings on the alphabet $\{a, b\}$. Let $G(\Sigma)$ be the set of rewriting rules

$$S \rightarrow c, \quad S \rightarrow x_i S y_i^* \quad (1 \leq i \leq n) \quad (32)$$

generating the language $L[G(\Sigma)]$. $G(\Sigma)$ is thus a linear context-free grammar with a single nonterminal. Consider now the question whether $G(\Sigma)$ generates a string $z c z^*$. Clearly, this will be true just in case there is an index sequence (i_1, \dots, i_m) for Σ such that

$$z = x_{i_1} \dots x_{i_m} = y_{i_1} \dots y_{i_m}. \quad (33)$$

Thus the problem of determining whether an arbitrary linear grammar G (with one nonterminal symbol) generates a string of the form $z c z^*$, for some z , is just the Post correspondence problem (cf. Schützenberger, 1961c). Consequently, there is no general method for determining, given a context-free grammar G (which may even be linear, with one nonterminal symbol), whether G generates a string $z c z^*$, hence an infinite number of such strings (as previously noted), or whether it generates no such string.

But now let G_m be the grammar

$$S \rightarrow c, \quad S \rightarrow a S a, \quad S \rightarrow b S b \quad (34)$$

generating the mirror-image language $L(G_m) = \{z c z^*\}$. Given Σ as in Eq. 31, consider the question,

$$\text{What is the cardinality of } L(G_m) \cap L[G(\Sigma)]? \quad (35)$$

But this is just the question, How many strings of the form $z c z^*$ does $G(\Sigma)$ generate? We know that the answer is, Either none or an infinite number, and we have just discovered that there is no algorithm for determining which of these is the case. Therefore, there is no algorithm for determining, given arbitrary Σ , whether $L(G_m) \cap L[G(\Sigma)]$ is empty or infinite.

Observe further that no infinite subset of $L(G_m)$ is a regular language. Consequently, the intersection $L(G_m) \cap L[G(\Sigma)]$ is regular just in case it is finite, that is, empty. Since there is no algorithm for determining this, there is no algorithm for determining whether for arbitrary Σ , $L(G_m) \cap L[G(\Sigma)]$ is a regular language.

Theorem 23. *There is no algorithm for determining, given the context-free grammars G_1 and G_2 , whether $L(G_1) \cap L(G_2)$ is empty, infinite, or regular. (Bar-Hillel, Perles, & Shamir, 1960.)*

In fact, this is true even when G_1 is fixed as in Grammar 34 and G_2 is linear with a single nonterminal symbol (as is G_1 also). Linear grammars with one nonterminal are the simplest systems beyond finite automata in our framework; and it is well known that there is an algorithm for determining whether the intersection of two regular languages is empty or infinite (it is always regular).

We described G_m in Grammar 34 as the grammar generating the mirror-image language with a defined midpoint. Let G_m^2 be the grammar consisting of the rules in Grammar 34 and, in addition, the rule

$$S_1 \rightarrow ScS. \quad (36)$$

Let S_1 be the initial symbol of G_m^2 . Thus $L(G_m^2)$ consists of all strings of the form xcx^*cy^* , where x and y are strings in the alphabet $\{a, b\}$.

Given Σ again, as in Example 31, define $G_m(\Sigma)$ as the grammar containing Rules 32 of $G(\Sigma)$ and, in addition, the rules

$$S_1 \rightarrow aS_1a, \quad S_1 \rightarrow bS_1b, \quad S_1 \rightarrow cSc, \quad (37)$$

where S_1 is the initial symbol. $G_m(\Sigma)$ is the grammar that embeds $L[G(\Sigma)]$, as defined by Rules 32, into the mirror-image language. That is, $L[G_m(\Sigma)]$ consists of exactly those strings of the form $xcyczc^*$, where ycz is generated by $G(\Sigma)$; $G_m(\Sigma)$ is basically an amalgam of G_m and $G(\Sigma)$.

Consider now the intersection of the languages generated by G_m^2 and $G_m(\Sigma)$ {just as before we considered the intersection of $L(G_m)$ and $L[G(\Sigma)]$ }. This is the set $L(G_m^2) \cap L[G_m(\Sigma)]$ consisting of exactly those strings x meeting the following conditions:

- (i) $x = x_1cx_2cx_3cx_4$ (x_i a string on $\{a, b\}$)
- (ii) $x_1 = x_2^*$ and $x_3 = x_4^*$ (since $x \in L(G_m^2)$)
- (iii) $x_1 = x_4^*$ and $x_2cx_3 \in L[G(\Sigma)]$ (since $x \in L[G_m(\Sigma)]$).

In particular, then, $x_1 = x_3$, $x_2 = x_4$, and $x_2 = x_3^*$. Consequently, $L(G_m^2) \cap L[G_m(\Sigma)]$ will be empty just in case there is no index sequence satisfying Σ and infinite otherwise, exactly as before (since, as we have already observed, the question whether there is a string $x_2cx_3 \in L[G(\Sigma)]$, where $x_2 = x_3^*$, is exactly the correspondence problem for Σ). Consequently, there can be no algorithm for determining whether $L(G_m^2) \cap L[G_m(\Sigma)]$ is empty or infinite (these being the only possibilities).

Each string of $L(G_m^2) \cap L[G_m(\Sigma)]$ is of the form xcx^*cx^* , and it is easy to show that no infinite set of strings of this form constitutes a context-free language. Consequently, $L(G_m^2) \cap L[G_m(\Sigma)]$ is a context-free language just in case it is finite, that is, empty, and since the question of emptiness is, as just observed, undecidable, the question whether $L(G_m^2) \cap L[G_m(\Sigma)]$ is a context-free language is undecidable.

Theorem 24. *There is no algorithm for determining, given the context-free grammars G_1 and G_2 , whether $L(G_1) \cap L(G_2)$ is a context-free language. (Bar-Hillel, Perles, & Shamir, 1960.)*

In fact, this is true even when G_1 is fixed as G_m^2 and G_2 is meta-linear (note that G_m^2 is also meta-linear). Note that Theorem 23 follows from the argument that proves Theorem 24.

We have considered languages with the alphabet $\{a, b, c\}$, but by appropriate coding, we can easily extend these unsolvability results to languages with alphabets of two or more symbols (cf. Bar-Hillel, Perles, & Shamir, 1960).

In Bar-Hillel, Perles, and Shamir (1960) the proof of Theorems 23 and 24 proceeds essentially as follows. Consider Σ , as in Example 31. Let L_Σ be the language consisting of all strings

$$ab^{i_k} \dots ab^{i_1}cx_{i_1} \dots x_{i_k}cy_{j_1}^* \dots y_{j_1}^*cb^{j_1}a \dots b^{j_1}a, \tag{39}$$

where (i_1, \dots, i_k) and (j_1, \dots, j_l) are index sequences for Σ . It is easy to show that L_Σ is a context-free language generated by a grammar G_Σ . G_Σ , so defined, plays the role of $G_m(\Sigma)$ in the proof previously sketched.

Consider now the language L_M consisting of just the strings

$$x_1cx_2cx_2^*cx_1^*, \tag{40}$$

where x_1 and x_2 are strings on $\{a, b\}$. L_M is also a context-free language, generated by a grammar G_M ; this grammar plays the role of G_m^2 in the proof sketched previously.

We observe now that $L_M \cap L_\Sigma$ is the set of all strings

$$ab^{i_k} \dots ab^{i_1}cx_{i_1} \dots x_{i_k}cy_{i_k}^* \dots y_{i_1}^*cb^{i_1}a \dots b^{i_k}a, \tag{41}$$

where $x_{i_1} \dots x_{i_k} = y_{i_1} \dots y_{i_k}$; that is, where i_1, \dots, i_k is an index sequence that satisfies Σ . Consequently, by the argument given previously, if there is an index sequence satisfying Σ , then $L_M \cap L_\Sigma$ is infinite and is not a regular language or a context-free language; if there is no index sequence satisfying Σ , then $L_M \cap L_\Sigma$ is empty (and therefore, trivially, is a regular language and a context-free language). Hence unsolvability of the Post correspondence problem implies unsolvability of the problem of determining, for arbitrary Σ , whether $L_M \cap L_\Sigma$ is empty, finite, a regular language, or a context-free language (Theorems 23 and 24).

By a construction too complex to reproduce here, it is shown in Bar-Hillel, Perles, and Shamir (1960) that, for each Σ , the complement \bar{L}_Σ of L_Σ is a context-free language. It is easy to show that the complement \bar{L}_M of L_M is a context-free language. The union of two context-free languages is context-free. Therefore, for each Σ the language $\bar{L}_M \cup \bar{L}_\Sigma$

$= \overline{L_M \cap L_\Sigma}$ is a context-free language. The complement of this language is $L_M \cap L_\Sigma$, and we have just observed that there is no algorithm for determining whether this is empty, finite, regular, or context-free. Each step is constructive. Consequently, we have the result:

Theorem 25. *There is no algorithm for determining, given a context-free grammar G , whether the language $\overline{L(G)}$ (the complement of the language generated by G with respect to the alphabet $\{a, b, c\}$) is empty, finite, regular, or context-free. (Bar-Hillel, Perles, & Shamir, 1960.)*

In particular, there is no algorithm for determining whether an arbitrary context-free grammar generates all strings in its terminal vocabulary, that is, whether it is universal; and there is no algorithm for determining whether an arbitrary context-free grammar generates a regular language (since this will be true just in case the complement of the language it generates is a regular language).

At the end of Sec. 4.3 we observed that it is not known whether the complement of a language generated by a linear grammar with one nonterminal (the simplest type of nonregular language) is a context-free language. If the answer to this question is positive, then the complement of $L[G(\Sigma)]$ generated by $G(\Sigma)$ of Example 32 is context-free. It is easy to show that the complement of $L(G_m)$ generated by G_m of Example 34 is context-free. Consequently, by the argument given, we can show that there is no algorithm for determining whether the complement of a context-free language (namely, the union of the complements of these linear languages) is empty, finite, or regular. Furthermore, the complement of $L(G_m^2)$ (cf. Rule 36) is context-free, and the complement of $L[G_m(\Sigma)]$ is context-free if the complement of $L[G(\Sigma)]$ is also. Hence we would be able to prove Theorem 25 in full without considering L_Σ of Example 39, or the construction that gives its complement, if it is true that the complement of a language generated by a linear grammar with one nonterminal is context-free. This is apparently not a simple question, however.

The construction of the complement of L_Σ given by Bar-Hillel, Perles, and Shamir (1960) amounts to a proof that for a certain type of linear grammar with one nonterminal, the complement is linear. In Schützenberger and Chomsky (1962) it is proved that for a more general class of linear grammars with designated center symbols and one nonterminal (namely, those for which the string to the right of the center symbol in the generated language is uniquely determined by the string to its left) it is still true that the complement is linear. But the general question remains open, even for linear grammars with a single nonterminal and a designated center symbol.

From Theorem 25 it follows immediately that:

Theorem 26. *There is no algorithm for determining, given context-free*

grammars G_1 and G_2 , whether $L(G_1) = L(G_2)$. (Bar-Hillel, Perles, & Shamir, 1960.)

If there were such an algorithm, then it would be possible to determine in general whether $L(G_1)$ is universal, contrary to Theorem 25. It also follows immediately that the problem of determining whether $L(G_1) \subset L(G_2)$ is recursively unsolvable {this, in fact, follows also from Theorem 23, since $\overline{L(G_m)}$ is context-free, and $L[G(\Sigma)]$ is included in it just in case its intersection with $L(G_m)$ is null}.

One further immediate consequence of these undecidability results deserves mention here. We have already observed (last paragraph of Sec. 1.5) that the language L is regular if and only if there is a transducer T such that $T(U) = L$, where U is the set of all strings on V_T . As we have seen, Theorem 25 implies that there is no algorithm for determining whether an arbitrary context-free language L is a regular language; that is, whether there is a transducer T such that $T(U) = L$. Since U is a context-free (in fact, regular) language, we have the following theorem: Theorem 27. *There is no algorithm for determining, given two context-free languages L_1 and L_2 , whether there is a finite transducer T such that $T(L_1) = L_2$.* (S. Ginsburg, personal communication.)

4.5 Structural Ambiguity

We say that a context-free grammar G is *ambiguous* if it generates a string x in two essentially different ways, that is, if it assigns to x two distinct structural descriptions (cf. p. 367). The topic of structural ambiguity is important from many points of view, and, though it has been very little studied so far, there are some suggestive results.

Consider this question first. Is there an algorithm to determine whether a context-free grammar is ambiguous? A negative answer follows directly from the unsolvability of the correspondence problem. Suppose in fact that $\Sigma = \{(x_i, y_i) \mid 1 \leq i \leq n\}$, where x_i and y_i are again strings on some alphabet, say, the alphabet $\{a, b\}$. Select n new symbols d_1, \dots, d_n and construct the two grammars G_x and G_y as follows:

$$\begin{aligned} G_x: S_x &\rightarrow c, & S_x &\rightarrow d_i S_x x_i^* & (1 \leq i \leq n), \\ G_y: S_y &\rightarrow c, & S_y &\rightarrow d_i S_y y_i^* & (1 \leq i \leq n). \end{aligned} \tag{42}$$

Clearly, G_x and G_y are unambiguous, but note that there is an index sequence (i_1, \dots, i_m) such that $x_{i_1} \dots x_{i_m} = y_{i_1} \dots y_{i_m}$ if and only if G_x and G_y both generate the string z , where

$$z = d_{i_1} \dots d_{i_m} c x_{i_m}^* \dots x_{i_1}^* = d_{i_1} \dots d_{i_m} c y_{i_m}^* \dots y_{i_1}^*. \tag{43}$$

Thus the correspondence problem for Σ has a positive solution if and only if there is a z generated by both G_x and G_y , that is, if the grammar G_{xy} is ambiguous, where G_{xy} contains the rules of G_x , the rules of G_y , and, in addition, the rules $S \rightarrow S_x, S \rightarrow S_y$, S being the initial symbol of G_{xy} . Consequently, there is no algorithm for determining whether, for arbitrary Σ , the grammar G_{xy} constructed in the manner indicated is ambiguous.

Theorem 28. *There is no algorithm for determining whether a context-free grammar is ambiguous.* (Schützenberger, personal communication.)

Note that the grammar G_{xy} belongs to a class of grammars G meeting the following condition:

$$\begin{aligned} G \text{ is linear with at least three nonterminals} \\ \text{and terminating rules all of the form } \alpha \rightarrow c, \\ \text{where } c \text{ does not appear in any nonterminating} \\ \text{rule of } G. \end{aligned} \quad (44)$$

Thus we see that the ambiguity problem is unsolvable for grammars meeting this condition.

It has recently been shown that through a generalization of the correspondence problem, Condition 44 can be weakened to the case of one nonterminal without affecting the unsolvability of the ambiguity problem. In effect, this generalization permits S , S_x , and S_y to be identified in G_{xy} (Greibach, forthcoming). Thus Theorem 28 holds of the class of *minimal linear grammars* G meeting the condition:

$$\begin{aligned} G \text{ is linear with a single nonterminal } S \text{ and a} \\ \text{single terminating rule } S \rightarrow c, \text{ where } c \text{ does} \\ \text{not appear in any nonterminating rule of } G. \end{aligned} \quad (45)$$

Suppose that G is a minimal linear grammar with the nonterminating rules $S \rightarrow x_i S y_i^*$, $1 \leq i \leq n$, associated with the set of pairs

$$\Sigma = \{(x_i, y_i) \mid 1 \leq i \leq n\}.$$

G is ambiguous if and only if there are two index sequences I, J

$$\begin{aligned} I = (i_1, \dots, i_p), \quad J = (j_1, \dots, j_q), \\ I \neq J, \quad 1 \leq i_m, j_m \leq n, \end{aligned} \quad (46)$$

such that $x_{i_1} \dots x_{i_p} c y_{i_p}^* \dots y_{i_1}^* = x_{j_1} \dots x_{j_q} c y_{j_q}^* \dots y_{j_1}^*$. This, in turn, is true if and only if

$$\begin{aligned} \text{(i)} \quad & x_{i_1} \dots x_{i_p} = x_{j_1} \dots x_{j_q}, \\ \text{(ii)} \quad & y_{i_1} \dots y_{i_p} = y_{j_1} \dots y_{j_q}. \end{aligned} \quad (47)$$

But consider the question:

$$\begin{aligned} \text{Given } \Sigma, \text{ do there exist index sequences } I, J \\ \text{as in Example 46, that satisfy Condition 47?} \end{aligned} \quad (48)$$

From Theorem 28, extended to minimal linear grammars, it follows that this question concerning unique decipherability is undecidable.

We observed in Sec. 1.2 that corresponding to every finite automaton we can construct an equivalent deterministic finite automaton (and, of course, the question whether two finite automata are equivalent is decidable). In other words, given a one-sided linear grammar G , we can find an equivalent unambiguous one-sided linear grammar. An obvious question is whether this is also true of context-free grammars in general. It has been proven by Parikh (1961) that it is not.

Theorem 29. *There are context-free languages that cannot be generated by any unambiguous context-free grammar.*

Parikh proves that the language

$$L = \{x \mid x = a^n b^m a^{n'} b^m \text{ or } x = a^n b^m a^{n'} b^{m'}; n, n', m, m' \geq 1\} \quad (49)$$

cannot be generated by an unambiguous context-free grammar, although it is generated by the set of rules

$$\begin{aligned} S &\rightarrow AB, & S &\rightarrow DC, \\ A &\rightarrow aAa, & A &\rightarrow aBa, \\ C &\rightarrow bCb, & C &\rightarrow bDb, \\ B &\rightarrow b, & B &\rightarrow bB, \\ D &\rightarrow a, & D &\rightarrow aD. \end{aligned} \quad (50)$$

There is, in fact, a linear grammar equivalent to the Grammar 50. The degree of ambiguity it assigns to strings is 2. Another example of such a language is the set $\{a^n b^m a^p \mid n = m \text{ or } n = p\}$. It is an interesting open problem to find languages of higher (perhaps unbounded) degree of inherent ambiguity. Many open and important questions can immediately be raised concerning the question of inherent ambiguity, its scale, its relation to decidability of ambiguity, and the level of richness of grammar at which it arises (e.g., are there minimal linear languages that are inherently ambiguous or is there inherent ambiguity at the level of context-sensitive grammars?), etc.

Theorem 29 is a suggestive result. It is an interesting question why natural languages have as much structural ambiguity as they do. We might hope to obtain an answer to this question that would take the following form:

1. Grammars of natural languages are drawn from the class Γ of generative processes.
2. The language L is rich enough in expressive power to contain the set of grammatical devices Δ but not Δ' (e.g., the class of sentences Σ but not Σ').

3. A grammar $G \in \Gamma$ that expresses Δ but not Δ' (that generates Σ but not Σ') must be ambiguous, that is, the language it generates is inherently ambiguous with respect to Γ .

If an argument of this kind could be provided, it would be important not only as an explanation for the existence of structural ambiguities in L , but it would provide striking evidence of an indirect (hence quite interesting) kind for the validity of the general linguistic theory that makes the claim (1). It need hardly be emphasized that we are far from being able to provide such an argument for natural language, but Theorem 29 may represent a first step in this direction.

4.6 Context-Free Grammars and Finite Automata

We have seen that a finite automaton can be represented as a one-sided linear grammar and that such a device is much more restricted in generative capacity than a context-free or even a linear grammar may be. We have also observed that such elementary formal properties of natural languages as recursive nesting of dependencies make it impossible for them to be generated by finite automata, although these properties do not exclude them from the class of context-free (even linear) languages. From these observations we must conclude that the *competence* of the native speaker cannot be characterized by a finite automaton. The grammar stored in his brain cannot be a one-sided linear grammar, a fact that is not in the least surprising. Nevertheless, the *performance* of the speaker or hearer must be representable by a finite automaton of some sort. The speaker-hearer has only a finite memory, a part of which he uses to store the rules of his grammar (a set of rules for a device with unbounded memory), and a part of which he uses for computation in actually producing a sentence or "perceiving" its structure and understanding it.

These considerations are sufficient to show the importance of gaining a better understanding of the source and extent of the excess generative power of context-free grammars over finite automata (even though context-free grammars are demonstrably not fully adequate for the grammatical description of natural languages). We turn now to an investigation of this problem.

Let us first review what we have so far found concerning the relation of context-free grammars to restricted-infinite and finite automata. In Secs. 1.6 and 4.2 we have stated several results (which are contingent, in part, on results yet to be proved in this section) having to do with this question. In particular, we observed that context-free languages are the sets that are accepted by a class of restricted-infinite automata that we

called pushdown storage (PDS) automata. We showed that from this fact it follows that there is an extremely close relation between regular (one-sided linear) languages and context-free languages. Namely, let us extend the vocabulary V from which all context-free languages are constructed to a vocabulary V' containing V and α' for each $\alpha \in V$. Let us define K as the set of strings on V' that reduce to e by successive cancellation of $\alpha\alpha'$ and $\alpha'\alpha$ (i.e., by treating α and α' as inverses). Let us define $\phi(\alpha) = \alpha$ for $\alpha \in V_T$, $\phi(\alpha) = e$ for $\alpha \notin V_T$. For any language L , let us define $\psi(L) = \phi(K \cap L)$. Then, for each regular language L , $\psi(L)$ is a context-free language, and each context-free language is determined as $\psi(L)$ for some choice of a regular language L . Hence the family λ_1 of regular languages is mapped onto the family γ of context-free languages by the mapping ψ .

Continuing with the investigation of the relation between context-free and regular languages, we note first that there are context-free languages that are, in a sense, much 'bigger' than any regular language.

Theorem 30. *There is a context-free grammar G generating $L(G)$ with the following property: given a finite automaton A_1 generating $L(A_1) \subset L(G)$, we can construct a finite automaton A_2 generating $L(A_2)$ such that (i) $L(A_1) \subset L(A_2) \subset L(G)$ and (ii) $L(A_2)$ contains infinitely many strings not in $L(A_1)$. (Parikh, 1961.)*

Parikh shows that this result holds for a context-free grammar that provides the relations

$$S \Rightarrow A^n B^m A^n, \quad A \Rightarrow ce^k c, \quad B \Rightarrow df^k d \quad (m, n, k \geq 1). \quad (51)$$

[In fact, it is also true of the simpler grammar G generating just $L(G) = \{a^n b^m a^n \mid m, n \geq 1\}$ (S. Ginsburg, personal communication).] From Theorem 30 we see that we cannot, in general, approach a context-free language L as the limit of an increasing sequence of regular languages, each containing an infinite number of sentences not in the preceding one and all contained in L .

Suppose now that we have a context-free grammar G generating $L(G)$ and a finite transducer T with initial state S_0 . We shall construct a new context-free grammar G' which, in fact, generates $T[L(G)]$. Suppose, first, that T is bounded. We can, then, assume that it contains no instructions of the form $(e, S_i) \rightarrow (S_j, x)$. Let us construct the new context-free grammar G' with the output vocabulary of T as its terminal vocabulary and with nonterminal symbols represented as triples (S_i, α, S_j) , where S_i, S_j are states of T and $\alpha \in V$.⁴

⁴ The construction that follows is due to Bar-Hillel, Perles, and Shamir (1960) who use it only to prove what we give here as Theorem 32. Our Theorem 31 is proved in essentially this way in Ginsburg and Rose (1961). This construction is closely related to the representation of transducers by matrices in Schützenberger (1961a).

The initial symbol of G' is S' . The rules of G' are determined by the following principle:

- (i) $S' \rightarrow (S_0, S, S_i)$ is a rule of G' , for each i .
- (ii) If $A \rightarrow \alpha_1 \dots \alpha_k$ is a rule of G , then for each $i, j, \beta_1, \dots, \beta_{k-1}$, G' contains the rule

$$(S_i, A, S_j) \rightarrow (S_i, \alpha_1, S_{\beta_1})(S_{\beta_1}, \alpha_2, S_{\beta_2}) \dots (S_{\beta_{k-1}}, \alpha_k, S_j). \quad (52)$$
- (iii) If $(a, S_i) \rightarrow (S_j, x)$ is a rule of T , then G' contains the rule

$$(S_i, a, S_j) \rightarrow x.$$

To preserve the condition that a nonterminal symbol is one that appears to the left of a rule $\alpha \rightarrow \phi$, we can require also that G' contain the rules

$$(S_i, a, S_j) \rightarrow (S_i, a, S_j)a, \quad (53)$$

for each a, i, j not involved in step iii.

The terminating rules of G' are those given by step (iii) of Construction 52. Carrying a derivation of G' as far as we can without applying any terminating rules, we have, as final line, a string

$$(S_{i_0}, \alpha_1, S_{i_1})(S_{i_1}, \alpha_2, S_{i_2}) \dots (S_{i_{k-1}}, \alpha_k, S_{i_k})a, \quad (54)$$

where $i_0 = 0$, $\alpha_j \in V_T$ for each j , and G generates $\alpha_1 \dots \alpha_k$. Furthermore, if G generates $\alpha_1 \dots \alpha_k$ ($\alpha_j \in V_T$), then for each i_1, \dots, i_k , the String 54 is a line of a derivation of G' . But a derivation with the String 54 as final line will terminate with the terminal string z if and only if there are strings x_1, \dots, x_k such that $x_1 \dots x_k = z$ and for each j , $(\alpha_j, S_{i_{j-1}}) \rightarrow (S_{i_j}, x_j)$ is an instruction of T ; that is, if and only if T maps the input string $\alpha_1 \dots \alpha_k$ into z , passing through states $S_0, S_{i_1}, \dots, S_{i_k}$ in the process. Thus G generates the language $T[L(G)]$. Note that G' may have rules of the form $\alpha \rightarrow e$ (where $x = e$ in step iii of Construction 52). However, as we observed at the outset of Sec. 4, rules of this kind do not permit the generation of noncontext-free languages (aside from the language $\{e\}$). Consequently, we have the following result, where T is a bounded transducer.

Theorem 31. *If L is a context-free grammar and T a transducer, then $T(L)$ is a context-free language (or $T(L) = \{e\}$).*

Suppose that R is a regular language accepted by the automaton F with initial state S_0 . Construct the transducer T with the instruction $(a_i, S_j) \rightarrow (S_k, a_i)$ whenever F has the instruction (i, j, k) , that is, whenever F goes from state S_j to state S_k on reading the input a_i . We can assume, with no loss of generality, that F is deterministic (cf. Theorem 1) and that T is bounded. Let G be a context-free grammar generating $L(G)$. Construct G' by the construction previously given, but with the revision that in

place of step i of Construction 52 we have the single rule $S' \rightarrow (S_0, S, S_0)$. Now it is easy to show that G' generates the intersection of R with $L(G)$ by the argument that leads to Theorem 31.

Theorem 32. *If R is a regular language and L a context-free language, then the intersection $L \cap R$ is a context-free language.*

To drop the requirement of boundedness of the transducer T in Theorem 31, we amend the construction as follows. Given the context-free grammar G generating $L(G)$, first replace each rule $A \rightarrow x_1 \dots x_k$ by the rule $A \rightarrow qx_1qx_2q \dots qx_kq$, where q is some new symbol. Then apply the construction in (52) and (53), as before, to give G' . Now define Q_{ij} as

$$Q_{ij} = \{z \mid \text{for some } i_0, \dots, i_m, x_1, \dots, x_m, z = x_1 \dots x_m \\ \text{and for each } k, 1 \leq k \leq m, T \text{ has the rule} \quad (55) \\ (e, S_{i_{k-1}}) \rightarrow (S_{i_k}, x_k), \text{ where } i_0 = i \text{ and } i_m = j\}.$$

Clearly Q_{ij} is regular. Therefore, we can add to G' rules that provide for the relations

$$(S_i, q, S_j) \rightarrow e \quad \text{and} \quad (S_i, q, S_j) \Rightarrow x, \quad (56)$$

for each i, j and each $x \in Q_{ij}$. G' , so extended, generates the language $T[L(G)]$. Hence Theorem 31 holds without restriction on T .

For additional results concerning the effect of various operations on context-free languages and related systems, see Ginsburg and Rose (1961) and Schützenberger and Chomsky (1962).

Let us now restrict our attention to rules of the forms

- (i) $A \rightarrow BC,$
 - (ii) $A \rightarrow aB$ (right-linear),
 - (iii) $A \rightarrow Ba$ (left-linear),
 - (iv) $A \rightarrow a.$
- (57)

Recall that a *normal grammar* contains only rules of the types (i) and (iv); a *linear grammar* can be described, without loss of generality, as one that contains only rules of the forms (ii), (iii), and (iv); a *finite automaton* contains only rules of the forms (ii) and (iv), or only rules of the form (iii) and (iv). Recall also that a normal grammar can generate any context-free language and that a linear grammar, although more limited in generative capacity, can generate languages beyond the capacity of finite automata (Theorem 16). Thus we gain generative capacity over finite automata by permitting both right- and left-linear rules and, still beyond this, by allowing nonlinear rules to appear in the grammar. Of course, some normal grammars and some linear grammars may generate only regular languages, although there is no algorithm to determine when this is true

in either case (Theorem 25). The question remains, then, under what conditions is a normal or linear grammar richer than any finite automaton in generative capacity?

In order to study this question, it is useful to consider again the classification of recursive elements given in Chapter 11, Sec. 3. A finite automaton contains either all right-recursive or all left-recursive elements. A linear grammar may contain both right-recursive and left-recursive elements, as may a normal grammar. Furthermore, both linear and normal grammars may contain self-embedding elements. It turns out to be the latter property that accounts for the excess generative capacity over finite automata.

Definition 8. *A grammar is self-embedding if it contains a nonterminal symbol A such that for some nonnull strings ϕ, ψ , $A \Rightarrow \phi A \psi$.*

A self-embedding grammar is, in other words, one that contains self-embedding elements. It can now be shown:

If G is a nonself-embedding, context-free grammar
generating $L(G)$, then there is a finite automaton (58)
that generates $L(G)$.

From this we derive the following result immediately:

Theorem 33. *The language L is not regular if and only if all of its context-free grammars are self-embedding.* (Chomsky, 1959a, 1959b; Bar-Hillel, Perles, & Shamir, 1960.)

Thus L is a regular language just in case it has a nonself-embedding grammar.

Clearly, there is an algorithm to determine whether a context-free grammar contains self-embedding elements (similarly, whether it contains right-recursive and left-recursive elements). If we apply this test to a grammar G and discover that it has no self-embedding symbols, then we can conclude that G has the capacity of a finite automaton (although it may have both right-recursive and left-recursive symbols). If, on the other hand, we find that G has self-embedding symbols, we do not know whether G has the capacity of a finite automaton. This depends on the answer to the question whether there is a grammar G' that generates the same language as G with no self-embedding symbols. As we have seen, there is no mechanical procedure by which this can be determined for arbitrary G . Thus, although Theorem 33 provides an effective *sufficient condition* for a context-free grammar to be equivalent to a finite automaton in generative capacity (and, furthermore, a condition met by some grammar of each regular language), we know that it cannot be strengthened to an effective *criterion*, that is, an effective necessary and sufficient condition.

Proposition 58 follows directly from elementary properties of finite automata. Suppose that G is a nonself-embedding grammar generating $L(G)$, where the nonterminals of G are A_1, \dots, A_n . Call G *connected* if for each i, j there are strings ϕ, ψ such that $A_i \Rightarrow \phi A_j \psi$. Suppose that G is connected. If there are i, j, k, l such that $A_i \Rightarrow \phi_1 A_j \phi_2$ and $A_k \Rightarrow \psi_1 A_l \psi_2$, where $\phi_1 \neq \epsilon \neq \psi_2$, then it is immediate that G is self-embedding, contrary to assumption. Therefore there can be no such i, j, k, l , and thus each nonterminating rule of G is right-linear or each rule is left-linear. In either case, G generates a regular language.

Suppose now that $n = 1$. Then either $L(G)$ is finite or G is connected and generates a regular language, as just noted.

Suppose that Proposition 58 is true for all grammars containing less than n nonterminals. Suppose that A_1 is the initial symbol of G , which has nonterminals A_1, \dots, A_n . We may assume that G contains no redundant symbols and that it is not connected. Thus for some particular j there is no ϕ, ψ such that $A_j \Rightarrow \phi A_1 \psi$. Suppose $j \neq 1$. Form G' by deleting from G each rule $A_j \rightarrow \phi$ and replacing A_j elsewhere in the rules by a new symbol a . By the inductive hypothesis the language L' generated by G' is regular, as is the set $K = \{x \mid A_j \Rightarrow x\}$. It is obvious that if L_1 and L_2 are regular and L_3 consists of all those strings formed from $x \in L_1$ by replacing each a (if any) in x by some string of L_2 , then L_3 is regular. $L(G)$ is formed in this way from L' and K and is therefore regular.

Suppose $j = 1$. Let ϕ_1, \dots, ϕ_r be the strings such that $A_1 \rightarrow \phi_i$. For each i let $K_i = \{x \mid \phi_i \Rightarrow x\}$. Suppose that $\phi_i = \alpha_1 \dots \alpha_m$. By the inductive hypothesis the set $L_j = \{x \mid \alpha_j \Rightarrow x\}$ is regular. By Theorem 2 it follows that K_i , hence L , is also regular. This establishes Proposition 58. This observation also follows immediately from the much stronger result to which we turn next.

We have considered so far only the question of *weak equivalence* among grammars, that is, the question whether they generate the same language. We have also defined a relation of *strong equivalence* holding between two grammars that not only generate the same language, but also the same set of structural descriptions (see Chapter 11, Sec. 5.1). Little is known about strong equivalence. However, it is important to observe that we can extend Proposition 58 to the effect that, given a nonself-embedding grammar G , we can construct a finite transducer T that, in a sense that we shall make precise, is strongly equivalent to G . We can also strengthen this result to any finite degree of self-embedding. It has certain consequences to which we return in Chapter 13. See also Chomsky (1961) for further discussion.

To make this result precise, let us consider more closely the class of finite transducers. We can assume that the input alphabet of each transducer

T is a subset of V_T (the fixed and universal set of terminal symbols of all context-free grammars). We assume that the output alphabet of each transducer is a subset of some fixed set A_0 . Given T , we say that T *accepts* x and *assigns to it the structural description* y [briefly, T *generates* (x, y)] just in case T begins its computation in state S_0 with a blank storage tape scanning the leftmost symbol of the string $x\#$ on the input tape, and terminates its computation on its first return to S_0 scanning $\#$ on the input tape and with y as the contents of the storage tape.⁵ Thus, if T generates (x, y) , then T accepts x in the manner of a finite automaton with no output (cf. Sec. 1.2) and maps it into y in the manner of a transducer.

In order to compare the generative capacity of transducers, so regarded, with that of context-free grammars, let us assume that we are given an effective one-one mapping, Φ , which maps the set of structural descriptions given by context-free grammars (defined, let us say, as strings with labeled bracketing—cf. p. 367) into the set of strings in A_0 . We can now define strong equivalence as a relation between context-free grammars and finite transducers:

Definition 9. *Given the context-free grammar G and the finite transducer T , then G and T are strongly equivalent if and only if the following condition is met: T generates $(x, \Phi(y))$ just in case G generates x with the structural description y .*

Thus, if T is strongly equivalent to G , T accepts just those strings generated by G and maps each such string into each of the structural descriptions assigned to it by G and nothing else.

We can now state a precise form of the generalization of Proposition 58: **Theorem 34.** *There is an effective procedure Ψ such that, given a normal nonself-embedding context-free grammar G , $\Psi(G)$ is a finite transducer that is strongly equivalent to G . (Chomsky, 1959a).*

As previously observed, this procedure can easily be generalized to any finite degree of self-embedding in a manner that we will describe more carefully. Clearly, Proposition 58 and Theorem 33 follow from Theorem 34.

Theorem 34 has actually been proved only for normal grammars meeting the additional condition that if $A \rightarrow BC$ is a rule then $B \neq C$, and if $A \rightarrow \phi B \psi$ and $A \rightarrow \chi B \omega$ are rules then $\phi = \chi$ and $\psi = \omega$. It is not difficult to show that these additional conditions have no effect on generative capacity. Furthermore, it is merely a matter of added detail to drop these restrictions (and, in fact, many, if not all of the restrictions that give normality).

The proof of Theorem 34 is too complex to be given here, but we present the procedure Ψ and illustrate it by an example. Beforehand, however,

More precisely, in view of the account of transducers given in Sec. 1.5, we should say that T begins its computation scanning σ on an otherwise blank storage tape and terminates its computation with σy as the contents of the storage tape.

note how Theorem 34 differs from Theorem 19 (Sec. 4.2), which states that, given any modified normal grammar G , there is a finite transducer T with the following property: T maps x into a string z that reduces to e if and only if z is a structural description assigned to x by G . In this case Φ is the identity mapping; that is, the output of T is in the exact form of a structural description assigned by G . Furthermore, there is no limitation here to nonself-embedding grammars. However, the transducer T guaranteed by Theorem 19 is not strongly equivalent to G in the sense of Def. 9. Thus T can return to S_0 and terminate, with input x , with a string y on the storage tape that does not reduce to e (and is not a structural description assigned to x by G). In fact, the reason why the transducer T associated with a context-free grammar in Theorem 19 appears, superficially, more powerful than the transducer T' associated with a context-free grammar by Theorem 33 is that T , but not T' , is, in effect, using a potentially infinite memory in deciding when to accept a string with a given structural description, since this decision requires an analysis of the (unbounded) string on the storage tape to determine whether it reduces to e . We know from Theorem 33 that Theorem 34 is the strongest possible result concerning strong equivalence of context-free grammars and devices with strictly bounded memory.

To illustrate Theorem 34, suppose that G is a normal, nonself-embedding grammar meeting the additional condition stated directly after Theorem 34. Let K be the set of sequences $\{(A_1, \dots, A_m)\}$ meeting the following condition: for each i, j such that $1 \leq i < j \leq m$, $A_i \rightarrow \phi A_{i+1}\psi$ and $A_i \neq A_j$. We can now construct the grammar G^* with nonterminal symbols represented in the form $[B_1, \dots, B_n]_i$ ($i = 1, 2$), where the B_j 's are nonterminal symbols of G :⁶

Suppose that $(B_1, \dots, B_n) \in K$.

- (i) If $B_n \rightarrow a$, then $[B_1 \dots B_n]_1 \rightarrow a[B_1 \dots B_n]_2$.
- (ii) If $B_n \rightarrow CD$, where $C \neq B_i \neq D$ ($i \leq n$), then
 - (a) $[B_1 \dots B_n]_1 \rightarrow [B_1 \dots B_n C]_1$,
 - (b) $[B_1 \dots B_n C]_2 \rightarrow [B_1 \dots B_n D]_1$
 - (c) $[B_1 \dots B_n D]_2 \rightarrow [B_1 \dots B_n]_2$.
- (iii) If $B_n \rightarrow CD$, where $B_i = D$ for some $i \leq n$, then (59)
 - (a) $[B_1 \dots B_n]_1 \rightarrow [B_1 \dots B_n C]_1$,
 - (b) $[B_1 \dots B_n C]_2 \rightarrow [B_1 \dots B_i]_1$.
- (iv) If $B_n \rightarrow CD$, where $B_i = C$ for some $i \leq n$, then
 - (a) $[B_1 \dots B_i]_2 \rightarrow [B_1 \dots B_n D]_1$,
 - (b) $[B_1 \dots B_n D]_2 \rightarrow [B_1 \dots B_n]_2$.

⁶ We can use the symbol \rightarrow unambiguously for both G and G^* , since the forms of their nonterminal symbols differ.

We can now prove that there is an S -derivation of z in G if and only if there is a $[S]_1$ -derivation of $z[S]_2$ in G^* (Theorem 10 in Chomsky, 1959a), where S is the initial symbol of G .

The rules of G^* are all of the form $A \rightarrow aB$, where $a = \epsilon$ unless the rule in question was formed by step (i) of the Construction 59. We can, therefore, regard G^* as a finite automaton. Suppose that we now supply the automaton with a new state S_0 and the additional rules

$$S_0 \rightarrow [S]_1, \quad [S]_2 \rightarrow S_0. \quad (60)$$

Taking S_0 as its initial state, the device is now weakly equivalent to G . To convert this device to a transducer, we must supply it with rules stating the output symbol it produces as it moves from state Q_i to state Q_j reading symbol a_k . We take the output alphabet to be the set of non-terminal symbols of G^* (that is, the set of symbols of the form $[B_1 \dots B_n]_i$, which now designate states of the automaton). We shall say that when the device switches into the state Q it prints the output symbol Q . This completes the construction Ψ required in Theorem 34, which associates a finite transducer T with a nonself-embedding grammar G . If G generates x , T maps the input string x into the output string σ , where σ is a record of the successive states that T has traversed in accepting (generating) x . This sequence σ actually contains a complete account of a structural description assigned to x by G , and for each such structural description assigned to x by G there is a sequence of states σ , into which x is mapped by T , that preserves the structure of this structural description exactly. The point is that the names of the states of T actually contain information about certain subtrees of the labeled tree associated with x by G ; from the sequence of these states this labeled tree can be completely reconstructed. It is possible to construct a procedure Φ of the type specified in Def. 9 that will convert an output of T into a structural description (e.g., a labeled bracketing) of its input and conversely. Such a construction is carried out in detail by Langendoen (1961).⁷ Consequently, T as constructed by the procedure Ψ of Construction 59 and 60 is strongly equivalent to G .

The properties of the construction Ψ can be clarified by an example. Consider the grammar in (61), which meets the conditions assumed for the construction Ψ and Theorem 34:

$$\begin{aligned} S &\rightarrow AB, & A &\rightarrow SC, & B &\rightarrow DB, \\ A &\rightarrow a, & B &\rightarrow b, & C &\rightarrow c, & D &\rightarrow d. \end{aligned} \quad (61)$$

⁷ Langendoen, in fact, constructs a procedure Φ that operates in real time, as it were; that is to say, the labeled bracketing of a string x generated by G is produced by Φ from the output of the transducer T associated with G in the course of the computation of T on x .

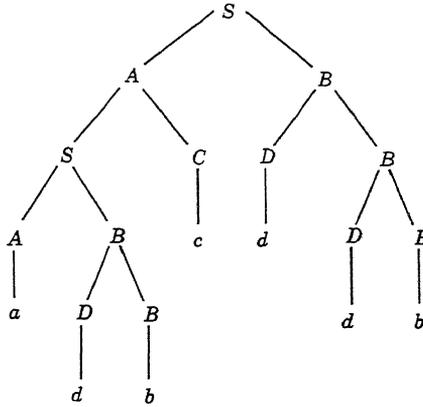


Fig. 10. Structural description for a sentence generated by the grammar (61).

This grammar generates the language consisting of all strings $a(d^i b c)^j d^k b$ and assigns to them such structural descriptions as the one shown in Fig. 10 (for the case $i = j = 1, k = 2$). Note that G contains both left-recursive and right-recursive elements, although it contains no self-embedding elements, and that, in this case, the left-recursive element A dominates the right-recursive element B . This example serves to illustrate the point, sometimes overlooked or misunderstood, that although the finite transducer T , which interprets sentences in the manner of G , of course reads these sentences from left-to-right in one pass, it does not follow that there must be any left-right asymmetry in the structural descriptions of the sentences that T accepts and interprets in the manner of G .

The class K of sequences constructed from the grammar G of (61) consists of the sequences (S) , (S, A) , (S, B) , (S, A, C) , and (S, B, D) . The construction Ψ now provides these rules:

$$\begin{aligned}
 & \left\{ \begin{array}{l} [S]_1 \rightarrow [SA]_1 \\ [SA]_2 \rightarrow [SB]_1 \\ [SB]_2 \rightarrow [S]_2 \end{array} \right\} && \text{(by step ii of Construction 59)} \\
 & [SA]_1 \rightarrow a[SA]_2 && \text{(by step i of Construction 59)} \\
 & \left\{ \begin{array}{l} [S]_2 \rightarrow [SAC]_1 \\ [SAC]_2 \rightarrow [SA]_2 \end{array} \right\} && \text{(by iv of Construction 59)} \\
 & [SB]_1 \rightarrow b[SB]_2 && \text{(by i of Construction 59)} \\
 & \left\{ \begin{array}{l} [SB]_1 \rightarrow [SBD]_1 \\ [SBD]_2 \rightarrow [SB]_1 \end{array} \right\} && \text{(by iii of Construction 59)} \\
 & \left\{ \begin{array}{l} [SAC]_1 \rightarrow c[SAC]_2 \\ [SBD]_1 \rightarrow d[SBD]_2 \end{array} \right\} && \text{(by i of Construction 59).}
 \end{aligned} \tag{62}$$

These constitute the grammar G^* provided by Ψ . Corresponding to Fig. 10, we have the derivation

$$\begin{array}{ll}
 [S]_1 & abc[SA]_2 \\
 [SA]_1 & abc[SB]_1 \\
 a[SA]_2 & abc[SBD]_1 \\
 a[SB]_1 & abcd[SBD]_2 \\
 a[SBD]_1 & abcd[SB]_1 \\
 ad[SBD]_2 & abcd[SBD]_1 \\
 ad[SB]_1 & abcdd[SBD]_2 \\
 adb[SB]_2 & abcdd[SB]_1 \\
 adb[S]_2 & abcddb[SB]_2 \\
 adb[SAC]_1 & abcddb[S]_2 \\
 abc[SAC]_2 &
 \end{array} \tag{63}$$

It is clear that the sequence of nonterminal symbols produced in this derivation enables us to reconstruct uniquely the structural description in Fig. 10. In fact, the automaton with the rules of Example 62 generates the sentence $abcddb$, essentially, by tracing systematically through the labeled tree of Fig. 10. This is a representative example; it illustrates how a device with finite memory can associate with each string x generated by a nonself-embedding normal grammar G the structural description assigned to x by G , where this structural description may be of arbitrary complexity.

Suppose now that we were to apply this construction to a self-embedding normal grammar G . Let us say that a transducer T generates (x, y) in the manner of G if T generates (x, y) in the sense previously defined (i.e., maps x into y while accepting x in the manner of a finite automaton), where $\Phi^{-1}(y)$ is a structural description assigned to x by G . Then the transducer $\Psi(G)$ constructed by Constructions 59 and 60 will, in fact, generate (x, y) in the manner of G for each pair (x, y) such that y is a structural description assigned to x by G , where y involves no self-embedding. Furthermore, by increasing the memory of T (conceptually, the easiest way to do this is to provide it with a bounded pushdown storage), we can allow it to relabel self-embedded symbols up to any bounded degree of self-embedding and then operate as before. In this case it can generate (x, y) in the manner of G for any pair (x, y) such that y is a structural description assigned to x by G that involves no more than some bounded degree of self-embedding. Beyond this we cannot go with a finite device, as we know from Theorem 33. It is clear from the results of Sec. 1.6 and 4.2 that if we allow the transducer an unbounded pushdown storage memory then it can be made

strongly equivalent to any given normal grammar G —observe, in particular that Constructions 18 and 19 in Sec. 4.2 are, in effect, the trivial special case of Construction 59 involving only steps i and ii.

These observations give us a precise indication of the extent to which sentences generated by a context-free grammar can be handled (i.e., accepted and interpreted) by a device with finite memory or a person with no (or fixed) supplementary aids to computation. We return to this question again in Chapter 13.

4.7 Definability of Languages by Systems of Equations

Suppose that G is a context-free grammar with nonterminals ordered as A_1, \dots, A_n , where A_1 is the designated initial symbol. With each A_i associate the set Σ_i of terminal strings dominated by A_i ; that is, $\Sigma_i = \{x \mid A_i \Rightarrow x\}$, using the notations to which we have adhered throughout. We thus associate with the grammar G the sequence of sets $(\Sigma_1, \dots, \Sigma_n)$, each a set of terminal strings, where Σ_1 is the terminal language generated by G . We say that this sequence of sets *satisfies* the grammar G , with the given ordering of nonterminals. Clearly, each term of the satisfying sequence is the terminal language generated by some context-free grammar, in fact, a grammar differing from G only in choice of initial symbol.

Suppose that we now regard the nonterminal symbols of G as variables ranging over sets of strings in the terminal vocabulary. We define a *polynomial expression* in the variables A_1, \dots, A_n as an expression of the form

$$\phi_1 + \dots + \phi_k, \tag{64}$$

where each ϕ_i is a string in V and the only nonterminal symbols appearing in Expression 64 are A_1, \dots, A_n . A polynomial expression such as Expression 64 can be regarded as defining a function f which maps a sequence of sets of strings in V_T onto a set of strings in V_T in the following manner. Given the sequence $(\Sigma_1, \dots, \Sigma_n)$, where Σ_i is a set of strings in V_T , let $f(\Sigma_1, \dots, \Sigma_n)$ be the set of strings formed by replacing each occurrence of A_i in Expression 64 by the symbol $\bar{\Sigma}_i$ designating the set Σ_i , then interpreting $+$ as set union and concatenation as set (Cartesian) product—that is to say, where A and B are sets, $AB = \{yz \mid y \in A \text{ and } z \in B\}$; $xA = \{xy \mid y \in A\}$; $Ax = \{yx \mid y \in A\}$. For example, the function $f(A, B)$ defined by the polynomial expression

$$a + Aa + BaA \tag{65}$$

maps the pair of sets $\{x, y\}, \{z, w\}$ onto the set $\{a, xa, ya, zax, zay, wax, way\}$.

Given the context-free grammar G with nonterminals A_1, \dots, A_n , we associate with each A_i the polynomial expression $\phi_1 + \dots + \phi_k$, where $A_i \rightarrow \phi_1, \dots, A_i \rightarrow \phi_k$ are all of the rules of G with A_i as the left-hand member. Consider now the system of equations

$$\begin{aligned} A_1 &= f_1(A_1, \dots, A_n) \\ &\vdots \\ &\vdots \\ A_n &= f_n(A_1, \dots, A_n), \end{aligned} \tag{66}$$

where f_i is the function defined by the polynomial expression associated with A_i . It is well known that such a system of equations has a unique minimal solution; that is, there is a unique sequence of sets, $\Sigma_1, \dots, \Sigma_n$, which satisfies this system of equations (as values for A_1, \dots, A_n , respectively), such that if $\Sigma_1', \dots, \Sigma_n'$ is another solution then $\Sigma_i \subset \Sigma_i'$ for each i . Furthermore, it is clear that the sequence of sets that constitutes the minimal solution for Eqs. 66 (what we shall henceforth call *the solution*) is the sequence of sets that satisfies G in the sense of the first paragraph of this section and Σ_1 is the terminal language generated by G .

Putting the same remark in different language, we can regard Eqs. 66 as defining a function f such that

$$f(A_1, \dots, A_n) = [f_1(A_1, \dots, A_n), \dots, f_n(A_1, \dots, A_n)]. \tag{67}$$

A *fixed point* of the function f is a sequence $(\Sigma_1, \dots, \Sigma_n)$ such that $f(\Sigma_1, \dots, \Sigma_n) = (\Sigma_1, \dots, \Sigma_n)$. Then there is a unique minimal fixed point of the function f , which is identical with the solution to Eqs. 66; that is, it is the sequence of sets that satisfies G .

The solution to Eqs. 66 can be determined by the following recursive procedure. We construct a sequence $\sigma_0, \sigma_1, \dots$, in which each term σ_i is an n -tuple of sets, in the following way: σ_0 is the n -tuple $(\emptyset, \dots, \emptyset)$, where \emptyset is the null set. For each $i \geq 0$, let $\sigma_{i+1} = f(\sigma_i)$. Where $\sigma_i = (\sigma_1^i, \dots, \sigma_n^i)$, define $\sigma = \lim_{i \rightarrow \infty} \sigma_i = (\sigma_1^\omega, \dots, \sigma_n^\omega)$, where $\sigma_j^\omega = \sum_k \sigma_j^k$.

Then σ is the solution to Eqs. 66. It is the minimal fixed point of f of Eq. 67, the sequence of sets satisfying G . And σ_1^ω is the terminal language generated by G . We say that each σ_i^ω is *definable* from Eqs. 66; a *definable language* is a set that is definable from some such system of equations. Clearly, the definable languages are exactly the context-free languages.

The point of view that we have just sketched is developed in Ginsburg & Rice (1962) and is the basis for the investigations carried out there and continued in Ginsburg & Rose (1963a, b). This work was motivated originally by an investigation of problem-oriented computer languages,

in particular, ALGOL, and has led to several interesting observations concerning these systems, to which we return in Sec. 4.8.

This approach to the study of context-free languages has been placed in a more general setting by Schützenberger (see in particular, Schützenberger, 1961c, 1962b, and Schützenberger & Chomsky, 1962). Suppose that we have a mapping f which assigns to each string x in V_T a non-negative integer $f(x)$. We can represent f as a formal power series r :

$$r = \sum_x \langle r, x \rangle x \tag{68}$$

in the elements $a_i \in V_T$, where the integral coefficient $\langle r, x \rangle = f(x)$. We say that the formal power series r is *characteristic* if, for each x , $\langle r, x \rangle$ is either 0 or 1, that is, if it represents a characteristic function. We define the *support* of the formal power series r [$= \text{sup}(r)$] as the set of strings x such that $\langle r, x \rangle \neq 0$. Thus we obtain the support of r by regarding $+$ as ordinary set union and nx , where n is the coefficient of x , as $x + \dots + x$, n times (which amounts to identifying nx with x for $n \neq 0$).

Note that a formal power series becomes an ordinary power series in the variables $a_i \in V_T$ if we regard them as commutative, that is, if we identify any two strings that can be obtained from one another by permutation.

The set of formal power series is closed under the following operations (among others):

- (i) Multiplication by an integer: the coefficient $\langle nr, x \rangle$ of x in nr is $n\langle r, x \rangle$.
- (ii) Addition: the coefficient $\langle r + r', x \rangle$ of x in $r + r'$ is $\langle r, x \rangle + \langle r', x \rangle$.
- (iii) Multiplication: the coefficient $\langle rr', x \rangle$ of x in rr' is obtained by factoring x into $yz = x$ in all possible ways and taking the sum of all the integers $\langle r, y \rangle \langle r', z \rangle$; that is, $\langle rr', x \rangle = \sum \langle r, y \rangle \langle r', z \rangle$, for all y, z such that $yz = x$. (69)

Note that addition is analogous to set union and multiplication, to formation of the set product. Thus we have

$$\text{sup}(r + r') = \text{sup}(r) \cup \text{sup}(r'), \quad \text{sup}(rr') = \text{sup}(r) \cdot \text{sup}(r'). \tag{70}$$

We can also define an operation analogous to set intersection, namely, the operation \otimes such that $r \otimes r'$ is the power series in which the coefficient $r \otimes r'$ of x is $\langle r, x \rangle \langle r', x \rangle$. There are also easily defined operations corresponding to universal closure and, in certain cases, to complement.

Given the grammar G with nonterminals A_1, \dots, A_n , let $f_i(x)$ be the number of different structural descriptions assigned to x by the grammar G_i formed by taking A_i as the initial symbol of G [we can, to make this

precise, take structural descriptions to be labeled bracketings generated by G in the manner described on p. 367, in which case $f_i(x)$ is the number of strings generated by G_i from which x can be formed by dropping brackets]. Let r_i be the formal power series that assigns to a string x the coefficient $\langle r_i, x \rangle = f_i(x)$. Thus $[\text{sup}(r_1), \dots, \text{sup}(r_n)]$ is the sequence of sets that satisfies G in the sense previously defined; $\text{sup}(r_1)$ is the terminal language $L(G)$ generated by G ; $\langle r_1, x \rangle = 0$ just in case $x \notin L(G)$; $\langle r_1, x \rangle = n$ just in case G provides n nonequivalent derivations, that is, n distinct structural descriptions, for x . We say that the sequence (r_1, \dots, r_n) satisfies G . The definition of satisfaction given previously is the special case in which we consider only those formal power series formed from ordinary formal power series by identifying all coefficients greater than zero.

We can, in fact, obtain the sequence (r_1, \dots, r_n) which satisfies G by an iterative procedure exactly as before. Regarding G again as the sequence of Eqs. 66, we construct an infinite sequence $\sigma_0, \sigma_1, \dots$, in which $\sigma_i = (r_1^i, \dots, r_n^i)$ and r_j^i is a formal power series (with, in fact, only a finite number of terms; that is, it is a polynomial in the terminal symbols of G). We again regard Eqs. 66 as defining a function f which, in this case, maps a sequence (r_1, \dots, r_n) into the sequence $[f_1(r_1, \dots, r_n), \dots, f_n(r_1, \dots, r_n)]$. The function f_i is defined, as before, by the polynomial expression $\phi_1 + \dots + \phi_k$, where $A \rightarrow \phi_i (1 \leq i \leq k)$ are all the rules in G that contain A on the left. We now interpret $+$ and concatenation not as set union and complex product but as the corresponding operations on power series, as in Definition 69. Take σ_0 as the sequence (r_1^0, \dots, r_n^0) , where each r_i^0 is the null power series in which every coefficient is zero. Take $\sigma_{i+1} = f(\sigma_i)$. As before, take $\sigma = \lim_{i \rightarrow \infty} \sigma_i = (r_1^\omega, \dots, r_n^\omega)$, where r_j^ω is defined as follows. Suppose that x is a string of length k and r_j^k is the j th term of σ_k , as above. Then the coefficient of x in the power series r_j^ω is determined by the following condition:

$$\langle r_j^\omega, x \rangle = \langle r_j^k, x \rangle. \quad (71)$$

In fact, it is not difficult to show that in the sequence $\sigma_0, \sigma_1, \dots$ the coefficients assigned to words of length k do not change past σ_k . Consequently, σ is well defined as the limit of this sequence; r_1^ω is the formal power series generated by G , where A_1 is the designated initial symbol of G and its support, $\text{sup}(r_1)$, is the language generated by G , in the former sense; r_1^ω also assigns to each string x belonging to $\text{sup}(r_1^\omega)$ its coefficient $\langle r_1^\omega, x \rangle$, which is a measure of the degree of ambiguity assigned to x by G . We speak of r_1^ω as *the* power series that satisfies G .

Suppose, for example, that we have the grammar G with the rules

$$A \rightarrow AA, \quad A \rightarrow a, \quad A \rightarrow b. \quad (72)$$

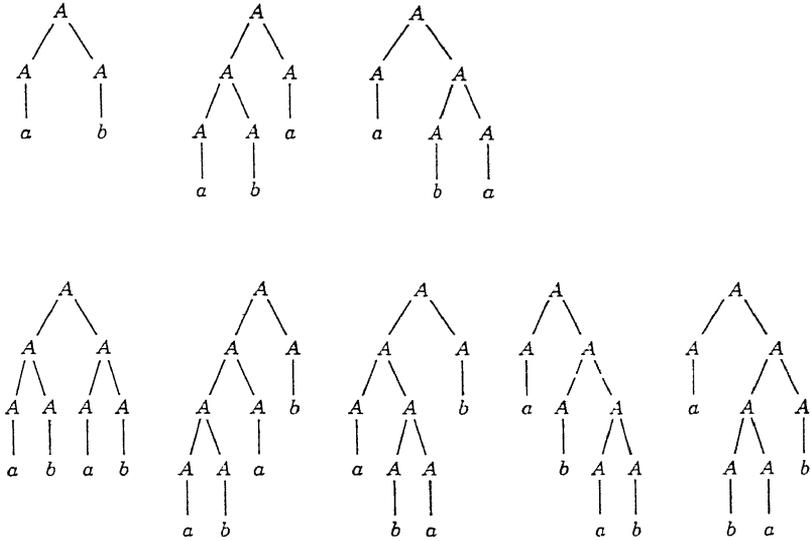


Fig. 11. *P*-markers illustrating the ambiguity of the grammar of (72).

This corresponds to the system of equations consisting of just

$$A = a + b + A^2. \tag{73}$$

In this extremely simple case we have $\sigma_i = (r_1^i)$, where

$$\begin{aligned} r_1^1 &= a + b + \emptyset^2 = a + b, \\ r_1^2 &= a + b + (r_1^1)^2 = a + b + (a + b)^2 \\ &= a + b + a^2 + ab + ba + b^2, \\ r_1^3 &= a + b + (r_1^2)^2 = \sum \langle r_1^3, x \rangle x, \text{ where} \\ &\quad \langle r_1^3, x \rangle = 1 \text{ for each string } x \text{ of length 1, 2 or 4,} \\ &\quad \langle r_1^3, x \rangle = 2 \text{ for each string } x \text{ of length 3,} \\ r_1^4 &= a + b + (r_1^3)^2, \text{ etc.} \end{aligned} \tag{74}$$

The coefficients of each string of length 3 will continue to be 2 in each r_1^i ($i > 3$), and the coefficients will increase with the length of the string. Thus in this grammar there is exactly one way to generate each string of length 2, exactly two ways to generate each string of length 3, exactly 5 ways to generate each string of length 4, etc., as can be seen from the examples in Fig. 11. The power series r_1 , which is the limit of the r_1^i 's, so defined, is the solution of Eq. 73. Its support is the terminal language $L(G)$ generated by G .

In this case $L(G)$ is the set of all strings in the alphabet $\{a, b\}$, and we know that there is an equivalent grammar G^* which will have as its solution a characteristic power series (i.e., with all coefficients = 0 or 1—in this case all = 1 except for the coefficient of e) with $L(G)$ as its support. An example is the grammar G^* represented as the equation

$$A = a + b + aA + bA. \quad (75)$$

In fact, we have observed more generally (Sec. 1.2, Theorem 1) that corresponding to every finite automaton there is a deterministic finite automaton. Rephrased in our present terminology, every regular language is generated by a grammar G that is satisfied by a characteristic formal power series. Clearly, the language generated by the grammar of Example 72 is a regular language, as we can see from the fact that it is generated by the one-sided linear grammar of Eq. 75. However, Theorem 29 of Sec. 4.5 asserts that there are context-free languages that cannot be represented in this way by a characteristic power series; that is, Theorem 29 asserts the existence of a language L which is the support of a power series r that satisfies a context-free grammar but which is not the support of any *characteristic* power series satisfying a context-free grammar.

As a second example to illustrate these notions, consider the following grammars:

$$S \rightarrow a + bSS, \quad (76)$$

$$S \rightarrow a + SbS. \quad (77)$$

Interpreting b as the sign for conditional and a as a variable, we see that Grammar 76 generates the set of well-formed formulas of the implicational calculus with one free variable in Polish parenthesis-free notation and correspondingly has a solution that is characteristic. Grammar 77 generates the set of strings of this calculus in ordinary notation, without parentheses, and its solution is the power series in which the coefficient of a string is the number of distinct ways in which it can be parenthesized to yield a well-formed formula of this system.

Schützenberger's notion of representing sets enumerated by a generative process in terms of formal power series is well motivated for the study of language. As has been mentioned several times, we are ultimately interested in studying processes that generate systems of structural descriptions rather than sets of strings; that is, we are ultimately interested in strong rather than weak generative capacity. The framework just sketched provides a first step toward this goal, since it takes account of the number of structural descriptions assigned to a string (though not the structural descriptions themselves). It also provides a particularly natural way of approaching the study of nondeterministic transduction. Recall that a

transducer can, in general, have two kinds of indeterminacy. When in state S_i reading the symbol a , it can have the option of switching to one of several states. If it switches to state S_j , it can have the further option of printing one of several strings on its output tape. Let us say that the string $x = b_1 \dots b_m$ (where b_i is a symbol of the input alphabet) carries the transducer T from state S_i to S_j with output $x = x_1 \dots x_m$ if T has the rules $(b_k, S_{i_k}) \rightarrow (S_{i_{k+1}}, x_k)$ for some i_1, \dots, i_{m+1} ($i_1 = i$; $i_{m+1} = j$) and for each $k \leq m$. Then a string x may carry T from S_i to S_j with many different outputs, and it may carry T from S_i to S_j with the output x in many different ways (i.e., with different factorizations of x). The natural way to represent the effect of the input string x in carrying T from S_i to S_j is therefore by a polynomial $\pi(x, i, j) = \sum \langle \pi(x, i, j), z \rangle z$, where $\langle \pi(x, i, j), z \rangle$ is the number of different ways in which z can be given as output as x carries T from S_i to S_j . We can then represent an n -state transducer T by a homomorphism μ mapping V_T into the ring of $n \times n$ matrices with polynomials in the output alphabet of T as entries. Then μx will be the matrix with entries $(\mu x)_{ij} = \pi(x, i, j)$, which represent the behavior of T as x carries it from S_i to S_j . Many problems involving transduction thus become problems in manipulation of matrices that can be handled by familiar techniques (cf. Schützenberger 1961a, 1962c). Moreover, several new questions suggest themselves in this more general framework. Thus we have restricted ourselves in this discussion to the positive power series that has only nonnegative coefficients. More generally, we can consider the *algebraic elements* (of the ring of power series), which have positive or negative coefficients and which satisfy systems of equations that may have negative coefficients in the polynomial expressions. We can think of a power series r with positive or negative integral coefficients as being the difference of two positive power series r' and r'' . The coefficient of the string x in r is the difference between the number of times that x is generated by the grammar corresponding to r' and the grammar corresponding to r'' . The support of r is the class of strings that is not generated the same number of times by these two grammars. Schützenberger has studied the family of formal power series $r = r' - r''$, where r' and r'' satisfy one-sided linear grammars and thus have regular languages as their supports (these are the formal power series that correspond to rational functions when we identify strings that differ only by permutations) and has characterized the supports of such formal power series in terms of acceptability by a certain class of restricted-infinite automata (cf. Schützenberger, 1961a). He has also shown that such an r may have as support a noncontext-free language and that there are some context-free languages that do not constitute the support of any such power series (Schützenberger, 1961c). For further discussion of these and

related questions, see these papers and Schützenberger & Chomsky (1962).

Relating to the questions of ambiguity for context-free languages raised in Sec. 4.5, we have the following general result concerning regular languages, which makes use of some of these notions.

Theorem 35. *Let G be a one-sided linear grammar satisfied by the formal power series r and generating the language $L = \text{sup}(r)$. Let $L_k = \{x \mid \langle r, x \rangle \leq k\}$. Then L_k is a regular language for each k . (Schützenberger, personal communication.)*

Let N be the set of nonnegative integers, k a fixed integer, and N^M the semiring of the $M \times M$ matrices with entries in N . It is proved in Schützenberger (1962c) that where G and r are as in the statement of Theorem 35 and U is the set of all strings on V_T (the free semigroup with generators $a \in V_T$) then

$$\begin{aligned} &\text{there is an } M < \infty \text{ and a homomorphism } \mu \\ &\text{of } U \text{ into } N^M \text{ such that } \langle r, x \rangle = (\mu x)_{1,M}. \end{aligned} \quad (78)$$

Let $K = \{i \mid 0 \leq i \leq k\}$ and $\beta: N \rightarrow K$ be defined by

$$\beta(n) = n, \quad \text{for } n \leq k; \quad \beta(n) = k, \quad \text{for } n > k. \quad (79)$$

Define an addition and a multiplication for K by setting

$$i \oplus j = \beta(i + j), \quad i \otimes j = \beta(ij). \quad (80)$$

K is a semiring, and it is easily shown that β is a homomorphism mapping N onto K . Let K^M be the set (in fact, semiring) of the $M \times M$ matrices with entries in K . Then β extends in a natural fashion to a homomorphism $\beta: N^M \rightarrow K^M$.

Define $\phi(x) = \beta[\mu(x)]$, where μ is as in Proposition 78. Thus $\phi(x)$ is an element of K^M for $x \in U$. Then, clearly,

$$\begin{aligned} (\phi x)_{1,M} &= \langle r, x \rangle, \quad \text{if } \langle r, x \rangle \leq k, \\ (\phi x)_{1,M} &= k, \quad \text{if } \langle r, x \rangle > k. \end{aligned} \quad (81)$$

But K^M is a multiplicative semigroup of finite cardinality, and, for $k' < k$, $L_{k'} = \{x \mid \langle r, x \rangle \leq k'\}$ (by definition) $= \{x \mid (\mu x)_{1,M} \leq k'\}$ (by Proposition 78) $= \{x \mid \phi(x) \in Q\}$, where Q is the subset of $\phi(U)$ containing $\phi(x)$ just in case $(\phi x)_{1,M} \leq k'$. It is well known that where ψ is a homomorphism mapping a subset of U into a finite semigroup H , then $\psi^{-1}(H) = \{x \mid \psi(x) \in H\}$ is a regular language. Consequently, $L_{k'}$ is a regular language.

From the fact that L_k is regular for each k , it follows by elementary properties of regular sets (cf. Theorem 1) that for each k the set of strings x such that $\langle r, x \rangle = k$ and the set of strings x such that $\langle r, x \rangle \geq k$ are each regular. In particular, the set of strings x such that $\langle r, x \rangle \geq 2$ is a regular

language. This is the set of strings that is ambiguous with respect to G in the sense of Sec. 4.5.

Suppose, in particular, that $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$ are two sets of strings. Define L_X as the set of all strings $z = x_{k_1} \dots x_{k_s}$ ($k_s \leq n$), that is, as the set of all strings factorizable into strings of X —in the notation of Sec. 1.2, L_X would be represented $(x_1, \dots, x_n)^*$. Define L_Y similarly. Let us call X a *code* (cf. Chapter 11, Sec. 2) if each string of L_X is uniquely factorizable (decipherable) into members of X (similarly, Y). Now consider the grammar G_X with the rules $S \rightarrow x_i S$ and G_Y with the rules $S \rightarrow y_i S$ (and the rule $S \rightarrow e$). By Theorem 35 the set of strings which are ambiguous with respect to G_X (similarly, with respect to G_Y) is regular, and consequently there is a procedure for determining whether it is empty. Hence there is an algorithm for determining whether a set of strings constitutes a code (cf. Sardinas & Patterson, 1953). However, we cannot decide whether X and Y fail to be codes in the same way (cf. Sec. 4.5).

Suppose now that G and r are as in Theorem 35 and that G' is a second one-sided linear grammar satisfied by r' . It follows from a theorem of Markov and from Proposition 78 that there is no algorithm for determining in an arbitrary case of this sort whether there is an x such that $\langle r, x \rangle = \langle r', x \rangle$. Theorem 35 implies that, given k , there is an algorithm for determining whether $L_k \cap L'_k$ is nonempty, where $L_k = \{x \mid \langle r, x \rangle \leq k\}$ and $L'_k = \{x \mid \langle r', x \rangle \leq k\}$. We see, then, that there is no algorithm for determining whether there is a k such that $L_k \cap L'_k$ is nonempty (Schützenberger, personal communication).

4.8 Programming Languages

A program for a digital computer can be regarded as a string of symbols in some fixed alphabet. A programming language can be regarded as an infinite set of strings, each of which is a program. A programming language has a grammar that specifies precisely the alphabet and the set of techniques for constructing programs. Ginsburg and Rice have pointed out that the language ALGOL has a context-free grammar, though not a one-sided linear or even a sequential grammar. This observation suggests that it might be of some interest to interpret the results obtained in the general study of context-free languages, taking them as constituting a class of potential “problem oriented” programming languages.

Note, in particular, that a programming language must have an unambiguous grammar in the sense defined in Sec. 4.5, above. If the set of techniques that is available for the construction of programs constitutes

an ambiguous grammar, then the programmer may construct a program that he intends the machine to interpret in a certain way, but the machine may interpret it in quite a different way. We have seen, however, that there are certain context-free languages that are inherently ambiguous with respect to the class of context-free grammars (Theorem 29, Sec. 4.5). Hence there is at least an abstract possibility that a certain infinite class of "programs" may not be characterizable by an unambiguous grammar when the techniques for constructing programs are limited to those expressible within the framework of context-free grammar. Furthermore, we have observed that there is no algorithm for determining whether a context-free grammar is ambiguous (Theorem 28, Sec. 4.5). Thus in particular cases the problem of determining whether a given grammar is ambiguous (whether a proposed programming language is minimally adequate) may be quite a difficult one.

Consider now the problem of translating from a programming language L_1 into another language L_2 (e.g., machine code or another higher order programming language). We can regard this as the problem of constructing a finite transducer (a "compiler") T such that $T(L_1) = L_2$. There is no reason to assume, in general, that such a transducer exists. Furthermore, we have seen that the general problem of determining for given context-free languages L_1 and L_2 whether there exists a transducer T such that $T(L_1) = L_2$ is recursively unsolvable (Theorem 27, Sec. 4.4). Hence the problem of translating between arbitrary systems of this sort seems to raise potentially quite difficult questions (Ginsburg, personal communication).

5. CATEGORIAL GRAMMARS

Traditional grammatical analysis is concerned with the division of sentences into phrases and subphrases, down to word categories, where these phrases belong to a finite number of types (noun phrases, predicates, nouns, etc.). In the last twenty years there have been various attempts in descriptive linguistics to codify and clarify the traditional approach. One might mention here in particular Harris (1946), elaborated further in Harris (1951, Chapter 16), Wells (1947), and the recent work of Pike and his colleagues in Tagmemics (cf., e.g., Elson & Pickett, 1960). The generative systems studied in Secs. 3 and 4 represent one attempt to give precise expression to some of these ideas. There have been several other, more or less related, approaches which we mention here only briefly.

Several attempts have been made to develop systematic procedures that might lead from a set of sentences to a categorization of substrings into

phrase types (e.g., Harris, 1946; Harris, 1951; Chomsky, 1953; Kulagina, 1958). These approaches are conceptually related to one another and to the systems we have discussed, but the exact nature of this relation has not been explored. For a somewhat different approach to systematic categorization of phrase types see Hiž (1961).

A second approach arose from the theory of semantical categories of Lesniewski, which was developed for the study of formalized languages. A modification, based on the formulation in Ajdukiewicz (1935), was suggested by Bar-Hillel (1953) as a precise explication of the immediate constituent analysis of recent linguistics. Lambek (1958, 1959, 1961) has also developed several systems of this general type, with certain additional modifications, and he has examined their applicability to linguistic material. Similar approaches are also discussed in Wundheiler and Wundheiler (1955), Suszko (1958), Curry and Feys (1958), and Curry (1961). We follow here the exposition in Bar-Hillel, Gaifman, and Shamir (1960).

We can establish a system of categories in the following way. Select a finite number of *primitive* categories (e.g., the category s of sentences and the category n of nouns and noun phrases, which were the only primitive categories envisaged in the system of Ajdukiewicz). All primitive categories are categories. When α and β are categories, then $[\alpha/\beta]$ and $[\alpha\backslash\beta]$ are also categories—call them *derived* categories. Thus we can have such categories as $[n/s]$, $[s/[n/s]]$, and $[[n/n]\backslash[s\backslash n]]$. These are the only categories. Each member of V_T is assigned to one or more categories. The set of categories to which elements of V_T are assigned, with a list of their members, constitutes the grammar G . We have the following two *rules of resolution*:

- (i) Resolve a sequence of two category symbols of the form $[\alpha/\beta]$, β to α .
- (ii) Resolve a sequence of two category symbols of the form α , $[\alpha\backslash\beta]$ to β .

These rules suggest cancellation in arithmetic, which was, in fact, the motivation for the notation.

Given a string x of elements of V_T , replace each symbol of V_T in x by its category symbol, thus giving a sequence of category symbols. There may be several such associated sequences of category symbols, since a member of V_T may belong to several categories. Denote these sequences as $C_1(x)$, \dots , $C_n(x)$. By successive application of the rules of resolution to $C_i(x)$, we may find either that $C_i(x)$ resolves ultimately to s or that it resolves ultimately to some sequence of (one or more) category symbols distinct from s . If, for some i , $C_i(x)$ resolves to s , we say that the grammar G *generates* x ; if there is no such i , G does not generate x . The set of strings generated by G is the *language generated by* G . As in the case of the other generative

grammars we have discussed, it is merely a notational question whether we think of G as a grammar that generates sentences or that accepts strings as inputs and determines whether they are sentences (i.e., a recognition device). Generally, the latter phraseology has been used for categorial grammars of the type just described.

The functioning of such a grammar can be clarified by an example. Suppose that our grammar contains the primitive categories n and s , the words *John*, *Mary*, *loves*, *died*, *is*, *old*, *very*, and the following category assignment: *John*, *Mary* to n ; *died* to $[n \setminus s]$; *loves* to $[n \setminus s]/n$; *old* to $[n/n]$; *very* to $[n/n]/[n/n]$; *is* to $[n \setminus s]/[n/n]$. Thus intransitive verbs (such as *died*) are regarded as "operators" that "convert" nouns appearing to their left to sentences; transitive verbs (*loves*) are regarded as operators that convert nouns appearing to their right to intransitive verbs; adjectives are regarded as operators that convert nouns appearing to their right to nouns; *very* is regarded as an operator that converts an adjective appearing to its right to an adjective; *is* is regarded as an operator that converts an adjective appearing to its right to an intransitive verb. Such strings as the following resolve to s as indicated:

$$(i) \text{ John died} \\ \frac{n, [n \setminus s]}{s}$$

$$(ii) \text{ John loves Mary} \\ \frac{n, \frac{[n \setminus s]/n, n}{[n \setminus s]}}{s}$$

(83)

$$(iii) \text{ John is very old} \\ \frac{n, [n \setminus s]/[n/n], \frac{[n/n]/[n/n], [n/n]}{[n/n]}}{[n \setminus s]} \\ \frac{\quad}{s}$$

A grammar of the type just described we call a *bidirectional categorial grammar*. If all of the derived categories of the grammar are of the type $[\alpha \setminus \beta]$ or if all are of the type $[\alpha/\beta]$, we call the system a *unidirectional categorial grammar*. Ajdukiewicz considered only the second form, since he was primarily concerned with systems using Polish parenthesis-free notation, in which functors precede arguments.

It is possible, of course, to regard both unidirectional and bidirectional

categorial systems as generative grammars, and we can ask how they are related to one another and to the systems we have discussed. Bar-Hillel, Gaifman, and Shamir (1960) have shown the following:

Theorem 36. *The families of unidirectional categorial grammars, bidirectional categorial grammars, and context-free grammars are weakly equivalent.*

If G is a bidirectional categorial grammar, there is a context-free grammar that generates the language generated by G ; and if G is a context-free grammar there is a unidirectional categorial grammar that generates the language generated by G . From this follows the somewhat surprising corollary that the class of unidirectional categorial grammars is equal in generative capacity to the full class of bidirectional categorial grammars.

Shamir has recently observed (personal communication) that Theorem 36 can be established by a proof very much like that of the proof of equivalence of context-free grammars and PDS automata.

It should be emphasized that the relation studied in Theorem 36 is weak equivalence. It does not follow that, given a grammar of one of these kinds, a grammar of one of the other kinds can be found that will involve a category assignment of comparable complexity or naturalness or that will assign the same bracketing (constituent structure) to substrings. It seems, in fact, that for those subparts of actual languages that can be described in a fairly natural way by context-free grammars, a corresponding description in terms of bidirectional categorial systems becomes complex fairly rapidly (and, of course, a natural description with a unidirectional categorial grammar is generally quite out of the question).

The systems that Lambek has developed differ in several respects from the one just described—in particular, they allow a greater degree of flexibility in category assignment. Thus his rules of resolution assert that a category α is also at the same time a category of the form $\beta/[\alpha\backslash\beta]$, so that in this and other ways it is possible to increase the complexity and length of the sequence of category symbols associated with a string by application of rules of resolution. Consequently, it is not immediately obvious, as it is in the case of the system just sketched, that the language generated is recursive. Lambek has shown, however, that the systems he has studied are, in fact, decidable. It is not known how Lambek's system is related to bidirectional categorial systems or context-free grammars, although one would expect to find that the relation is quite close, perhaps as close as weak equivalence.

The interest of the various kinds of categorial grammars is that they contain no grammatical rules beyond the lexicon; that is to say, where G is an assignment of the words of a finite vocabulary V_T to a finite number of categories, primitive and derived, it is possible to determine for each

string x on the vocabulary V_T whether G generates x by a computational procedure that uses the rules of resolution, which are uniform for all grammars of the given type, hence need not be stated as part of the grammar G . There is, in fact, a traditional view that identifies grammar with the set of grammatical properties of words or morphemes (cf. de Saussure, 1916, p. 149), and it might reasonably be maintained that the approach just outlined gives one precise expression to this notion.

Matthews has recently investigated a generalization of the theory of constituent-structure grammar in which certain types of discontinuity are permitted (Matthews, 1963b). Continuing to follow the notational convention of Chapter 11, Sec. 4, let us consider rules of the form $A \rightarrow \varphi_1[n]\varphi_2$, where $n \geq 0$. We interpret such a rule as applying to a string $\psi A \alpha_1 \dots \alpha_n \chi$ to form $\psi \varphi_1 \alpha_1 \dots \alpha_n \varphi_2 \chi$ (where $\alpha_i \neq e$) and as applying to a string $\psi A \alpha_1 \dots \alpha_m$ ($m < n$) to form $\psi \varphi_1 \alpha_1 \dots \alpha_m \varphi_2$. Under this convention, we can regard a context-free grammar as one containing only rules of the form $A \rightarrow \varphi_1[0]\varphi_2$. He has also generalized this in a natural way to the case of discontinuous context-sensitive rules. For any grammar, we now define a *left-to-right derivation* in the manner presented explicitly in Sec. 4.2, p. 373, and a *left-to-right discontinuous grammar* as a grammar with rules of the form just given (or of the more general context-sensitive discontinuous type) and with rule applications so restricted as to allow only left-to-right derivations, and all of these (cf. Sec. 4.2). Matthews has shown that a left-to-right discontinuous grammar can generate only context-free languages, so that these generalizations do not increase generative capacity. Obviously, the same is therefore true of *right-to-left discontinuous grammars* which provide derivations in the manner also described on p. 373 (i.e., only the right-most nonterminal symbol is rewritten at each stage) and in which a rule of the form $A \rightarrow \varphi_1[n]\varphi_2$ is interpreted as placing $\varphi_1 n$ symbols to the left (or to the extreme left) as A is rewritten, instead of n symbols to the right (or to the extreme right) as in the case of a left-to-right discontinuous grammar (similarly for context-sensitive discontinuous rules). Matthews has also extended this result to rules which permit multiple discontinuities and has observed that allowing even two-way discontinuous rules does not extend the capacity of context-sensitive grammars.

Various other models of linguistic structure have been proposed, but insofar as they can be interpreted as specifying a form of generative grammar (i.e., insofar as they specify grammars that provide information about sentence structure in an explicit manner) they seem to fall largely within the scope of the theory of constituent-structure grammar, or even, quite generally, the theory of context-free grammar. For discussion, see Gross (1962) and Postal (forthcoming).

This concludes our survey of formal properties of grammars. It is hardly necessary to stress the preliminary character of most of these investigations. As is apparent from the appended bibliography, the whole subject is, properly speaking, only five or six years old, and much of this survey has in fact dealt with work in progress. It is important to reiterate that the systems that have so far proved amenable to serious abstract study are undoubtedly inadequate to represent the full complexity and richness of the syntactic devices available in natural language, in particular, because of the restriction to rewriting systems that do not incorporate grammatical transformations of the kind discussed in Chapter 11, Sec. 5. Nevertheless, they do appear to have the scope of the theories of grammatical structure that have been proposed in traditional and modern linguistics or in recent work on computable sentence analysis or that are implicit in traditional and modern descriptive studies, with the exception of the theoretical and descriptive studies involving transformations. Certain basic properties of natural languages (e.g., bracketing into continuous phrases, categorization into lexical and phrase types, nesting of dependencies) appear in systems of the kind that we have surveyed. Hence the study of these systems has some direct bearing on the character of natural language. Furthermore, it is clear that profitable abstract study of systems as rich and intricate as natural language, or of organisms sufficiently complex to master and use such systems, will require sharper tools and deeper insights into formal systems than we now possess, and these can be acquired only through study of language-like systems that are simpler than the given natural languages. Whether these richer systems will yield to serious abstract study is, of course, a question about which we can at present only speculate.

References

- Ajdukiewicz, K. Die syntaktische Konnexität. *Studia Philosophica*, 1935, 1, 1-27.
- Bar-Hillel, Y. A quasi-arithmetical notation for syntactic description. *Language*, 1953, 29, 47-58.
- Bar-Hillel, Y., Gaifman, C., & Shamir, E. On categorial and phrase structure grammars. *Bull. Res. Council of Israel*, 9F, 1960, 1-16.
- Bar-Hillel, Y., Perles, M., & Shamir, E. On formal properties of simple phrase structure grammars. Tech. Rept. No. 4, Office of Naval Research, Information Systems Branch, 1960. (Also published in *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 1961, 14, 143-172.)
- Bar-Hillel, Y., & Shamir, E. Finite state languages: formal representation and adequacy problems. *Bull. Res. Council of Israel*, 1960, 8F, 155-166.
- Chomsky, N. Systems of syntactic analysis. *J. Symbolic Logic*, 1953, 18, 242-256.
- Chomsky, N. Three models for the description of language. *IRE Trans. on Inform. Theory*, 1956, IT-2, 113-124.

- Chomsky, N. On certain formal properties of grammars. *Information and Control*, 1959, 2, 137-167. (a)
- Chomsky, N. A note on phrase structure grammars. *Information and Control*, 1959, 2, 393-395. (b)
- Chomsky, N. On the notion "Rule of grammar." In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Sympos. in Appl. Math.* Providence, R.I.: American Mathematical Society, 1961. Pp. 6-24. Reprinted in J. Katz & J. Fodor (Eds.), *Readings in the Philosophy of Language*. New York: Prentice-Hall, 1963.
- Chomsky, N. Context-free grammars and pushdown storage. *RLE Quart. Prog. Rept. No. 65*. Cambridge, Mass.: M.I.T. March 1962. (a)
- Chomsky, N. The logical basis for linguistic theory. *Proc. IXth Int. Cong. of Linguists*, 1962. (b). Reprinted in J. Katz & J. Fodor (Eds.), *Readings in the Philosophy of Language*. New York: Prentice-Hall, 1963.
- Chomsky, N., & Miller, G. A. Finite state languages. *Information and Control*, 1958, 1, 91-112.
- Čulík, K. Some notes on finite state languages and events represented by finite automata using labelled graphs. *Časopis pro pěstování matematiky*, 1961, 86, 43-55 (Prague).
- Curry, H. Some logical aspects of grammatical structure. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Sympos. in Appl. Math.* Providence, R.I.: American Mathematical Society, 1961. Pp. 56-68.
- Curry, H., & Feys, R. *Combinatory logic*. Amsterdam: North-Holland, 1958.
- Davis, M. *Computability and unsolvability*. New York: McGraw-Hill, 1958.
- Elson, B., & Pickett, V. B. *Beginning morphology-syntax*. Summer Institute of Linguistics, Santa Ana, Calif., 1960.
- Floyd, R. W. Mathematical induction on phrase structure grammars. *Information and Control*, 1961, 4, 353-358.
- Ginsburg, S., & Rice, H. G. Two families of languages related to ALGOL. *J. Assoc. Computing Machinery*, 1962, 10, 350-371.
- Ginsburg, S., & Rose, G. F. Some recursively unsolvable problems in ALGOL-like languages. *J. Assoc. Computing Mach.*, 1963, 10, 29-47. (a)
- Ginsburg, S., & Rose, G. F. Operations which preserve definability in languages. *J. Assoc. Computing Mach.*, 1963, 10, 175-195. (b)
- Greibach, S. Undecidability of the ambiguity problem for minimal linear grammars. *Information and Control* (in press).
- Gross, M. *On the equivalence of models of languages used in the fields of mechanical translation and information retrieval*. Mimeographed. Cambridge: Mass. Inst. Tech., 1962.
- Harris, Z. S. From morpheme to utterance. *Language*, 1946, 22, 161-183.
- Harris, Z. S. *Methods in structural linguistics*. Chicago: Univer. of Chicago Press, 1951.
- Harris, Z. S. Co-occurrence and transformation in linguistic structure. *Language*, 1957, 33, 283-340.
- Hiž, H. Congrammaticality. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Sympos. in Appl. Math.* Providence, R.I.: American Mathematical Society, 1961. Pp. 43-50.
- Katz, J., & Fodor, J. The structure of a semantic theory. To appear in *Language*. Reprinted in J. Katz & J. Fodor (Eds.), *Readings in the Philosophy of Language*. New York: Prentice-Hall, 1963.
- Kleene, S. C. Representation of events in nerve nets and finite automata. In C. E.

- Shannon & J. McCarthy (Eds.), *Automata Studies*. Princeton: Princeton Univer. Press, 1956. Pp. 3-41.
- Köhler, W. *The place of value in a world of fact*. New York: Liveright, 1938.
- Kulagina, O. S. Ob odnom sposobe opredelenija grammatičeskix panjatij na baze teorij množestv. (On one method of defining grammatical categories on the basis of set theory.) *Problemy kibernetiki*, 1, Moscow, 1958.
- Lambek, J. The mathematics of sentence structure. *Amer. Math. Monthly*, 1958, **65**, 154-170.
- Lambek, J. Contributions to a mathematical analysis of the English verb phrase. *J. Canadian Linguistic Assoc.*, 1959, **5**, 83-89.
- Lambek, J. On the calculus of syntactic types. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Symp. in Appl. Math.* Providence, R.I.: American Mathematical Society, 1961. Pp. 166-178.
- Landweber, P. S. Three theorems on phrase structure grammars of type 1. *Information and Control* (in press).
- Langendoen, T. *Structural descriptions for sentences generated by non-self-embedding constituent grammars*. Undergraduate Honors Thesis, Mass. Inst. of Tech., 1961.
- Lashley, K. S. Learning: I. Nervous mechanisms of learning. In C. Murchison (Ed.), *The foundations of experimental psychology*. Worcester, Mass.: Clark Univer. Press, 1929. Pp. 524-563.
- Lashley, K. S. The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior*. New York: Wiley, 1951. Pp. 112-136.
- Matthews, G. H. *Hidatsa syntax*. Mimeographed. Cambridge; Mass. Inst. Tech., 1962a.
- Matthews, G. H. Discontinuity and asymmetry in phrase structure grammars. *Information and Control* (in press).
- Matthews, G. H. A note on asymmetry of phrase structure grammars. *Information and Control* (in press).
- Miller, G. A., & Selfridge, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, **63**, 176-185.
- Myhill, J. *Linear bounded automata*. WADD Technical note 60-165. Wright Air Development Division, Wright-Patterson Air Force Base, Ohio, 1960.
- McNaughton, R. The theory of automata: a survey. In F. L. Alt (Ed.), *Advances in computers*, Vol. 2. New York: Academic Press, 1961.
- McNaughton, R., & Yamada, H. Regular expressions and state graphs for automata. *IRE Trans. on Electronic Computers*, 1960, **EC-9**, 39-47.
- Newell, A., Shaw, J. C., & Simon, H. A. Report on a general problem-solving program. In *Information Processing. Proc. International Conference on Information Processing, UNESCO*, Paris, June 1959. Pp. 256-264.
- Oettinger, A. Automatic syntactic analysis and the pushdown store. In R. Jakobson (Ed.), *Structure of language and its mathematical aspects, Proc. 12th Sympos. in Appl. Math.* Providence, R.I.: American Mathematical Society, 1961. Pp. 104-129.
- Parikh, R. Language-generating devices. *RLE Quart. Prog. Rept.*, No. 60, Cambridge, Mass.: M.I.T. January 1961, 199-212.
- Post, E. A variant of a recursively unsolvable problem. *Bull. Amer. Math. Soc.*, 1946, **52**, 264-268.
- Postal, P. On the limitations of context-free phrase structure description. *RLE Quart. Prog. Rept.*, No. 64, Cambridge, Mass.: M.I.T. January 1962, 231-238.
- Postal, P. Constituent analysis. *Int. J. Amer. Linguistics*, Supplement (to appear).
- Rabin, M., & Scott, D. Finite automata and their decision problems. *IBM J. Res. Develop.*, 1959, **3**, 114-125.

- Ritchie, R. W. *Classes of recursive functions of predictable complexity*. Doctoral dissertation, Dept. Math., Princeton Univer., 1960.
- Rogers, H. *Recursive functions and effective computability*. Mimeographed, Dept. Math., Mass. Inst. Tech., 1961.
- Sardinas, A. A., & Patterson, G. W. A necessary and sufficient condition for unique decipherability of coded messages. *IRE Convention Record*, 1953, 8, 104-108.
- Saussure, F. de. *Cours de linguistique générale*, Paris: 1916. (Translation by W. Baskin, *Course in general linguistics*, New York: Philosophical Library, 1959).
- Scheinberg, S. *Some properties of constituent structure grammars*. Unpublished paper, 1960. (a)
- Scheinberg, S. Note on the Boolean properties of context-free languages. *Information and Control*, 1960, 3, 372-375. (b)
- Schützenberger, M. P. *Un problème de la théorie des automates*. Séminaire Dubreil-Pisot, Paris, December 1959.
- Schützenberger, M. P. A remark on finite transducers. *Information and Control*, 1961, 4, 185-196. (a)
- Schützenberger, M. P. On the definition of a family of automata. *Information and Control*, 1961, 4, 245-270. (b)
- Schützenberger, M. P. Some remarks on Chomsky's context-free languages. *RLE Quart. Prog. Rept. No. 63*, Cambridge, Mass.: M.I.T. October 1961, 155-170. (c)
- Schützenberger, M. P. *On a family of formal power series*. Mimeographed, 1962. (a)
- Schützenberger, M. P. Certain families of elementary automata and their decision problems. To appear in *Proc. Sympos. on Math. Theory Automata*, Vol. XII, MRI Symposia Series, 1962. (b)
- Schützenberger, M. P. On a theorem of Jungen. *Proc. Amer. Math. Soc.*, 1962, 13, 885-890. (c)
- Schützenberger, M. P. *On context-free languages and push-down automata*. Research paper RC-793 of IBM Res. Lab., Yorktown Heights, New York, 1962. (d)
- Schützenberger, M. P. Finite counting automata. *Information and Control*, 1962, 5, 91-107. (e)
- Schützenberger, M. P., & Chomsky, N. The algebraic theory of context-free languages. *Computer programming and formal systems*. Amsterdam: North-Holland, 1963. Pp. 118-161.
- Shamir, E. *On sequential languages*. Tech. Rept. No. 7, Office of Naval Research, Information Systems Branch, 1961.
- Shamir, E. A remark on discovery algorithms for grammars. *Information and Control*, 1962, 5, 246-251.
- Shannon, C. E., & Weaver, W. *The mathematical theory of communication*. Urbana: University of Illinois Press, 1949.
- Shepherdson, J. C. The reduction of two-way automata to one-way automata. *IBM J. Res. Develop.*, 1959, 3, 198-200.
- Suszko, R. Syntactic structure and semantical reference I. *Studia logica*, 1958, 8, 213-244.
- Tolman, E. C. *Purposive behavior in animals and men*. New York: Appleton-Century-Crofts, 1932.
- Wells, R. Immediate constituents. *Language*, 1947, 23, 81-117.
- Wundheiler, L., & Wundheiler, A. Some logical concepts for syntax. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages*. Cambridge: Technology Press and Wiley, 1955. Pp. 194-207.
- Yamada, H. *Counting by a class of growing automata*. Doctoral dissertation, Univer. of Pennsylvania, Philadelphia, 1960.

13

Finitary Models of Language Users¹

George A. Miller
Harvard University

Noam Chomsky
Massachusetts Institute of Technology

1. *The preparation of this Chapter was supported in part by the U.S. Army, the Air Force Office of Scientific Research, and the Office of Naval Research; and in part by the National Science Foundation (Grants No. NSF G-16486 and No. NSF G-13903).*

Contents

| | |
|--|-----|
| 1. Stochastic Models | 421 |
| 1.1. Markov sources, | 422 |
| 1.2. k -Limited stochastic sources, | 427 |
| 1.3. A measure of selective information, | 431 |
| 1.4. Redundancy, | 439 |
| 1.5. Some connections with grammaticalness, | 443 |
| 1.6. Minimum-redundancy codes, | 450 |
| 1.7. Word frequencies, | 456 |
| 2. Algebraic Models | 464 |
| 2.1. Models incorporating rewriting systems, | 468 |
| 2.2. Models incorporating transformational grammars, | 476 |
| 3. Toward a Theory of Complicated Behavior | 483 |
| References | 488 |

Finitary Models of Language Users

In this chapter we consider some of the models and measures that have been proposed to describe talkers and listeners—to describe the users of language rather than the language itself. As was pointed out at the beginning of Chapter 12, our language is not merely the collection of our linguistic responses, habits, or dispositions, just as our knowledge of arithmetic is not merely the collection of our arithmetic responses, habits, or dispositions. We must respect this distinction between the person's knowledge and his actual or even potential behavior; a formal characterization of some language is not simultaneously a model of the users of that language.

When we turn to the description of a user, a severe constraint is placed on our formulations. We have seen that natural languages are not adequately characterized by one-sided linear grammars (finite automata), yet we know that they must be spoken and heard by devices with bounded memory. How might this be accomplished? No automaton with bounded memory can produce all and only the grammatical sentences of a natural language; every such device, man presumably included, will exhibit certain limitations.

In considering models for the actual performance of human talkers and listeners an important criterion of adequacy and validity must be the extent to which the model's limitations correspond to our human limitations. We shall consider various finite systems—both stochastic and algebraic—with the idea of comparing their shortcomings with those of human talkers and listeners. For example, the fact that people are able to produce and comprehend an unlimited variety of novel sentences indicates immediately that their capacities are quite different from those of an automaton that compiles a simple list of all the grammatical sentences it hears. This example is trivial, yet it illustrates the kind of argument we must be prepared to make.

1. STOCHASTIC MODELS

It is often assumed, usually by workers interested in only one aspect of communication, that our perceptual models for a listener will be rather different from any behavioral models we might need for a speaker.

That assumption was not adopted in our discussion of formal aspects of linguistic competence, and it will not be adopted here in discussing empirical aspects of linguistic performance. In proposing models for a *user* of language—a user who is simultaneously talker and listener—we have assumed instead that the theoretically significant aspects of verbal behavior must be common to both the productive and receptive functions.

Once a formal theory of communication or language has been constructed, it generally turns out to be equally useful for describing both sources and receivers; in order to describe one or the other we simply rename various components of the formal theory in an appropriate fashion. This is illustrated by the stochastic theories considered in this section.

Stochastic theories of communication generally assume that the array of message elements can be represented by a probability distribution and that various communication processes (coding, transmitting, and receiving) have the effect of operating on that a priori distribution to transform it according to known transitional probabilities into an a posteriori distribution. The basic mathematical idea, therefore, is simply the multiplication of a vector by a matrix. But the interpretation we give to this underlying mathematical structure differs, depending on whether we interpret it as a model of a source, a channel, or a receiver. Thus the distinction between talkers and listeners is in no way critical for the development of the basic stochastic theory of communication. The same neutrality also characterizes the algebraic models of the user that are discussed in Sec. 2 of this chapter.

Purely for expository purposes, however, it is often convenient to present the mathematical argument in a definite context. For that reason we have arbitrarily chosen here to interpret the mathematics as a model of the source. This choice should not be taken to mean that a stochastic theory of communication must be concerned solely, or even principally, with speakers rather than with transmitters or hearers. The parallel development of these models for a receiver would be simply redundant, since little more than a substitution of terms would be involved.

1.1 Markov Sources

An important function of much communication is to reduce the uncertainty of a receiver about the state of affairs existing at the source. In such task-oriented communications, if there were no uncertainty about what a talker would say, there would be no need for him to speak. From a receiver's point of view the source is unpredictable; it would seem to

be a natural strategy, therefore, to describe the source in terms of probabilities. Moreover, the process of transmission is often exposed to random and unpredictable perturbations that can best be described probabilistically. The receiver himself is not above making errors; his mistakes can be a further source of randomness. Thus there are several valid motives for the development of stochastic theories of communication.

A stochastic theory of communication readily accommodates an infinitude of alternative sentences. Indeed, there would seem to be far more stochastic sequences than we actually need. Since no grammatical sentence is infinitely long, there can be at most only a countable infinitude of them. In probability theory we deal with a random sequence that extends infinitely in both directions, past and future, and we consider the uncountable infinitude of all such sequences that might occur.² The events with which probability theory deals are subsets of this set of all sequences. A finite stochastic sentence, therefore, must correspond to a finite segment of the infinite random sequence. A probability measure is assigned to the space of all possible sequences in such a way that (in theory, at least) the probability of any finite segment can be computed.

If the process of manufacturing messages were completely random, the product would bear little resemblance to actual utterances in a natural language. An important feature of a stochastic model for verbal behavior is that successive symbols can be correlated—that the history of the message will support some prediction about its future. In 1948 Shannon revived and elaborated an early suggestion by Markov that the source of messages in a discrete communication system could be represented by a stationary stochastic process that selected successive elements of the message from a finite vocabulary according to fixed probabilities. For example, Markov (1913) classified 20,000 successive letters in Puskhin's *Eugene Onegin* as vowels v or consonants c , then tabulated the frequency N of occurrences of overlapping sequences of length three. His results are summarized in Table 1 in the form of a tree.

There are several constraints on the frequencies that can appear in such a tabulation of binary sequences. For example, $N(vc) = N(cv) \pm 1$, since the sequence cannot shift from vowels to consonants more often, ± 1 , than it returns from consonants to vowels. In this particular example the number of degrees of freedom is 2^{n-1} , where n is the length of the string that is analyzed and 2 is the size of the alphabet.

The tabulated frequencies enable us to estimate probabilities. For instance, the estimated probability of a vowel is $\hat{p}(v) = N(v)/N = 0.432$. If successive letters were independent, we would expect the probability of a vowel following a consonant $p(v | c)$ to be the same as the probability of a

² We assume that the stochastic processes we are studying are stationary.

vowel following another vowel $p(v | v)$, and both would equal $p(v)$. The tabulation, however, yields $\hat{p}(v | c) = 0.663$, which is much larger than $\hat{p}(v)$, and $\hat{p}(v | v) = 0.128$, which is much smaller. Clearly, Russian vowels are more likely to occur after consonants than after vowels. Newman (1951) has reported further data on the written form of several languages and has confirmed this general tendency for vowels and consonants to alternate. (It is unlikely that this result would be seriously affected if the analyses had been made with phonemes rather than with written characters.)

Table 1 Markov's Data on Consonant-Vowel Sequences in Pushkin's *Eugene Onegin*

| | | | | | | | | | |
|-----------------|---|---|----------------|---|---|-----------------|---|---|--------------|
| $N(vvv) = 115$ | } | — | $N(vv) = 1104$ | } | — | $N(v) = 8,638$ | } | — | $N = 20,000$ |
| $N(vvc) = 989$ | | | | | | | | | |
| $N(vcv) = 4212$ | } | — | $N(vc) = 7534$ | } | — | $N(v) = 8,638$ | } | — | $N = 20,000$ |
| $N(vcc) = 3322$ | | | | | | | | | |
| $N(cvv) = 989$ | } | — | $N(cv) = 7534$ | } | — | $N(c) = 11,362$ | } | — | $N = 20,000$ |
| $N(cvc) = 6545$ | | | | | | | | | |
| $N(ccv) = 3322$ | } | — | $N(cc) = 3827$ | } | — | $N(c) = 11,362$ | } | — | $N = 20,000$ |
| $N(ccc) = 505$ | | | | | | | | | |

Inspection of the message statistics in Table 1 reveals that the probability of a vowel depends on more than the one immediately preceding letter. Strictly speaking, therefore, the chain is not Markovian, since a Markov process has been defined in such a way (cf. Feller, 1957) that all of the relevant information about the history of the sequence is given when the single, immediately preceding outcome is known. However, the Markovian representation is readily projected to handle more complicated cases. We shall consider how this can be done.

But first we must clarify what is meant by a Markov source. Given a discrete Markov process with a finite number of states v_0, \dots, v_D and a probability measure μ , a *Markov source* is constructed by defining $V = \{v_0, \dots, v_D\}$ to be the vocabulary; messages are formed by concatenating the names of the successive states through which the system passes. In the terms used in Sec. 1.2 of Chapter 12 a Markov source is a special type of finite state automaton in which the triples that define it are all of the form (i, j, i) and in which the control unit has access to the conditional probabilities of all state transitions.

In Sec. 2 of Chapter 11, a state was defined as the set of all initial strings that were equivalent on the right. This definition must be extended for stochastic systems, however. We say that all the strings that allow the same

continuations with the same probabilities are stochastically equivalent on the right; then a state of a stochastic source is the set of all strings that are stochastically equivalent on the right.

If we are given a long but arbitrary sequence of symbols and wish to test whether it comprises a Markov chain, we must proceed to tabulate the frequencies of the possible pairs, triplets, etc. Our initial (Markovian) hypothesis in this analysis is that the symbol occurring at any given time can be regarded as the name of the state that the source is in at that time. Inspection of the actual sequence, however, may reveal that some of the hypothesized states are stochastically equivalent on the right (all possible continuations are assigned the same probabilities in both cases) and so can be parsimoniously combined into a single state. This reduction in the number of states implies that the state names must be distinguished from the terminal vocabulary. We can easily broaden our definition of a Markov source to include these simplified versions by distinguishing the set of possible states $\{S_0, S_1, \dots, S_m\}$ from the vocabulary $\{v_0, v_1, \dots, v_D\}$.

Since human messages have dependencies extending over long strings of symbols, we know that any pure Markov source must be too simple for our purposes. In order to generalize the Markov concept still further, therefore, we can introduce the following construction (McMillan, 1953): given a Markov source with a vocabulary V , select some different vocabulary W and define a homomorphic mapping of V into W . This mapping will define a new probability measure. The new system is a *projection* of a Markov source, but it may not itself be Markovian in the strict sense.

Definition 1. *Given a Markov source with vocabulary $V = \{v_0, \dots, v_D\}$, with internal states S_0, \dots, S_m , and with probability measure μ , a new source can be constructed with the same states but with vocabulary W and derived probability measure μ' , where $w_j \in W$ if and only if there is a $v_i \in V$, and a mapping θ such that $\theta(v_i) = w_j$. Any source formed from a Markov source by this construction is a projected Markov source.*

The effect of this construction is best displayed by an example. Consider the Markov source whose graph is shown in Fig. 1 and assume that appropriate probabilities are assigned to the indicated transitions, all other conceivable transitions having probability zero. The vocabulary is $V = \{1, 2, 3, 4\}$, and each symbol names the state that the system is in after that symbol occurs. We shall consider three different ways to map V into an alternative vocabulary according to the construction in Definition 1:

1. Let $\theta(1) = \theta(4) = v$ and $\theta(2) = \theta(3) = c$. Then the projected system is a *higher order Markov source* of the type required to represent the probabilities of consonant-vowel triplets in Table 1. Under this construction we would probably identify state 1 as $[v]$, state 2 as $[vc]$, state 3

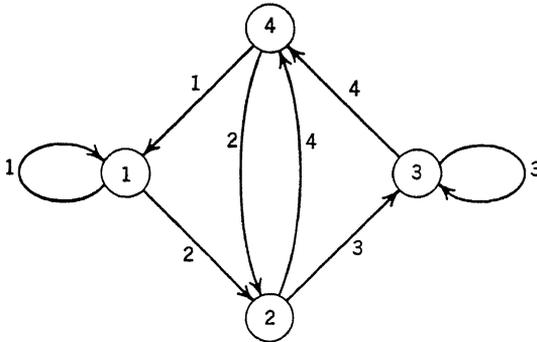


Fig. 1. Graph of a Markov source.

as $[cc]$, and state 4 as $[cv]$, thus retaining the convention of naming states after the sequences that lead into them, but now with the non-Markovian stipulation that more than one preceding symbol is implicated. In the terminology of Chapter 12, we are dealing here with a k -limited automaton, where $k = 2$.

2. Let $\theta(1) = \theta(2) = a$ and $\theta(3) = \theta(4) = b$. Then the projected system is ambiguous: an occurrence of a may leave the system in either state 1 or state 2; an occurrence of b may leave it in either state 3 or state 4. The states cannot be distinctively named after the sequences that lead into them.

3. Let $\theta(1) = -1$, $\theta(2) = \theta(4) = 0$, and $\theta(3) = +1$. With this projection we have a non-Markovian example mentioned by Feller (1957, p. 379). If we are given a sequence of independent random variables that can assume the values ± 1 with probability $\frac{1}{2}$, we can define the moving average of successive pairs, $X_n = (Y_n + Y_{n+1})/2$. The sequence of values of X_n is non-Markovian for an instructive reason; given a consecutive run of $X_n = 0$, how it will end depends on whether it contains an odd or an even number of 0's. After a run of an even number of occurrences of $X_n = 0$ the run must terminate as it began; after an odd number the run must terminate with the opposite symbol from the one with which it started. Thus it is necessary to remember how the system got into each run of 0's and how long the run has been going on. But, since there is no limit to how long a run of 0's may be, this system is not k -limited for any k . Thus it is impossible to produce the moving average by a simple Markov source or even by a higher order (k -limited) Markov process (which still must have finite memory), but it is quite simple to produce it with the projected Markov source constructed here.

By this construction, therefore, we can generalize the notion of a Markov

source to cover any kind of finite state system (regular event) for which a suitable probability measure has been defined.

Theorem 1. *Any finite state automaton over which an appropriate probability measure is defined can serve as a projected Markov source.*

Given any finite state automaton with an associated probability measure, assign a separate integer to each transition. The set of integers so assigned must form the vocabulary of a Markov source, and the rule of assignment defines a homomorphic mapping into a projected Markov source. This formulation makes precise the sense in which regular languages can be said to have Markovian properties.

All of our projected Markov sources will be assumed to operate in real time, from past to future, which we conventionally denote as left to right. Considered as rewriting systems, therefore, they contain only right-branching rules of the general form $A \rightarrow aB$, where A and B correspond to states of the stochastic system. The variety of projected Markov sources is, of course, extremely large, and only a few of the many possible types have been studied in any detail. We shall sample some of them in the following sections.

These same ideas could have been developed equally well to describe a receiver rather than a source. A projected Markov receiver is one that will accept as input only those strings of symbols that correspond to possible sequences of state transitions and that, through explicit agreement with the source or through long experience, has built up for each state an estimate of the probabilities of all possible continuations. As we have already noted, once the mathematical theory is fixed its specialization as a model for either the speaker or the hearer is quite simple. We are really concerned with ways to characterize the user of natural languages; the fact that we have here pictured him as a source is quite arbitrary.

1.2 k -Limited Stochastic Sources

One well-studied type of projected Markov source is known generally as a higher order, or k -limited, Markov source, which generates a $(k + 1)$ -order approximation to the sample of text from which it is derived. The states of the k -limited automaton are identified with the sequences of k successive symbols leading into them, and associated with each state is a probability distribution defined over the D different symbols of the alphabet. If there are D symbols in the alphabet, then a k -limited stochastic source will have (potentially) D^k different states. A 0-limited stochastic source has but one state and generates the symbols independently.

If k is small and if we consider an alphabet of only 27 characters (26

letters and a space), it is possible to estimate the transitional probabilities for a k -limited stochastic source by actually counting the number of $(k + 1)$ -tuplets of each type in a long sample of text. If we use these tabulations, it is then possible to produce $(k + 1)$ -order approximations to the original text by drawing successive characters according to the probability distribution associated with the state determined by the string of k preceding characters. It is convenient to define a zero-order approximation as one that uses the characters independently and equiprobably; a first-order approximation uses the characters independently; a second-order approximation uses the characters with the probabilities appropriate in the context of the immediately preceding letter; etc.

An impression of the kind of approximations to English that these sources produce can be obtained from the following examples, taken from Shannon (1948). In each case the $(k + 1)$ th symbol was selected with probability appropriate to the context provided by the preceding k symbols.

1. Zero-order letter approximation (26 letters and a space, independent and equiprobable): XFOML RXKHRJFFJUJ ZLPWCFWKC YJFFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order letter approximation (characters independent but with frequencies representative of English): OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order letter approximation (successive pairs of characters have frequencies representative of English text): ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order letter approximation (triplets have frequencies representative of English text): IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

A k -limited stochastic source can also be defined for the words in the vocabulary V in a manner completely analogous to that for letters of the alphabet A . When states are defined in terms of the k preceding words, the following kinds of approximations are obtained:

5. First-order word approximation (words independent, but with frequencies representative of English): REPRESENTING AND SPEEDILY IS AN GOOD APT OR CAME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation (word-pairs with frequencies

representative of English): THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The following two illustrations are taken from Miller & Selfridge (1950).

7. Third-order word approximation (word-triplets with frequencies representative of English): FAMILY WAS LARGE DARK ANIMAL CAME ROARING DOWN THE MIDDLE OF MY FRIENDS LOVE BOOKS PASSIONATELY EVERY KISS IS FINE.

8. Fifth-order word approximation (word quintuplets with frequencies representative of English): ROAD IN THE COUNTRY WAS INSANE ESPECIALLY IN DREARY ROOMS WHERE THEY HAVE SOME BOOKS TO BUY FOR STUDYING GREEK.

Higher-order approximations to the statistical structure of English have been used to manipulate the apparent meaningfulness of letter and word sequences as a variable in psychological experiments. As k increases, the sequences of symbols take on a more familiar look and—although they remain nonsensical—the fact seems to be empirically established that they become easier to perceive and to remember correctly.

We know that the sequences produced by k -limited Markov sources cannot converge on the set of grammatical utterances as k increases because there are many grammatical sentences that are never uttered and so could not be represented in any estimation of transitional probabilities. A k -limited Markov source cannot serve as a natural grammar of English no matter how large k may be. Increasing k does not isolate the set of grammatical sentences, for, even though the number of high-probability grammatical sequences included is thereby increased, the number of low-probability grammatical sequences excluded is also increased correspondingly. Moreover, for any finite k there would be ungrammatical sequences longer than k symbols that a stochastic user could not reject.

Even though a k -limited source is not a grammar, it might still be proposed as a model of the user. Granted that the model cannot isolate the set of all grammatical sentences, neither can we; inasmuch as our human limitations often lead us into ungrammatical paths, the real test of this model of the user is whether it exhibits the same limitations that we do.

However, when we examine this model, not as a convenient way to summarize certain statistical parameters of message ensembles, but as a serious proposal for the way people create and interpret their communicative

utterances, it is all too easy to find objections. We shall mention only one, but one that seems particularly serious: the k -limited Markov source has far too many parameters (cf. Miller, Galanter, & Pribram, 1960, pp. 145–148). As we have noted, there can be as many as D^k probabilities to be estimated. By the time k grows large enough to give a reasonable fit to ordinary usage the number of parameters that must be estimated will have exploded; a staggering amount of text would have to be scanned and tabulated in order to make reliable estimates.

Just how large must k and D be in order to give a satisfactory model? Consider a perfectly ordinary sentence: *The people who called and wanted to rent your house when you go away next year are from California.* In this sentence there is a grammatical dependency extending from the second word (the plural subject *people*) to the seventeenth word (the plural verb *are*). In order to reflect this particular dependency, therefore, k must be at least 15 words. We have not attempted to explore how far k can be pushed and still appear to stay within the bounds of common usage, but the limit is surely greater than 15 words; and the vocabulary must have at least 1000 words. Taking these conservative values of k and D , therefore, we have $D^k = 10^{45}$ parameters to cope with, far more than we could estimate even with the fastest digital computers.

Of course, we can argue that many of these 10^{45} strings of 15 words whose probabilities must be estimated are redundant or that most of them have zero probability. A more realistic estimate, therefore, might assume that what we learn are not the admissible strings of words but rather the “sentence frames”—the admissible strings of syntactic categories. Moreover, we might recognize that not all sequences of categories are equally likely to occur; as a conservative estimate (cf. Somers, 1961), we might assume that on the average there would be about four alternative categories that might follow in any given context. By such arguments, therefore, we can reduce D to as little as four, so that D^k becomes $4^{15} = 10^9$. That value is, of course, a notable improvement over 10^{45} parameters, yet, when we recall that several occurrences of each string are required before we can obtain reliable estimates of the probabilities involved, it becomes apparent that we still have not avoided the central difficulty—an enormous amount of text would have to be scanned and tabulated in order to provide a satisfactory empirical basis for a model of this type.

The trouble is not merely that the statistician is inconvenienced by an estimation problem. A learner would face an equally difficult task. If we assume that a k -limited automaton must somehow arise during childhood, the amount of raw induction that would be required is almost inconceivable. We cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds.

1.3 A Measure of Selective Information

Although the direct estimation of all the probabilities involved in a k -limited Markov model of the language user is impractical, other statistics of a more general and summary nature are available to represent certain average characteristics of such a source. Two of these with particular interest for communication research are *amount of information* and *redundancy*. We introduce them briefly and heuristically at this point.

The problem of measuring amounts of information in a communication situation seems to have been posed first by Hartley (1928). If some particular piece of equipment—a switch, say, or a relay—has D possible positions, or physical states, then two of the devices working together can have D^2 states, three can have D^3 states altogether, etc. The number of possible states of the total system increases exponentially as the number of devices increases linearly. In order to have a measure of information that will make the capacity of $2n$ devices just double the capacity of n of them, Hartley defined what we now call the information capacity of a device as $\log D$, where D is the number of different states the total system can get into. Hartley's proposal was later generalized and considerably extended by Shannon (1948) and Wiener (1948).

When applied to a communication channel, Hartley's notion of capacity refers to the number of different signals that might be transmitted in a unit interval of time. For example, let $N(T)$ denote the total number of different strings exactly T symbols long that the channel can transmit. Let D be the number of different states the channel has available and assume that there are no constraints on the possible transitions from one state to another. Then $N(T) = D^T$, or

$$\frac{\log N(T)}{T} = \log D,$$

which is Hartley's measure of capacity. In case there are some constraints on the possible transitions, $N(T)$ will still (in the limit) increase exponentially but less rapidly. In the general case, therefore, we are led to define channel capacity in terms of the limit:

$$\text{channel capacity} = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}. \quad (1)$$

This is the best the channel can do. If a source produces more information per symbol on the average, the channel will not be able to transmit it all—not, at least, in the same number of symbols. The practical problem,

therefore, is to estimate $N(T)$ from what we know about the properties of the channel.

Our present goal, however, is to see how Hartley's original insight has been extended to provide a measure of the amount of information per symbol contained in messages generated by stochastic devices of the sort described in the preceding sections of this chapter. We shall confine our discussion here to those situations in which the purpose of communication is to reduce a receiver's uncertainty. The amount of information he receives, therefore, must be some function of what he learns about the state of the source. And what he learns will depend on how ignorant he was to begin with. Let us assume that the source selects its message by any procedure, random or deterministic, but that all the receiver knows in advance is that the source will choose among a finite set of mutually exclusive messages M_1, M_2, \dots, M_D with probabilities $p(M_1), p(M_2), \dots, p(M_D)$, where these probabilities sum to unity. What Shannon and Wiener did was to develop a measure $H(M)$ of the receiver's uncertainty, where the argument M designates the choice situation:

$$M = \begin{pmatrix} M_1, & M_2, & \dots, & M_D \\ p(M_1), & p(M_2), & \dots, & p(M_D) \end{pmatrix}.$$

When the particular message is correctly received, a listener's uncertainty about it will be reduced from $H(M)$ to zero; therefore, the message conveyed $H(M)$ units of information. Thus $H(M)$ is a measure of the amount of information required (on the average) to select M_i when faced with the choice situation M .

We list as assumptions a number of properties that intuition says a reasonable measure of uncertainty ought to have for discrete devices. Then, following a heuristic presentation by Khinchin (1957), we shall use those assumptions to develop the particular H of Shannon and Wiener.

Our first intuitive proposition is that uncertainty depends only on what might happen. Impossible events will not affect our uncertainty. If a particular message M_i is known in advance to have $p(M_i) = 0$, it should not affect the measure H in any way if M_i is omitted from consideration. Assumption 1. *Adding any number of impossible messages to M does not change $H(M)$:*

$$H \begin{pmatrix} M_1, & \dots, & M_D, & M_{D+1} \\ p(M_1), & \dots, & p(M_D), & 0 \end{pmatrix} = H \begin{pmatrix} M_1, & \dots, & M_D \\ p(M_1), & \dots, & p(M_D) \end{pmatrix}.$$

Our second intuition is that people are most uncertain when the alternative messages are all equally probable. Any bias that makes one message

more probable than another conveys information in the sense that it reduces the receiver's total amount of uncertainty. With only two alternative messages, for example, a 50 : 50 split presents the least predictable situation imaginable. Since there are D different messages in M , when they are equiprobable $p(M_i) = 1/D$ for all i .

Assumption 2. $H(M)$ is a maximum when all the messages in M are equiprobable:

$$H\left(\begin{matrix} M_1, & \dots, & M_D \\ p(M_1), & \dots, & p(M_D) \end{matrix}\right) \leq H\left(\begin{matrix} M_1, & \dots, & M_D \\ 1/D, & \dots, & 1/D \end{matrix}\right).$$

Now let $L(D)$ represent the amount of uncertainty involved when all the messages are equiprobable. Then we have, by virtue of our two assumptions,

$$\begin{aligned} L(D) &= H\left(\begin{matrix} M_1, & \dots, & M_D, & M_{D+1} \\ 1/D, & \dots, & 1/D, & 0 \end{matrix}\right) \\ &\leq H\left(\begin{matrix} M_1, & \dots, & M_{D+1} \\ 1/D + 1, & \dots, & 1/D + 1 \end{matrix}\right) = L(D + 1). \end{aligned}$$

Therefore, we have established the following lemma:

Lemma 1. $L(D)$ is a monotonic increasing function of D .

That is to say, when all D of the alternative messages in M are equiprobable $H(M)$ is a nondecreasing function of D . Intuitively, the more different things that can happen, the more uncertain we are.

It is also reasonable to insist that the uncertainty associated with a choice should not be affected by making the choice in two or more steps, but should be the weighted sum of the uncertainties involved in each step. This critically important assumption can be stated:

Assumption 3. $H(M)$ is additive.

Let any two events of M be combined to form a single, compound event, which we designate as $M_1 \cup M_2$ and which has probability $p(M_1 \cup M_2) = p(M_1) + p(M_2)$. Thus we can decompose M into two parts:

$$M' = \left(\begin{matrix} M_1 \cup M_2, & M_3, & \dots, & M_D \\ p(M_1) + p(M_2), & p(M_3), & \dots, & p(M_D) \end{matrix} \right),$$

and

$$M'' = \left(\begin{matrix} M_1, & M_2, & M_3, & \dots, & M_D \\ \frac{p(M_1)}{p(M_1) + p(M_2)}, & \frac{p(M_2)}{p(M_1) + p(M_2)}, & 0, & \dots, & 0 \end{matrix} \right).$$

A choice from M is equivalent to a choice from M' followed (if $M_1 \cup M_2$ is chosen) by a choice from M'' . Assumption 3 means that $H(M)$ depends

on the sum of $H(M')$ and $H(M'')$. In calculating $H(M)$, however, $H(M'')$ should be weighted by $p(M_1) + p(M_2)$ because that represents the probability that a second choice will be required. Assumption 3 implies that

$$H(M) = H(M') + [p(M_1) + p(M_2)]H(M'').$$

If this equation holds whenever two messages of M are lumped together, then it can easily be generalized to any subset whatsoever, and it can be extended to more than one subset of messages in M .

In order to discuss this more general situation, we represent the messages in M by M_{ij} , where i is the first selection and j is the second. The first selection is made from A :

$$A = \begin{pmatrix} A_1, & \dots, & A_r \\ p(A_1), & \dots, & p(A_r) \end{pmatrix},$$

where

$$p(A_i) = \sum_j p(M_{ij}),$$

and the second choice depends (as before) on the outcome of the first; that is to say, the second choice is made from

$$B | A_i = \begin{pmatrix} B_1, & \dots, & B_s \\ p(B_1 | A_i), & \dots, & p(B_s | A_i) \end{pmatrix}.$$

The B_j have probabilities $p(B_j | A_i)$ that depend on A_i , the preceding choice from A . The two choices together are equivalent to—are a decomposition of—a single choice from M , where

$$A_i B_j = M_{ij},$$

and

$$p(A_i) p(B_j | A_i) = p(M_{ij}).$$

Now, by Assumption 3, $H(M)$ should be the sum of the two components. But that is a bit complicated, since $H(B | A_i)$ is a random variable depending on i . On the average, however, it will be

$$E\{H(B | A_i)\} = \sum_i p(A_i) H(B | A_i) = H(B | A). \quad (2)$$

In this situation, therefore, the assumption of additivity means that

$$H(M) = H(AB) = H(A) + H(B | A). \quad (3)$$

Of course, if A and B are independent, Eq. 3 becomes

$$H(AB) = H(A) + H(B), \quad (4)$$

and, if the messages are independent and equally probable, a sequence of s successive choices among D alternatives will give

$$L(D^s) = sL(D). \quad (5)$$

We shall now establish the following lemma:

Lemma 2. $L(D) = k \log D$, where $k > 0$.

Consider repeated independent choices from the same number D of equiprobable messages. Select m such that for any positive integers D, s, t

$$D^m \leq s^t \leq D^{m+1} \quad (6)$$

$$m \log D \leq t \log s \leq (m+1) \log D$$

$$\frac{m}{t} \leq \frac{\log s}{\log D} \leq \frac{m+1}{t} \quad (7)$$

From Eq. 6, and the fact that $L(D)$ is monotonic increasing, it follows that

$$L(D^m) \leq L(s^t) \leq L(D^{m+1}),$$

and from Eq. 5 we know that

$$\begin{aligned} m L(D) &\leq t L(s) \leq (m+1) L(D) \\ \frac{m}{t} &\leq \frac{L(s)}{L(D)} \leq \frac{m+1}{t}. \end{aligned} \quad (8)$$

Combining Eqs. 7 and 8, therefore,

$$\left| \frac{L(s)}{L(D)} - \frac{\log s}{\log D} \right| \leq \frac{1}{t}.$$

Since m is not involved, t can be chosen arbitrarily large, and

$$\frac{L(s)}{\log s} = \frac{L(D)}{\log D}.$$

Moreover, since D and s are quite arbitrary, these ratios must be constant independent of D ; that is to say,

$$\frac{L(D)}{\log D} = k, \quad \text{so} \quad L(D) = k \log D.$$

Of course, $\log D$ is nonnegative and therefore [since $L(D)$ is monotonic increasing] $k > 0$. This completes the proof of Lemma 2.

Ordinarily k is chosen to be unity when logarithms are taken to the base 2,

$$L(D) = \log_2 D, \quad (9)$$

that is to say, the unit of measurement is taken to be the amount of uncertainty involved in a choice between two equally possible alternatives. This unit is called a *bit*.

Next consider the general case with unequal, but rational, probabilities. Let

$$p(A_i) = \frac{g_i}{g} \quad (i = 1, \dots, r),$$

where the g_i are all positive integers and

$$\sum_i g_i = g.$$

The problem is to determine $H(A)$. In order to do this, we shall construct a second choice situation ($B | A_i$) in a special way so that the Cartesian product $M = A \times B$ will consist entirely of equiprobable alternatives.

Let ($B | A_i$) consist of g_i messages each with probability $1/g_i$. Therefore,

$$H(B | A_i) = H\left(\begin{array}{c} B_1, \dots, B_{g_i} \\ 1/g_i, \dots, 1/g_i \end{array}\right) = L(g_i) = c \log g_i. \quad (10)$$

From Eqs. 2 and 10 it follows that

$$\begin{aligned} H(B | A) &= \sum_i p(A_i) H(B | A_i) \\ &= \sum_i p(A_i) c \log g_i \\ &= c \sum_i p(A_i) \log p(A_i) g \\ &= c \log g + c \sum_i p(A_i) \log p(A_i). \end{aligned} \quad (11)$$

Consider next the compound choice $M = A \times B$. Since

$$p(A_i B_j) = p(A_i) p(B_j | A_i) = \frac{g_i}{g} \cdot \frac{1}{g_i} = \frac{1}{g},$$

it must follow that for this specially contrived situation there are g equally probable events and

$$H(A \times B) = H(AB) = L(g) = c \log g. \quad (12)$$

When we substitute Eqs. 11 and 12 into Eq. 3 we obtain

$$c \log g = H(A) + c \log g + c \sum_i p(A_i) \log p(A_i).$$

We have now established the theorem:

Theorem 2. For rational probabilities,

$$H(A) = -c \sum_i p(A_i) \log p(A_i). \quad (13)$$

Since Eq. 13 can be interpreted as the mean value $E\{-\log p(A_i)\}$, the measure of uncertainty thus turns out to be the mean logarithmic probability—a quantity familiar to physicists under the name entropy. It is as

though we had defined the amount of information in message A_i to be $-\log p(A_i)$, regardless of what the probability distribution might be for the other messages. The assumption that the amount of information conveyed by one particular message is independent of all the other possible messages is what Luce (1960) has called the assumption of independence from irrelevant alternatives; he remarks (Luce, 1959) that it is characteristic—either explicitly or implicitly—of most theories of choice behavior.

Finally, in order to make $H(B)$ a continuous function of the probabilities, we need a fourth assumption of continuity. Since it is felt intuitively that a small change in probabilities should result in a small change in $H(M)$, this final assumption needs little comment here. It will not play a critical role in the discussion that follows.

Next, we want to use H to measure the uncertainty associated with the projected Markov sources. Suppose we have a stationary source with a finite number of states A_1, \dots, A_n , with an alphabet B_1, \dots, B_D , and with the matrix of transitional probabilities $p(B_j | A_i)$. When the system is in state A_i , the choice situation is

$$B | A_i = \begin{pmatrix} B_1, & B_2, & \dots, & B_D \\ p(B_1 | A_i), & p(B_2 | A_i), & \dots, & p(B_D | A_i) \end{pmatrix}.$$

By Theorem 2 the amount of information involved in this choice must be

$$H(B | A_i) = -c \sum_j p(B_j | A_i) \log p(B_j | A_i).$$

This quantity is defined for each state. In order to obtain an average value to represent the amount of information that we can expect for the source, regardless of the state it is in, we must average over i :

$$\begin{aligned} E\{H(B | A_i)\} &= \sum_i p(A_i) H(B | A_i) \\ &= -c \sum_i \sum_j p(A_i, B_j) \log p(B_j | A_i) = H(B | A). \end{aligned} \quad (14)$$

Now we can regard $H(B | A)$ as a measure of the average amount of information obtained when the source moves one step ahead by choosing a letter from the set $\{B_i\}$. [In the special case in which successive events in the chain are independent, of course, $H(B | A)$ reduces to $H(B)$.] A string of N successive choices, therefore, will yield $NH(B | A)$ units of information on the average.

In general, $H(AB) \leq H(A) + H(B)$; equality obtains only when A and B are independent. This fact can be demonstrated as follows: the familiar expansion

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \quad (x \geq -1),$$

can be used to establish that $e^x \geq 1 + x$. If we set $t = 1 + x$, this inequality can be written as

$$t - 1 \geq \log_e t, \quad (t \geq 0).$$

Now put $t = p(A_i)p(B_j)/p(A_iB_j)$:

$$\frac{p(A_i)p(B_j)}{p(A_iB_j)} - 1 \geq \log_e p(A_i) + \log_e p(B_j) - \log_e p(A_iB_j),$$

and take expected values over the distribution $p(A_iB_j)$:

$$\begin{aligned} \sum_j \sum_i p(A_iB_j) \frac{p(A_i)p(B_j)}{p(A_iB_j)} - 1 &\geq \sum_j \sum_i p(A_iB_j) \log_e p(A_i) \\ &\quad + \sum_j \sum_i p(A_iB_j) \log_e p(B_j) \\ &\quad - \sum_j \sum_i p(A_iB_j) \log_e p(A_iB_j), \end{aligned}$$

so

$$1 - 1 \geq -H(A) - H(B) + H(AB),$$

which is the result we wished to establish:

$$H(A) + H(B) \geq H(AB). \quad (15)$$

If we compare Eq. 15 with the assumption of additivity expressed in Eq. 3, we see that we have also established the following theorem:

Theorem 3. $H(B) \geq H(B | A).$ (16)

This important inequality can be interpreted to mean that knowledge of the choice from A cannot increase our average uncertainty about the choice from B . In particular, if A represents the past history of some message and B represents the choice of the next message unit, then the average amount of information conveyed by B can never increase when we know the context in which it is selected.

It is important to remember that H is an average measure of *selective* information, based on the assumption that the improbable event is always the most informative, and is not a simple measure of semantic information (cf. Carnap & Bar-Hillel, 1952). An illustration may suggest the kind of problems that can arise: in ordinary usage *It is a man* will generally be judged to convey more information than *It is a vertebrate*, because the fact that something is a man implies that it is a vertebrate, but not vice versa. In the framework of selective information theory, however, the situation is reversed. According to the tabulations of the frequencies of English words, *vertebrate* is a less probable word than *man*, and its selection in English discourse must therefore be considered to convey more information.

Because many psychological processes involve selective processes of one kind or another, a measure of selective information has proved to be of some value as a way to characterize this aspect of behavior. Surveys of various applications of information measures to psychology have been prepared by Attneave (1959), Cherry (1957), Garner (1962), Luce (1960), Miller (1953), Quastler (1955), and others. Not all applications of the mean logarithmic probability have been carefully considered and well motivated, however. As Cronbach (1955) has emphasized, in many situations it may be advisable to develop alternative measures of information based on intuitive postulates that are more closely related to the particular applications we intend to make.

1.4 Redundancy

Since $H(B) \geq H(B|A)$, where equality holds only for sequentially independent messages, any sequential dependencies that the source introduces will act to reduce the amount of selective information the message contains. The extent to which the information is reduced is a general and interesting property of the source. Shannon has termed it the *redundancy* and has defined it in the following way.

First, consider the amount of information that could be encoded in the given alphabet (or vocabulary) if every atomic symbol were used independently and equiprobably. If there are D atomic symbols, then the informational capacity of the alphabet will be $L(D) = \log_2 D$ bits per symbol. Moreover, this value will be the maximum possible with that alphabet. Now, if we determine that the source is producing an amount $H(M)$ that is actually less than its theoretical maximum per symbol, $H(M)/L(D)$ will be some fraction less than unity that will represent the *relative* amount of information from the source. One minus the relative information is the redundancy:

$$\text{per cent redundancy} = 100 \left(1 - \frac{H(M)}{\log D} \right). \quad (17)$$

The relative amount of information per symbol is a measure of how efficiently the coding alphabet is being used. For example, if the relative information per symbol is only half what it might be, then on the average the messages are twice as long as necessary. Shannon (1948), on the basis of his observation that a highly skilled subject could reconstruct passages from which 50% of the letters had been removed, estimated the efficiency of normal English prose as something less than 50%. But, when Chapanis (1954) tried to repeat this observation with other subjects and other passages, he found that if letters are randomly deleted and the text is shortened

so that no indication is given of the location of the deletion few people are able to restore more than 25% of the missing letters in a short period of time. However, these are difficult conditions to impose on subjects. In order to estimate the coding efficiency of English writing, we should first make every effort to optimize the conditions for the person who is trying to reconstruct the text. For example, we might tell him in advance that all spaces between words and all vowels have been deleted. This form of abbreviation shortens the text by almost 50%, yet Miller and Friedman (1957) found that the most highly skilled subjects were able to restore the missing characters if they were given sufficient time and incentive to work at the task. We can conclude, therefore, that English is at least 50% redundant and perhaps more.

Why do we bother with such crude bounds? Why not compute redundancy directly from the message statistics for printed English? As we noted at the end of Sec. 1.2, the direct approach is quite impractical, for there are too many parameters to be estimated. However, we can put certain rough bounds on the value of H by limiting operations that use the available message statistics directly for short sequences of letters in English (Shannon, 1948). Let $p(x_i)$ denote the probability of a string x_i of k symbols from the source and define

$$G_k = -\frac{1}{k} \sum_i p(x_i) \log_2 p(x_i), \quad (18)$$

where the sum is taken over all strings x_i containing exactly k symbols. Then G_k will be a monotonic decreasing function of k and will approach H in the limit.

An even better estimate can be obtained with conditional probabilities. Consider a matrix P whose rows represent the D^k possible strings x_i of k symbols and whose columns represent the D different symbols a_j . The elements of the matrix are $p(a_j | x_i)$, the conditional probabilities that a_j will occur as the $(k + 1)$ st symbol given that the string x_i of k symbols just preceded it. For each row of this matrix

$$-\sum_j p(a_j | x_i) \log_2 p(a_j | x_i)$$

measures our uncertainty regarding what will follow the particular string x_i . The expected value of this uncertainty defines a new function,

$$F_{k+1} = -\sum_i \sum_j p(x_i) p(a_j | x_i) \log_2 p(a_j | x_i), \quad (19)$$

where $p(x_i)$ is the probability of string x_i . Since $p(x_i) p(a_j | x_i) = p(x_i a_j)$, we can show that

$$\begin{aligned} F_{k+1} &= (k + 1)G_{k+1} - kG_k \\ &= (k + 1)(G_{k+1} - G_k) + G_k. \end{aligned}$$

Therefore, as G_k approaches H , F_k must also approach H . Moreover,

$$G_{k+1} - F_{k+1} = \frac{k}{k+1} G_k \geq 0,$$

so we know that

$$G_k \geq F_k.$$

Thus F_k converges on H more rapidly than G_k as k increases.

Even F (and similar functions using the message statistics) converges quite slowly for natural languages, so Shannon (1951) proposed an estimation procedure using data obtained with a guessing procedure. We consider here only his procedure for determining an upper bound for H (and thus a lower bound for the redundancy).

Imagine that we have two identical k -limited automata that incorporate the true probabilities of English strings. Given a finite string of k symbols, these devices assign the correct probabilities for the $(k+1)$ st symbol. The first device is located at the source. As each symbol of the message is produced, the device guesses what the next symbol will be. It guesses first the most probable symbol, second the next most probable, and so on, continuing in this way until it guesses correctly. Instead of transmitting the symbol produced by the source, we transmit the number of guesses that the device required.

The second device is located at the receiver. When the number j is received, this second device interprets it to mean that the j th guess (given the preceding context) is correct. The two devices are identical and the order of their guesses in any context will be identical; the second machine decodes the received signal and recovers the original message. In that way the original message can be perfectly recovered, so the sequence of numbers must contain the same information—therefore no less an *amount* of information—as the original text. If we can determine the amount of information per symbol for the reduced text, we shall also have determined an upper bound for the original text.

What will the reduced text look like? We do not possess two such k -limited automata, but we can try to use native speakers of the language as a substitute. Native speakers do not know all the probabilities we need, but they do know the syntactic and semantic rules which lead to those probabilities. We can let a person know all of the text up to a given point, then on the basis of that and his knowledge of the language ask him to guess the next letter. Shannon (1951) gives the following as typical of the results obtained:

T H E R E # I S # N O # R E V E R S E # O N # A # . . .

1 1 1 5 1 1 2 1 2 1 1 1 5 1 1 7 1 1 1 2 1 3 2 1 2 2 . . .

The top line is the original message; below it is the number of guesses required for each successive letter.

Note that most letters are guessed correctly on the first trial—approximately 80% when a large amount of antecedent context is provided. Note also that in the reduced text the sequential constraints are far less important; how many guesses the n th letter took tells little about how many will be needed for the $(n + 1)$ st. It is as if the sequential redundancy of the original text were transformed into a nonsequential favoritism for small numbers in the reduced text. Thus we are led to consider the quantity

$$E_{k+1} = - \sum_{j=1}^{27} q_k(j) \log_2 q_k(j), \quad (20)$$

where $q_k(j)$ is the probability of guessing the $(k + 1)$ st letter of a string correctly on exactly the j th guess. If k is large, and if our human subject is a satisfactory substitute for the k -limited automaton we lack, then E_k should be fairly close to H .

Can we make this idea more precise? Suppose we reconsider the $D^k \times D$ matrix P whose elements $p(a_j | x_i)$ are the conditional probabilities of symbol a_j , given the string x_i . What the k -limited automaton will do when it guesses is to map a_j into the digit $\theta(a_j)$ for each row, where the character with the largest probability in the row would be coded as 1, the next largest as 2, and so on. Consider, therefore, a new $D^k \times D$ matrix Q whose rows are the same but whose columns represent the first D digits in order. Then in every row of this new matrix the conditional probabilities $q[\theta(a_j) | x_i]$ would be arranged in a monotonically decreasing order of magnitude from left to right. Note that we have lost nothing in shifting from P to Q ; θ has an inverse, so F_k can be computed from Q just as well as from P .

Now suppose we ignore the context x_i ; that is to say, suppose we simply average all the rows of Q together, weighting them according to their probability of occurrence. This procedure will yield $q_k(j)$, the average probability of being correct on the j th guess. From Theorem 3 we know that $F_k \leq E_k$. Therefore, E_k must also be an upper bound on the amount of information per symbol.

Moreover, this bound holds even when we use a human substitute for our hypothetical automaton, since people can err only in the direction of greater uncertainty (greater E_k) than would an ideal device. We can formulate this fact rigorously: suppose the true probabilities of the predicted symbols are p_i but that our subject is guessing on the basis of some (not necessarily accurate; cf. Toda, 1956) estimates \hat{p}_i , derived somehow from his knowledge of the language and his previous experience with the source. Let $\sum p_i = \sum \hat{p}_i = 1$, and consider the mean value of the

quantity $a_i = \hat{p}_i/p_i$. From the well-known theorem of the arithmetic and geometric means (see, e.g., Hardy, Littlewood, & Polya, 1952, Chapter 2), we know that

$$(a_1)^{p_1} \dots (a_D)^{p_D} \leq p_1 a_1 + \dots + p_D a_D,$$

from which we obtain directly

$$\left(\frac{\hat{p}_1}{p_1}\right)^{p_1} \dots \left(\frac{\hat{p}_D}{p_D}\right)^{p_D} \leq 1,$$

with equality only if $\hat{p}_i = p_i$ for all i . Taking logarithms,

$$\sum_{i=1}^D p_i \log \frac{\hat{p}_i}{p_i} \leq 0,$$

which gives the desired inequality

$$-\sum p_i \log \hat{p}_i \geq -\sum p_i \log p_i. \quad (21)$$

Any inaccuracy in the subject's estimated probabilities can serve only to increase the estimate of the amount of information. The more ignorant he is, the more often the source will surprise him.

The guessing technique for estimating bounds on the amount of selective information contained in redundant strings of symbols can be performed rapidly, and the bounds are often surprisingly low. The technique has been useful even in nonlinguistic situations.

Shannon's (1951) data for a single, highly skilled subject gave $E_{100} = 1.3$ bits per character. For a 27-character alphabet the maximum possible would be $\log_2 27 = 4.73$ bits per character. The lower bound for the redundancy, therefore, is $1 - (1.3/4.73) = 0.73$. This can be interpreted to mean that, for the type of prose passages Shannon used, at least 73 of every 100 characters on the page could have been deleted if the same alphabet had been used most efficiently, that is, if all the characters were used independently and equiprobably. Burton and Licklider (1955) confirmed this result and added that E_k has effectively reached its asymptote by $k = 32$; that is to say, measurable effects of context on a person's guesses do not seem to extend more than 32 characters (about six words) back into the history of the message.

The lower bound on redundancy depends on the particular passage used. In some situations—air-traffic-control messages to a pilot landing at a familiar airport—redundancy may rise as high as 96% (Frick & Sumbly, 1952; Fritz & Grier, 1955).

1.5 Some Connections with Grammaticalness

In Sec. 3 of Chapter 11 we mentioned the difficult problem of assigning degrees of grammaticalness to strings in a way that would reflect the

manner and extent of their deviation from well-formedness in a given language. Some of the concepts introduced in the present chapter suggest a possible approach to this problem.³

Suppose we have a grammar G that generates a fairly narrow (though, of course, infinite) set $L(G)$ of well-formed sentences. How could we assign to each string not generated by the grammar a measure of its deviation in at least one of the many dimensions in which deviation can occur? We might proceed in the following way: select some unit—for concreteness, let us choose word units and, for convenience, let us not bother to distinguish in general between different inflectional forms (e.g., between *find*, *found*, *finds*). Next, set up a hierarchy \mathcal{C} of classes of these units, where $\mathcal{C} = \mathcal{C}_1, \dots, \mathcal{C}_N$, and for each $i \leq N$

$$\begin{aligned} \mathcal{C}_i = \{C_1^i, \dots, C_{a_i}^i\}, \text{ where: } & a_1 > a_2 > \dots > a_N = 1, \\ & C_j^i \text{ is nonnull,} \\ & \text{for each word } w, \text{ there is a } j \text{ such that} \\ & \quad w \in C_j^i, \\ & \text{and } C_j^i \subseteq C_k^i \text{ if and only if } j = k. \quad (22) \end{aligned}$$

\mathcal{C}_1 is the most highly differentiated class of categories; \mathcal{C}_N contains but a single category. Other conditions might be imposed (e.g., that \mathcal{C}_i be a refinement of \mathcal{C}_{i+j}), but Condition 22 suffices for the present discussion.

\mathcal{C}_i is called the *categorization of order i* ; its members are called *categories of order i* . A sequence $C_{b_1}^i, \dots, C_{b_q}^i$ of categories of order i is called a *sentence-form of order i* ; it is said to generate the string of words $w_1 \dots w_q$ if, for each $j \leq q$, $w_j \in C_{b_j}^i$. Thus the set of all word strings generated by a sentence-form is the complex (set) product of the sequence of categories.

We have described \mathcal{C} and G independently; let us now relate them. We say that a set Σ of sentence-forms of order i *covers G* if each string of $L(G)$ is generated by some member of Σ . We say that a sentence-form is *grammatical* with respect to G if one of the strings that the sentence-form generates is in $L(G)$ —*fully grammatical*, with respect to G , if each of the strings that it generates is in $L(G)$. We say that \mathcal{C} is *compatible* with G if for each sentence w of $L(G)$ there is a sentence-form of order one that generates w and that is fully grammatical with respect to G . Thus, if \mathcal{C} is compatible with G , there is a set of fully grammatical sentence-forms of order one that covers G . We might also require, for compatibility, that \mathcal{C}_1 be the smallest set of word classes to meet this condition. Note in this case that the categories of \mathcal{C}_1 need not be pairwise disjoint. For example,

³ The idea of using information measures to determine an optimal set of syntactic categories, as outlined here, was suggested by Peter Elias. This approach is developed in more detail, with some supporting empirical evidence, in Chomsky (1955, Chapter 4).

know will be in C_i^1 and *no* in C_j^1 , where $i \neq j$, although they are phonetically the same. If two words are mutually substitutable throughout $L(G)$, they will be in the same category C_j^1 , if it is compatible with G , but the converse is not necessarily true.

We say that a string w is *i*-grammatical (has degree of grammaticalness i) with respect to G , \mathcal{C} if i is the least number such that w is generated by a grammatical sentence-form of order i . Thus the strings of the highest degree of grammaticalness are those of order 1, the order with the largest number of categories. All strings are grammatical of order N or less, since \mathcal{C}_N contains only one category.

These ideas can be clarified by an example. Suppose that G is a grammar of English and that \mathcal{C} is a system of categories compatible with it and having a structure something like this:

$$\begin{aligned}
 \mathcal{C}_1: N_{\text{hum}} &= \{\text{boy, man, } \dots\} \\
 N_{\text{ab}} &= \{\text{virtue, sincerity, } \dots\} \\
 N_{\text{comp}} &= \{\text{idea, belief, } \dots\} \\
 N_{\text{mass}} &= \{\text{bread, beef, } \dots\} \\
 N_{\text{comm}} &= \{\text{book, chair, } \dots\} \\
 V_1 &= \{\text{admire, dislike, } \dots\} \\
 V_2 &= \{\text{annoy, frighten, } \dots\} \\
 V_3 &= \{\text{hit, find, } \dots\} \\
 V_4 &= \{\text{sleep, reminisce, } \dots\} \\
 &\text{etc.}
 \end{aligned} \tag{23}$$

$$\begin{aligned}
 \mathcal{C}_2: \text{Noun} &= N_{\text{hum}} \cup N_{\text{ab}} \cup \dots \\
 \text{Verb} &= V_1 \cup V_2 \cup \dots \\
 &\text{etc.}
 \end{aligned}$$

$$\mathcal{C}_3: \text{Word.}$$

This extremely primitive hierarchy \mathcal{C} of categories would enable us to express some of the grammatical diversity of possible strings of words. Let us assume that G would generate *the boy cut the beef, the boy reminisced, sincerity frightens me, the boy admires sincerity, the idea that sincerity might frighten you astonishes me, the boy found a piece of bread, the boy found the chair, the boy who annoyed me slept here*, etc. It would not, however, generate such strings as *the beef cut sincerity, sincerity reminisced, the boy frightens sincerity, sincerity admires the boy, the sincerity that the idea might frighten you astonishes me, the boy found a piece of book, the boy annoyed the chair, the chair who annoyed me found here*, etc. Strings of

the first type would be one-grammatical (as are all strings generated by G); strings of the second type would be two-grammatical; all strings would be three-grammatical, with respect to this primitive categorization.

Many of the two-grammatical strings might find a natural use in actual communication, of course. Some of them, in fact, (e.g., *misery loves company*, etc.) might be more common than many one-grammatical strings (an infinite number of which have zero probability and consist of parts which have zero probability, effectively).

A speaker of English can impose an interpretation on many of these strings by considering their analogies and resemblances to those generated by the grammar he has mastered, much as he can impose an interpretation on an abstract drawing. One-grammatical strings, in general, like representational drawings, need have no interpretation *imposed* on them to be understood. With a hierarchy such as \mathcal{C} we could account for the fact that speakers of English know for example, that *colorless green ideas sleep furiously* is surely to be distinguished, with respect to well-formedness, from *revolutionary new ideas appear infrequently* on the one hand and from *furiously sleep ideas green colorless* or *harmless seem dogs young friendly* (which has the same pattern of grammatical affixes) on the other; and so on, in an indefinite number of similar cases.

Such considerations show that a generative grammar could more completely fulfil its function as an explanatory theory if we had some way to project, from the grammar, a certain compatible hierarchy \mathcal{C} in terms of which degree of grammaticalness could be defined. Let us consider now how this might be done.

In order to simplify exposition, we first restrict the problem in two ways. We shall consider only sentences of some fixed length, say length λ . Second, let us consider the problem of determining the system of categories $\mathcal{C}_i = \{C_1^i, \dots, C_{a_i}^i\}$, where a_i is fixed. The best choice of a_i categories is the one that in the appropriate sense maximizes substitutability relations among the categorized elements. The question, then, is how we can select the fixed number of categories which best mirror substitutability relations. Note that we are interested in substitutability not with respect to $L(G)$ but to contexts stated in terms of the categories of \mathcal{C}_i itself. To take an example, *boy* and *sincerity* are much more freely substitutable in contexts defined by the categories of \mathcal{C}_2 of (23) than in actual contexts of $L(G)$; thus we may find both words in the context Noun Verb Determiner—, but not in the context *you frightened the* —. Some words may not be substitutable at all in $L(G)$, although they are mutually substitutable in terms of higher order categories. This fact suggests that systematic procedures of substitution applied to successive words in some sequence of grammatical sentences will probably always fail—as, indeed, they always

have so far—since the maximization of substitutability, in the sense we intend here, is a property of the whole system of categories.

A better way to approach the problem is this: suppose that σ_1 is a sequence s_1, \dots, s_m of all sentences of length λ in $L(G)$ and that \mathcal{C}_i is a proposed set of a_i categories. Let σ_2 be a sequence of sentence-forms $\Sigma_1, \dots, \Sigma_m$, where, for each $j \leq m$, Σ_j generates s_j and Σ_j consists of categories of \mathcal{C}_i . There will, of course, be many repetitions in σ_2 , in general. Let σ_3 be the sequence t_1, \dots, t_n of all strings generated by the Σ_j 's in σ_2 , where σ_3 contains no repetitions. For example, if σ_1 contains *the boy slept* and *the period elapsed*, but not *the period slept* or *the boy elapsed*, and if σ_2 is based on a categorization into nouns and verbs [i.e., σ_2 contains (Determiner, Noun, Verb)], then σ_3 would contain all four of those sentences.

It seems reasonable to measure the adequacy of a system of categories by some function of the length of the sequence σ_3 . The number of generated sentences in σ_3 indicates the extent to which the categorization reflects substitutability relations not only with respect to the given set of sentences but also with respect to contexts defined in terms of the categories themselves. Thus particular nouns may not be substitutable with respect to the same verbs, but they do each occur in a given position relative to some verb so that they are substitutable with respect to the category Verb. The same is true of particular verbs, adjectives, etc. This approach permits us to set up all the categories simultaneously.

To evaluate a system \mathcal{C}_i of a_i categories, given a sequence σ_1 of actual sentences of length λ , we shall try to discover a sequence σ_2 that covers σ_1 , in the sense we have defined (more precisely, whose terms constitute a set that covers σ_1), and that is *minimal* in the sense that it generates the shortest sequence σ_3 . In case the categories of \mathcal{C}_i are pairwise disjoint, this procedure is perfectly simple; we merely replace each word in the strings of σ_1 by the category of \mathcal{C}_i to which it belongs, thus forming σ_2 . But, if the categories of \mathcal{C}_i overlap, there may be many covering sequences σ_2 ; we must find the minimal one in order to evaluate \mathcal{C}_i .

Categories overlap in the case of grammatical homonyms, as we have observed. Note that if a word is put into more than one category when we form \mathcal{C}_i the value of this categorization will always suffer a loss in one respect. Each time a category appears in a sentence-form of σ_2 a set of sentences of σ_3 is generated for each word in that category. Hence the more words in a category, the more sentences generated and the less satisfactory the categorization. However, if the word assigned to two categories is a bona fide homonym, there may be a compensating saving. Suppose, for example, that the phoneme sequence $/n\bar{o}/$ (*know*, *no*) is put only into the category of verbs. Then all verbs will be generated in σ_3 in

the position *there are* — *books on the table*. Similarly, if it is put only into the category of determiners, all will be generated in such contexts as *I* — *that he has been here*. If $/n\bar{o}/$ is assigned to both categories, a given occurrence of $/n\bar{o}/$ in σ_1 can be assigned to either verb or determiner. Since verbs will appear anyway in the context *I* — *that he has been here* and determiners in *there are* — *books on the table*, no new sentence forms are produced by assignment of $/n\bar{o}/$ to verb in the first case and to determiner in the second. There is thus a considerable saving in the sequence σ_3 of generated strings.

These observations suggest a way to decide when an element should in fact be considered a set of grammatical homonyms. We make this decision when the loss incurred automatically in assigning it to several categories is more than compensated for by the gain that can be achieved through the extra freedom in choosing the complete covering sequence σ_2 ; there is always a numerical answer to this question. It must be shown, of course, that in terms of presystematic criteria, the solution of the homonym problem given by this approach is the correct one. Certain preliminary investigations of this have been hopeful (cf. Chomsky, 1955), but the task of evaluating and improving this or any other conception of syntactic category is an immense one. Furthermore, several important distinctions have been blurred in this brief discussion.

Let us now return to our two assumptions: (a) that the length λ of sentences is fixed and (b) that the number a_i of categories is fixed. The first is easily dispensable. Given G , we can evaluate a set $\mathcal{C}_i = \{C_1^i, \dots, C_{a_i}^i\}$, where a_i is a fixed integer, in the following way. Select a new "word" $\#$ to indicate sentence boundary, $\# \notin C_j^i$ for any j . Define a *discourse* as a sequence of words $\#, w_1^1, \dots, w_{\alpha_1}^1, \#, w_1^2, \dots, w_{\alpha_2}^2, \#, \dots, \#, w_1^k, \dots, w_{\alpha_k}^k$, where for each j , $w_1^j \dots w_{\alpha_j}^j$ is a sentence of the language generated by G . This is a discourse of length $\alpha_1 + \dots + \alpha_k + k$. An *initial discourse* is an initial subsequence of a discourse. A *discourse form* is a sequence of categories $C_{\beta_1}^i, \dots, C_{\beta_q}^i$ of categories of \mathcal{C}_i or $\{\#\}$ such that there is a discourse w_1, \dots, w_q , where, for each j , $w_j \in C_{\beta_j}^i$, and an *initial discourse form* is an initial subsequence of a discourse form. Let Σ_λ be a set of initial discourse forms, each of length λ , which covers the set of initial discourses of length λ and is minimal from the point of view of generation, and let $N(\lambda)$ be the number of distinct strings generated by the members of Σ_λ . Then the natural way to define the value of the categorization \mathcal{C}_i is, by analogy with the definition of channel capacity in Eq. 1, p. 431, as

$$\text{Val}(\mathcal{C}_i) = \lim_{\lambda \rightarrow \infty} \frac{\log N(\lambda)}{\lambda}. \quad (24)$$

We choose as the best categorization into a_i categories that analysis \mathcal{C}_i for which $\text{Val}(\mathcal{C}_i)$ is minimal. In other words, we select the categorization into a_i categories that minimizes the information per word, that is, maximizes the redundancy, in the generated “language” of grammatical discourses (assuming independence of successive sentences). Thus we shall try to select a categorization that maximizes the contribution of the category analysis to the total set of constraints under which the source operates in producing discourses. In practice, this computation can be much simplified by assuming that successive choices of Σ_λ , for increasing λ , are not independent.

We have now proposed a definition of optimal categorization into n categories, for each n , which is independent of arbitrary decisions about sentence length. We must finally consider the assumption that we are given the integers a_1, \dots, a_N which determine the number of categories in Condition 22. Suppose, in fact, that we determine for each n the optimal categorization K_n into n categories, in the way previously sketched. To select from the set $\{K_n\}$ the hierarchy \mathcal{C} , we must determine for which integers a_i we will actually set up the optimal categorization K_{a_i} as an order \mathcal{C}_i of \mathcal{C} . We would like to select a_i in such a way that K_{a_i} will be clearly preferred to K_{a_i-1} but will not be much worse than K_{a_i+1} ; that is to say, we would like to select K_{a_i} in such a way that there will be a considerable loss in forming a system of categories with fewer than a_i categories but not much of a gain in adding a further category.

We might, for example, take $\mathcal{C}_i = K_{a_i}$ as an order of \mathcal{C} just in case the function $f(n) = n\text{Val}(K_n)$ has a relative minimum at $n = a_i$. (We might also be interested in the absolute minimum of f , defined in this or some more appropriate way—we might take this as defining an absolute order of grammaticalness and an overriding bifurcation of strings into grammatical and ungrammatical, with the grammatical including as a proper subclass those generated by the grammar.)

In the way just sketched we might prescribe a general procedure Ψ such that, given a grammar G , $\Psi(G)$ is a hierarchy \mathcal{C} of categories compatible with G , by which degree of grammaticalness is defined for each string in the terminal vocabulary of G . It would then be correct to say that a grammar not only generates sentences with structural descriptions but also assigns to each string, whether generated or not, a degree of grammaticalness that measures its deviation from the set of perfectly well-formed sentences as well as a partial structural description that indicates how this string deviates from well-formedness.

It is hardly necessary to emphasize that this proposal is, in its details, highly tentative. Undoubtedly there are many other ways to approach this complex question.

1.6 Minimum-Redundancy Codes

Before a message can be transmitted, it must be coded in a form appropriate to the medium through which it will pass. This coding can be accomplished in many ways; the procedure becomes of some theoretical interest, however, when we ask about its efficiency. For a given alphabet, what codes will, on the average, give the shortest encoded messages? Such codes are called *minimum-redundancy codes*. Natural languages are generally quite redundant; how to encode them to eliminate that redundancy poses a challenging problem.

The question of coding efficiency becomes especially interesting when we recognize that every channel is noisy, so that an efficient code must not only be short but at the same time must enable us to keep erroneous transmissions below some specified probability. The solutions that have been found for this problem constitute the real core of information theory as it is applied to many practical problems in communication engineering. Inasmuch as psychologists and linguists have not yet exploited these fundamental results for noisy channels, we shall limit our attention here to the simpler problem of finding minimum-redundancy codes for noiseless channels.

The problem of optimal coding can be posed as follows: we know from Sec. 1.3 that an alphabet is used most efficiently when each character occurs independently and equiprobably, that is, when all strings of equal length are equiprobable. So we must find a function θ that maps our natural messages into coded forms in which all sequences of the same length are equiprobable. For the sake of simplicity, let us assume that the messages can be divided into independent units that can be separately encoded. In order to be definite, let us imagine that we are dealing with printed English and that we are willing to assume that successive words are independent. Each time a space occurs in the text the text accumulated since the preceding space is encoded as a unit. For each word, therefore, we shall want to assign a sequence of code symbols in such a way that, on the average, all the code symbols will be used independently and equally often and in such a way that we shall be able to segment the coded messages to recover the original word units when the time comes to decode it.

First, we observe that in any minimum-redundancy code the length of a given coded word can never be less than the length of a more probable coded word. If the more probable word were longer, a saving in the average length could be achieved by simply reversing the codes assigned to the two words. We begin, therefore, by ranking the words in order of decreasing probability of occurrence. Let p_r represent the probability of

the word ranked r , and let c_r represent the length of its encoded representation; that is to say, we rank the words

$$p_1 \geq p_2 \geq \dots \geq p_{N-1} \geq p_N,$$

where N is the number of different words in the vocabulary. For a minimum redundancy code we must then have

$$c_1 \leq c_2 \leq \dots \leq c_{N-1} \leq c_N.$$

Note, moreover, that the mean length C of an encoded word will be

$$C = \sum_{r=1}^N p_r c_r. \quad (25)$$

Obviously, the mean length would be a minimum if we could use only one-letter words, but this would entail too large a number D of different code characters. Ordinarily, our choice of D is limited by the nature of the channel. Of course, it is not length per se that we want to minimize but length per unit of information transmitted. The problem is to minimize C/H , the length per bit (or to maximize H/C , the amount of information per unit length), subject to the subsidiary conditions that $\sum p_r = 1$ and that the coded message be uniquely decodable.

By virtue of Assumption 2 in Sec. 1.3 it would seem that H/C , the information per letter in the encoded words, cannot be greater than $\log D$, the capacity of the coding alphabet. From that fact we might try to move directly to a lower bound,

$$C \geq \frac{H}{\log D}. \quad (26)$$

Although this inequality is correct, it cannot be derived as a simple consequence of Assumption 2. Consider the following counter-example (Feinstein, 1958): we have a vocabulary of three words with probabilities $p_1 = p_2 = 2p_3 = 0.4$ and we code them into the binary alphabet $\{0, 1\}$ so that $\theta(1) = 0$, $\theta(2) = 1$, and $\theta(3) = 01$. Now we can easily compute that $C = 1.2$, $H = 1.52$, and $\log_2 D = 1$, so that the average length is less than the bound stated in Eq. 26. The trouble, of course, is that θ does not yield a true code, in the sense defined in Chapter 11; the coded messages are not uniquely decodable. If, however, we add to Assumption 2 the further condition of unique decodability, the lower bound stated in Eq. 26 can be established. The further condition is most easily phrased in terms of a left tree code, in which no coded word is an initial segment of any other coded word. By using Eq. 21 we can write

$$H = - \sum_{i=1}^N p_i \log p_i \leq - \sum_{i=1}^N p_i \log \frac{D^{-c_i}}{\sum_{l=1}^N D^{-c_l}} = \log \sum_{l=1}^N D^{-c_l} + \sum_{i=1}^N p_i c_i \log D.$$

For left tree codes we know, from Eq. 4, in Chapter 11, that $\sum D^{-c_i} \leq 1$; therefore,

$$\log \sum D^{-c_i} \leq \log 1 = 0,$$

so we can write

$$H \leq \sum_{i=1}^N p_i c_i \log D,$$

from which the desired inequality of Eq. 26 follows by rearranging terms.

If Eq. 26 sets a lower bound on the mean length C , how closely can we approach it? The following theorem, due to Shannon (1948), provides the answer:

Theorem 4. *Given a vocabulary V of N words with information H and a coding alphabet A of D code symbols, it is possible to code the words by finite strings of code symbols from A in such a way that C , the average number of code symbols per word, satisfies the inequality*

$$\frac{H}{\log D} \leq C < \frac{H}{\log D} + 1. \quad (27)$$

The proof has been published in numerous places; see, for example, Feinstein (1958, Chapter 2) or Fano (1961, Chapter 3).

Instead of proving here that such minimum-redundancy codes exist, we shall consider ways of constructing them. Both Shannon (1948) and Fano (1949) proposed methods of constructing codes that approach minimum redundancy asymptotically as the length of the coding unit is enlarged to include progressively longer sequences of words. In 1952, however, Huffman discovered a method of systematically constructing minimum-redundancy codes for finite vocabularies without resorting to any limiting operations.

Huffman assumes that the vocabulary to be encoded is finite, that the probability of each word is known in advance, that a left tree code can be used, and that all code symbols will be of unit length. Within these limits, let us now consider the special conditions that a minimum-redundancy code must satisfy:

1. No two words can be represented by identical strings of code symbols.
2. It must be possible for a receiver to segment coded messages into the coded words that comprise them. (This restriction is discussed in Chapter 11, Sec. 2.) The printer's use of a special symbol (space) to mark word boundaries in a natural code is in general too inefficient for minimum redundancy codes. Proper segmentation in the sense of boundary markers is ensured, however, by the assumption that it must be a left tree code.
3. If the words are ranked in order of decreasing probability p_r , then the length of the r th word, c_r , must satisfy the inequalities

$$c_1 \leq c_2 \leq \dots \leq c_{N-1} = c_N.$$

Because all the code symbols are equally long, c_r can be interpreted simply as the number of symbols used to code the r th word. In a minimum redundancy tree code $c_{N-1} = c_N$ because the first c_{N-1} symbols used to code the N th word cannot be the coded form of any other word; that is to say, the coded forms of words $N - 1$ and N must differ in their first c_{N-1} symbols, and, if they do, no additional symbols are needed to encode word N .

4. At least two (and not more than D) words of code length c_N have codes that are identical except for their final digits. Imagine a minimum redundancy tree code in which this was not true; then the final code symbols could be deleted, thus shortening the average length of a coded word and so leading to a contradiction.

5. Each possible string of $c_N - 1$ code symbols must be used either to represent a word or some initial segment of the sequence must represent a word. If such a string of symbols existed and was not used, the average length of the coded words could be reduced by using it in place of some longer string.

These restrictions are sufficient to determine the following procedure, which we shall outline for a binary coding alphabet, $D = 2$. List the words from most probable to least probable. By (3), $c_{N-1} = c_N$, and, by (4), there are exactly two words of code length c_N that must be identical except for their final symbols. So we can assign 0 as the final digit of the $(N - 1)$ th word and 1 as the final digit of the N th word. Once this has been done, the $(N - 1)$ th and N th words taken together are equivalent to a single composite message; its code will be the common (but still unknown) initial segment of length $c_N - 1$ and its probability will be the sum of the probabilities of the two words comprising it. By combining these two words, we create a new vocabulary with only $N - 1$ words in it. Suppose we now reorder the words as before and repeat the whole procedure. We can continue to do so until the reduced vocabulary contains only two words, at which point we assign 0 to one and 1 to the other and the code is completed.

An illustration of this procedure, using a binary code, is shown in Table 2. A vocabulary of nine words is given in order of decreasing probability. The first step is to assign 0 to word h and 1 to word i (or conversely) as their final code symbols, then to combine h and i into a single item in a new derived distribution. The procedure is then repeated for the two least probable items in this new distribution, etc., until all the code symbols have been assigned. The result is to produce a coding tree; it can be seen with difficulty in Table 2, in which its trunk is on the right and its branches extend to the left, or more easily in Fig. 2, in which it has been redrawn in the standard way.

Table 2 Huffman's Method of Constructing a Minimum-Redundancy Code

| Word | Coded Form | Original Distribution | First Derived Distribution | Second | Third | Fourth | Fifth | Sixth | Seventh | Final |
|----------|------------|-----------------------|----------------------------|--------|-------|--------|-------|-------|---------|-------|
| <i>a</i> | 01 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.30 | 0.43 | 0.57 | →1.00 |
| <i>b</i> | 10 | 0.23 | 0.23 | 0.23 | 0.23 | 0.27 | 0.27 | 0.30 | 0.43 | |
| <i>c</i> | 000 | 0.15 | 0.15 | 0.15 | 0.15 | 0.23 | 0.23 | 0.27 | 0.43 | |
| <i>d</i> | 110 | 0.10 | 0.10 | 0.10 | 0.15 | 0.20 | 0.23 | 0.27 | 0.43 | |
| <i>e</i> | 0010 | 0.08 | 0.08 | 0.10 | 0.15 | 0.20 | 0.23 | 0.27 | 0.43 | |
| <i>f</i> | 0011 | 0.07 | 0.07 | 0.10 | 0.15 | 0.20 | 0.23 | 0.27 | 0.43 | |
| <i>g</i> | 1110 | 0.05 | 0.05 | 0.08 | 0.10 | 0.15 | 0.20 | 0.27 | 0.43 | |
| <i>h</i> | 11110 | 0.03 | 0.05 | 0.07 | 0.10 | 0.15 | 0.20 | 0.27 | 0.43 | |
| <i>i</i> | 11111 | 0.02 | 0.05 | 0.07 | 0.10 | 0.15 | 0.20 | 0.27 | 0.43 | |

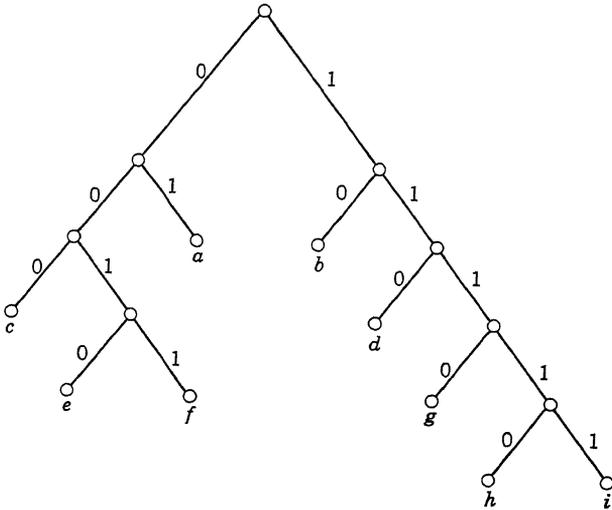


Fig. 2. The coding tree developed for the minimum-redundancy code of Table 2.

In order to evaluate the coding efficiency, we need to know $\log D$, C , and H . The coding alphabet is binary, $D = 2$, and its information capacity is $\log_2 D = 1$ bit per symbol. The average length of an encoded word can be easily computed from Table 2 by Eq. 25; the result is 2.80 binary code symbols per word. The amount of information in the original distribution or word probabilities can be computed by Eq. 13; the result is 2.781 bits per word. In terms of Theorem 4, therefore, we have

$$\frac{2.78}{1} \leq 2.80 < \frac{2.78}{1} + 1,$$

which indicates that for this example the average length is already quite close to its lower bound. The redundancy of the coded signal—as defined by Eq. 17—is less than 1%.

It should be obvious that errors in the transmission or reception of minimum-redundancy codes are difficult to detect. Every branch of the coding tree is utilized and errors convert any intended message into a perfectly plausible alternative message. Considerable study has been devoted to the most efficient ways to introduce redundancy into the code deliberately in order to make errors easier to detect and to correct. But these artificially redundant codes are not surveyed here. The point should be noted, however, that the redundancy of natural codes may not be so inefficient as it seems, for it can help us to communicate under less than optimal conditions.

The reason that minimum-redundancy codes are important is an economic one. There is a cost to communication and someone must pay for it. It is often appropriate to use C , the average length of the message, as a measure of the cost, since it takes either more time or more equipment to transmit more symbols. It should be recognized, however, that the economy achieved by minimizing C/H affects the supply price, not the demand price of this commodity (Marschak, 1960). The supply price is the lowest price the supplier is willing to charge; the demand price is the highest price the buyer is willing to pay. The demand price depends on the payoff that the customer expects to obtain by using the information; since that use will ordinarily involve the meaning of the message in an essential way, it takes us beyond the limits we have arbitrarily imposed on this chapter.

1.7 Word Frequencies

It is scarcely surprising to find that the various words of a natural language do not occur equally often in any reasonable sample of ordinary discourse. Some words are far more common than others. Psychologists have recognized the importance of these unequal probabilities for any kind of experimentation that uses words as stimuli. It is standard procedure for psychologists to try to control for the effects of unequal familiarity by selecting the words from some tabulation of relative frequencies of occurrence. For English the Thorndike-Lorge (1944) counts are probably the best known and most widely used. An extensive technical literature deals with the various statistics that have been compiled for the (usually written) languages of the world; we shall make no attempt to review or evaluate it in these pages. Instead, we shall concentrate our attention on certain statistical aspects of the vocabulary that seem theoretically most significant.

There is one particularly striking regularity that has been found in these various statistical explorations. The following is perhaps the simplest way to summarize it (Mandelbrot, 1959): consider a (finite or infinite) population of discrete *items*, each of which carries a *label* chosen from a discrete set. Let $n(f, s)$ be the number of different labels that occur exactly f times in a sample of s items. Then one finds that, for large s ,

$$n(f, s) = G(s)f^{-(1+\rho)}, \quad (28)$$

where $\rho > 0$ and $G(s)$ is a constant depending on the size of the sample.

If Eq. 28 is expressed as a probability density, then it is readily seen that the variance of f is finite if and only if $\rho > 2$ and that the mean of f is finite if and only if $\rho > 1$. In the cases of interest in this section it is often

true that $\rho < 1$, so we are faced with a law that is often judged anomalous (or even pathological) to those prejudiced in favor of the finite means and variances of normal distributions. In the derivation of the normal distribution function, however, it is necessary to assume that we are dealing with the sum of a large number of variables, each of which makes a small contribution relative to the total. When equal contributions are not assumed, however, it is still possible to have stable limit distributions, but either the second moment (and all higher moments) will be infinite, or all moments will be infinite (cf. Gnedenko & Kolmogorov, 1954, Chapter 7). Such is the distribution underlying Eq. 28.

Nonnormal limit distributions might be dismissed as mathematical curiosities of little relevance were it not for the fact that they have been observed in a wide variety of situations. As Mandelbrot (1958) has pointed out, these situations seem especially common in the social sciences. For example, if the items are quantities of money and the labels are the names of people who earn each item, then $n(f, s)$ will be the number of people earning exactly f units of money out of a total income equal to s . In this form the law was first stated (with $\rho > 1$) by Pareto (1897). Alternatively, if the items are taxonomic species and the labels are the names of genera to which they belong, then $n(f, s)$ will be the number of genera each with exactly f species. In this form the law was first stated by Willis (1922), then rationalized by Yule (1924), with $\rho < 1$ (and usually close to 0.5).

In the present instance, if the items are the consecutive words in a continuous discourse by a single author and the labels are sequences of letters used to encode words, then $n(f, s)$ will be the number of letter sequences (word types) that occur exactly f times in a text of s consecutive words (word tokens). In this form the law was first stated by Estoup (1916), rediscovered by Condon (1928), and intensively studied by Zipf (1935). Zipf believed that $\rho = 1$, but further analysis has indicated that usually $\rho < 1$. Considerable data indicating the ubiquity of Eq. 28 were provided by Zipf (1949), and empirical distributions of this general type have come to be widely associated with his name.

When working with word frequencies, it is common practice to rank them in order (as we did for the coding problem in the preceding section) from the most frequent to the least frequent. The rank r is then defined as the number of items that occur f times or more:

$$r = \sum_{j=f}^{\infty} n(j, s).$$

If we combine this definition with Eq. 28 and approximate the sum by an integral, then, for large f ,

$$r \sim \frac{G(s)}{\rho f^{\rho}},$$

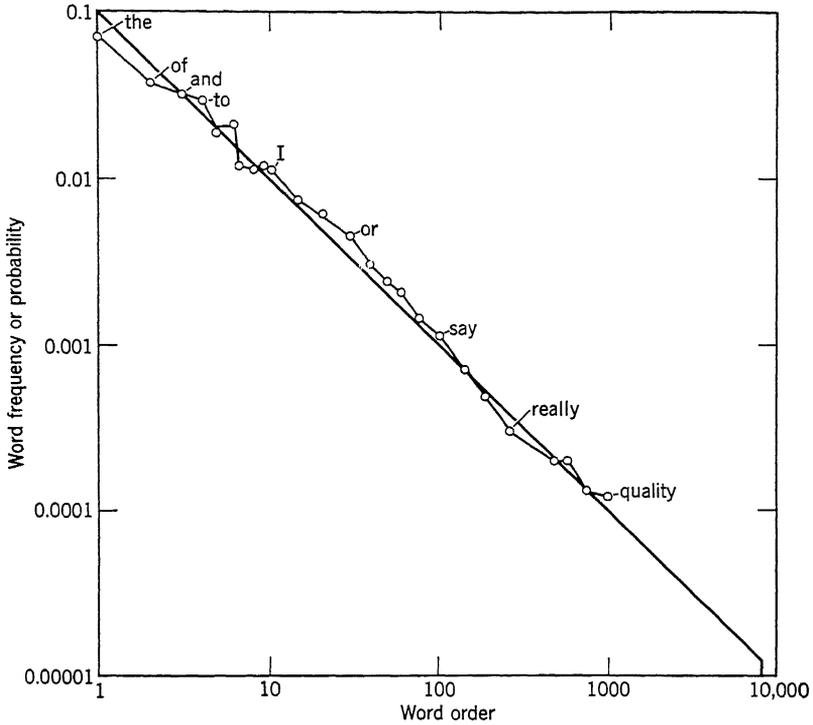


Fig. 3. The rank-frequency relation plotted on log-log coordinates.

which states a reciprocal relation between the ranks r and the frequencies f . We can rewrite this relation as

$$f \sim \left(\frac{G(s)}{\rho r} \right)^{1/\rho} = K' r^{-B}, \quad (29)$$

where $B = 1/\rho$. Therefore

$$\log f \sim K - B \log r,$$

which means that on log-log coordinates the rank-frequency relation should give a straight line with a slope of $-B$. It was with such a graph that the law was discovered, and it is still a popular way to display the results of a word count. An illustration is given in Fig. 3.

The persistent recurrence of stable laws of this nonnormal type has stimulated several attempts at explanation, and there has been considerable discussion of their relative merits. We shall not review that discussion here; the present treatment follows Mandelbrot (1953, 1957) but does little more than introduce the topic in terms of its simplest cases.

Imagine that the coded message is, in fact, a table of random (decimal)

digits. Let the digits 0 and 1 play the role of word-boundary markers; each time 0 or 1 occurs it marks the beginning of a new word. (In this code there are words of zero length; a minor modification can eliminate them if they are considered anomalous.) The probability of getting a particular word of exactly length i is (probability of symbol) ^{i} (probability of boundary marker) = $(0.8)^i(0.2)$, and the number of different words of length i is 8^i .

The critical point to note in this example is that when we order these coded words with respect to increasing length we have simultaneously ordered them with respect to decreasing probability. Thus it is possible to construct Table 3. The one word of zero length has a probability of 0.2

Table 3 The Rank-Frequency Relation for a Random Code

| Length i | Probability $p(w_i)$ | Number D^i | Ranks ΣD^i | Average Rank $r(w_i)$ |
|---------------|-------------------------|-----------------|-----------------------|--------------------------|
| 0 | 0.2 | 1 | 1 | 1 |
| 1 | 0.02 | 8 | 2-9 | 5.5 |
| 2 | 0.002 | 64 | 10-73 | 41.5 |
| 3 | 0.0002 | 512 | 74-585 | 329.5 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |

and, since it is the most probable word, it receives rank 1. The eight words one digit long all have a probability of 0.02 and share ranks 2 through 9; we assign them all the average rank 5.5; and so the table continues. When we plot these values on log-log coordinates, we obtain the function shown in Fig. 4. Visual inspection indicates that the slope is slightly steeper than -1 , which is also characteristic of many natural-language texts.

It is not difficult to obtain the general equation relating probability to average rank for this simple random case (Miller, 1957). Let $p(\#)$ be the probability of a word-boundary marker, and let $1 - p(\#) = p(L)$ be the probability of a letter. If the alphabet (excluding $\#$) has D letters, then $p(L)/D$ is the probability of any particular letter, and $p(w_i) = p(\#)p(L)^i D^{-i}$ is the probability of any particular word of length i ($= 0, 1, \dots$). This quantity will prove to be more useful when written

$$\begin{aligned}
 p(w_i) &= p(\#)e^{-i \log D} e^{i \log p(L)} \\
 &= p(\#)(e^{i \log D})^{-\left(1 - \frac{\log p(L)}{\log D}\right)} \\
 &= p(\#)(e^{i \log D})^{-B}.
 \end{aligned}
 \tag{30}$$

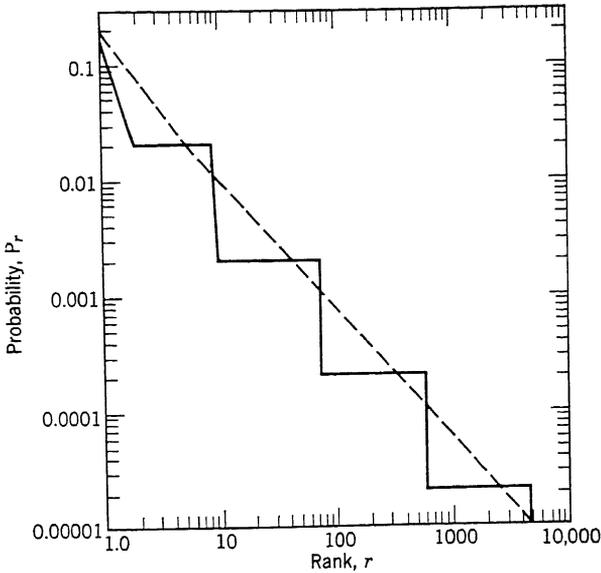


Fig. 4. The rank-frequency relation for strings of random digits occurring between successive occurrences of 0 or 1. The solid line represents the expected function and the dashed line represents the average ranks.

Since there are D^j different words of exactly length j , there must be $\sum_{j=0}^i D^j$ of them equal to or shorter than i , so that when we rank them in order of increasing length the D^j words of length j will receive ranks $1 + \sum_{j=0}^{i-1} D^j$ to $\sum_{j=0}^i D^j$. The average rank will be

$$\begin{aligned} r(w_i) &= \frac{1}{2} \left(1 + \sum_0^{i-1} D^j + \sum_0^i D^j \right) \\ &= \frac{1}{2} \left(1 + \frac{D^i - 1}{D - 1} + \frac{D^{i+1} - 1}{D - 1} \right) \\ &= D^i \frac{D + 1}{2(D - 1)} + \frac{D - 3}{2(D - 1)}, \end{aligned}$$

which will prove more useful if we write

$$D^i = e^{i \log D} = \frac{2(D - 1)}{D + 1} \left[r(w_i) - \frac{D - 3}{2(D - 1)} \right] = \frac{2(D - 1)}{D + 1} [r(w_i) - c], \quad (31)$$

for now Eqs. 30 and 31 combine to give

$$p(w_i) = p(\#) \left\{ \frac{2(D-1)}{D+1} [r(w_i) - c] \right\}^{-B} = K' [r(w_i) - c]^{-B}, \quad (32)$$

which can be recognized as a variant form of Eq. 29, where

$$B = 1 - \frac{\log p(L)}{\log D}, \quad c = \frac{D-3}{2(D-1)}, \quad \text{and} \quad K' = p(\#) \left[\frac{2(D-1)}{D+1} \right]^{-B}.$$

Thus a table of random numbers can be seen to follow the general type of law that has been found for word frequencies. If we take $D = 26$ and $p(\#) = 0.18$ to represent written English, then

$$B = 1 - \frac{\log 0.82}{\log 26} = 1.06,$$

$$c = \frac{26-3}{50} = 0.46,$$

and

$$K' = 0.18 \left(\frac{50}{27} \right)^{-1.06} = 0.09,$$

so we have

$$p(w_i) = 0.09 [r(w_i) - 0.46]^{-1.06}.$$

Since $c = 0.46$ will quickly become negligible as $r(w_i)$ increases, we can write

$$p(w_i) \sim 0.09 r(w_i)^{-1.06},$$

which is, in fact, close to the function that has been observed to hold for many normal English texts (Zipf, for example, liked to put $K' = 0.1$ and $B = 1$).

The hypothesis that word boundaries occur more or less at random in English text, therefore, has some reasonable consequences. It helps us to understand why the probability of a word decreases so rapidly as a function of its length—which is certainly true, on the average, for English. The critical step in the derivation of Eq. 32, however, occurs when we note that for the random message the rank with respect to increasing length and the rank with respect to decreasing probability are the same. In English, of course, this precise equivalence of rankings does not hold—otherwise we would never let our most frequent word *the* require three letters—but it holds approximately. Miller and Newman (1958) have verified the prediction that the *average* frequency of words of length i is a reciprocal function of their *average* rank with respect to increasing length, where the slope constant for the length-frequency relation on log-log coordinates is close to but perhaps somewhat smaller than B .

In Sec. 1.6 we noted that for a minimum-redundancy code the length of any given word can never be less than the length of a more probable word. Suppose, therefore, that we consider the rank-frequency relation for optimal codes, that is, for codes in which the lower bound on the average length C is actually realized, so that $C = H/\log D$. This optimal condition will hold when the length i of any given word is directly proportional to the amount of information associated with it:

$$i = \rho \frac{-\log p(w_i)}{\log D},$$

where ρ depends on the choice of scale units. This equation can be rewritten as

$$p(w_i) = (e^{i \log D})^{-1/\rho},$$

which is Eq. 30 again, with $B = 1/\rho$. From here on the argument can proceed exactly as before. We see, therefore, that the rank-frequency relation holds quite generally for minimum-redundancy codes because such codes (like tables of random numbers) use all equally long sequences of symbols equally probably. The fact that both minimum-redundancy codes and natural languages (which are certainly far from minimum-redundancy) share the rank-frequency relation in Eq. 29 is interesting, of course, but it provides no basis whatsoever for any speculation that there is something optimal about the coding used in natural languages.

The choice of the digits 0 and 1 as boundary markers to form words in a table of random numbers was completely arbitrary; any other digits would have served equally well. If we generalize this observation to English texts, it implies that we might choose some character other than the space as a boundary marker. Miller and Newman (1958) have studied the rank-frequency relation for a (relatively small) sample of pseudo-words formed by using the letter E as the word boundary (and treating the space as just another letter). The null word EE was most frequent, followed closely by ERE , $E\#E$, and so on. As predicted, a function of the general type of Eq. 29 was also obtained for these pseudo-words (but with a slope constant B slightly less than unity, perhaps attributable to inadequate sampling).

There is an enormous psychological difference between the familiar words formed by segmenting on spaces and the apparently haphazard strings that result when we segment on E . Segmenting on spaces respects the highly overlearned strings—Miller (1956) has referred to them as chunks of information in order to distinguish sharply from the bits of information defined in Sec. 1.3—that normally function as unitary, psychological elements of language. It seems almost certain, therefore,

that an evolutionary process of selection must have been working in favor of short words—some psychological process that would not operate on the strings of characters between successive *Es*. Thus we find many more very long, very improbable pseudo-words.

In one form or another the hypothesis that we favor short words has been advanced by several students of language statistics. Zipf (1935) has referred to it as the *law of abbreviation*: whenever a long word or phrase suddenly becomes common, we tend to shorten it. Mandelbrot (1961) has proposed that historical changes in word lengths might be described as a kind of random walk. He reasons that the probability of lengthening a word and the probability of shortening it should be in equilibrium, so that a steady state distribution of word lengths could be maintained. If the probability of abbreviation were much greater than the probability of expansion, the vocabulary would eventually collapse into a single word of minimum length. If expansion were more likely than abbreviation, on the other hand, the language would evolve toward a distribution with $B < 1$, and, presumably, some upper bound would have to be imposed on word lengths in order for the series $p(w_i)$ to converge, so that $\sum p(w_i) = 1$. It should be noted, however, that the existence of a relation in the form of Eq. 29 does not depend in any essential way on some prior psychological law of abbreviation. The central import of Mandelbrot's earlier argument is that Eq. 29 can result from purely random processes. Indeed, if there is some law of abbreviation at work, it should manifest itself as a deviation from Eq. 29—presumably in a shortage of very long, very improbable words, a shortage that would not become apparent until extremely large samples of text had been tabulated.

The occurrence of the rank-frequency relation of Eq. 29 does not constitute evidence of some powerful and universal psychological force that shapes all human communication in a single mold. In particular, its occurrence does not constitute evidence that the signal analyzed must have come from some intelligent or purposeful source. The rank-frequency relation, Eq. 29 has something of the status of a null hypothesis, and, like many null hypotheses, it is often more interesting to reject than to accept.

These brief paragraphs should serve to introduce some of the theoretical problems in the statistical analysis of language. There is much more that might be said about the analysis of style, cryptography, estimations of vocabulary size, spelling systems, content analysis, etc., but to survey all that would lead us even further away from matters of central concern in Chapters 11, 12, and 13.

If one were to hazard a general criticism of the models that have been constructed to account for word frequencies, it would be that they are still far too simple. Unlike the Markovian models that envision D^k parameters,

explanations for the rank frequency relation use only two or three parameters. The most they can hope to accomplish, therefore, is to provide a null hypothesis and to indicate in a qualitative way (perhaps) the kind of systems we are dealing with. They can tell us, for example, that any grammatical rule regulating word lengths must be regarded with considerable suspicion—in an English grammar, at least.

The complexity of the underlying linguistic process cannot be suppressed very far, however, and examples of nonrandom aspects are in good supply. For example, if we partition a random population on the basis of some independent criterion, the same probability distribution should apply to the partitions as to the parent population. If, for example, we partitioned according to whether the words were an odd or an even number of running words away from the beginning of the text or according to whether their initial letters were in the first or the last half of the alphabet, etc., we would expect the same rank-frequency relation to apply to the partitions as to the original population. There are, however, several ways to partition the parent population that look as though they ought to be independent but turn out in fact not to be. Thus, for example, Yule (1944) established that the same distribution does not apply when different categories (nouns, verbs, and adjectives) are taken separately; Miller, Newman, and Friedman (1958) showed a drastic difference between the distributions of content words (nouns, verbs, adjectives, adverbs) and of function words (everything else), and Miller (1951, p. 93) demonstrated that the distribution can be quite different if we consider only the words that occur immediately following a given word, such as *the* or *of*. There is nothing in our present parsimonious theories of the rank-frequency relation that could help us to explain these apparent deviations from randomness.

In an effort to achieve a more appropriate level of complexity in our descriptions of the user, therefore, we turn next to models that take account of the underlying structure of natural languages—models that, for lack of a better name, we shall refer to here as algebraic.

2. ALGEBRAIC MODELS

If the study of actual linguistic behavior is to proceed very far, it must clearly pay more than passing notice to the competence and knowledge of the performing organism. We have suggested that a generative grammar can give a useful and informative characterization of the competence of the speaker-hearer, one that captures many significant and deep-seated aspects of his knowledge of his own language. The question is, therefore, how does he put his knowledge to use in producing a desired sentence or

in perceiving and interpreting the structure of presented utterances? How can we construct a model for the language user that incorporates a generative grammar as a fundamental component? This topic has received almost no study, so we can do little more than introduce a few speculations.

As we observed in the introduction to this chapter, models of linguistic performance can generally be interpreted interchangeably as depicting the behavior of either a speaker or a hearer. For concreteness, in the present sections we shall concentrate on the listener's task and frame our discussion largely in perceptual terms. This decision is, however, a matter of convenience, not of principle.

Unfortunately, the bulk of the experimental research on speech perception has involved the recognition of individual words spoken in isolation as part of a list (cf. Fletcher, 1953) and so is of little value to us in understanding the effects of grammatical structure on speech perception. That such effects exist is clear from the fact that the same words are easier to hear in sentences than in isolation (Miller, Heise, & Lichten, 1951; Miller, 1962a). How these effects are caused, however, is not at all clear.

Let us take as our starting point the sentence-recognizing device introduced briefly in Chapter 11, Sec. 6.4. Instead of a relatively passive process of acoustic analysis followed by identification and symbolic representation, we imagined (following Halle & Stevens, 1959, 1962) an active device that recognizes its input by discovering what must be done in order to generate a signal (in some possibly derived form) to match it. At the heart of this active device, of course, is a component M that contains rules for generating a matching signal. Associated with M would be components to analyze and (temporarily) to store the input, components that reflect various semantic and situational constraints suggested by the context of the sentence, a heuristic component that could make a good first guess, a component to make the comparison of the input and the internally generated signals, and perhaps others. On the basis of an initial guess, the device generates an internal signal according to the rules stored in M and tests its guess against the input signal. If the match is unsatisfactory, the discrepancy is used to make a better guess. In this manner the device proceeds to modify its own internal signal until the match is judged satisfactory or the input is dismissed as unintelligible. The program for generating the matching signal can be taken as the symbolic representation of the input.

If it is granted that such a sentence-recognizer can provide a plausible model for human speech perception, we can take it as our starting point and can proceed to try to specify it more precisely. In particular, the two parts of it that seem to perform the most important functions are the contextual component, which helps to generate a first guess, and the

grammatical component M , which imposes the rules for generating the internal signal. We should begin by studying those two components. Even if it were feasible, a study of the ways contextual information can be stored and brought to bear would lead us far beyond the limits we have placed on this discussion. With respect to M , however, the task seems easier. The way the rules for synthesizing sentences might operate is, of course, very much in our present line of sight.

We are concerned with a finite device M in which are stored the rules of a generative grammar G . This device takes as its input a string x of symbols and attempts to understand it; that is to say, M tries to assign to x a certain structural description $F(x)$ —or a set $\{F_1(x), \dots, F_m(x)\}$ of syntactic descriptions in the case of a sentence x that is structurally ambiguous in m different ways. We shall not try to consider all of those real but obscure aspects of understanding that go beyond the assignment of syntactic structural descriptions to sentences, nor shall we consider the situational or contextual features that may determine which of a set of alternative structural descriptions is actually selected in a particular case. There is no point of principle underlying this limitation to syntax rather than to semantics and to single sentences rather than their linguistic and extra-linguistic contexts—it is simply an unfortunate consequence of limitations in our current knowledge and understanding. At present there is little that can be said, with much precision, about those further questions. [See Ziff (1960) and Katz & Fodor (1962) for discussion of the problems involved in the development of an adequate semantic theory and some of the ways in which they can be investigated].

The device M must contain, in addition to the rules of G , a certain amount of computing space, which may be utilized in various different ways, and it must be equipped to perform logical operations of various sorts. We require, in particular, that M assign a structural description $F_i(x)$ to x only if the generative grammar G stored in the memory of M assigns $F_i(x)$ to x as a possible structural description. We say that the device M (*partially*) *understands the sentence x in the manner of G* if the set $\{F_1(x), \dots, F_m(x)\}$ of structural descriptions provided by M with input x is (included in) the set assigned to x by the generative grammar G . In particular, M does not accept as a sentence any string that is not generated by G . (This restriction can, of course, be softened by introducing degrees of grammaticalness, after the manner of Sec. 1.5, but we shall not burden the present discussion with that additional complication.) M is thus a finite transducer in the sense of Chapter 12, Sec. 1.5. It uses its information concerning the set of all strings in order to determine which of them are sentences of the language it understands and to understand sentences belonging to this language. This information, we assume, is represented in

the form of rules of the generative grammar G stored in the memory of M .

Before continuing, we should like to say once more that it is perfectly possible that M will not contain enough computing space to allow it to understand all sentences in the manner of the device G whose instructions it stores. This is no more surprising than the fact that a person who knows the rules of arithmetic perfectly may not be able to perform many computations correctly in his head. One must be careful not to obscure the fundamental difference between, on the one hand, a device M storing the rules G but having enough computing space to understand in the manner of G only a certain proper subset L' of the set L of sentences generated by G and, on the other hand, a device M^* designed specifically to understand only the sentences of L' in the manner of G . The distinction is perfectly analogous to the distinction between a device F that contains the rules of arithmetic but has enough computing space to handle only a proper subset Σ' of the set Σ of arithmetical computations and a device F^* that is designed to compute only Σ' . Thus, although identical in their behavior to F^* and M^* , F and M can improve their behavior without additional instruction if given additional memory aids, but F^* and M^* must be redesigned to extend the class of cases that they can handle. It is clear that F and M , the devices that incorporate competence whether or not it is realized in performance, provide the only models of any psychological relevance, since only they can explain the transfer of learning that we know occurs when memory aids are in fact made available.

In particular, if the grammar G incorporated in M exceeds any finite automaton in generative capacity, then we know that M will not be able to understand all sentences in the manner of G . There would be little reason to expect, a priori, that the natural languages learned by humans should belong to the special family of sets that can be generated by one-sided linear grammars (cf. Defs. 6 and 7, Chapter 12, Sec. 4.1) or by nonself-embedding context-free grammars (cf. Proposition 58 and Theorem 33, Chapter 12, Sec. 4.6). In fact, they do not, as we have observed several times. Consequently, we know that a realistic model M for the perceiver will incorporate a grammar G that generates sentences that M cannot understand in the manner of G (without additional aids). This conclusion should occasion no surprise; it leads to none of the paradoxical consequences that have occasionally been suggested. There has been much confusion about this matter and we should like to reemphasize the fact that the conclusion we have reached is just what should have been expected.

We can construct a model for the listener who understands a presented sentence by specifying the stored grammar G , the organization of memory, and the operations performable by M . We determine a class of perceptual models by stating conditions that these specifications must meet. In

Sec. 2.1 we consider perceptual models that store rewriting systems. Then in Sec. 2.2 we discuss possible features of perceptual models that incorporate transformational grammars.

2.1 Models Incorporating Rewriting Systems

Let us suppose that we have a language L generated by a context-sensitive grammar G that assigns to each sentence of L a P -marker—a labeled tree or labeled bracketing—in the manner we have already considered. What can we say about the understanding of sentences by the speaker of L ? For example, what can we say about the class of sentences of his language that this speaker will be able to understand at all? If we construct a finite perceptual device M that incorporates the rules of G in its memory, to what extent will M be able to understand sentences in the manner of G ?

In part, we answered this question in Sec. 4.6 of Chapter 12. Roughly, the answer was the following. Suppose that we say that *the degree of self-embedding of the P -marker Q is m* if m is the largest integer meeting the following condition: there is, in the labeled tree that represents Q , a continuous path passing through $m + 1$ nodes N_0, \dots, N_m , each with the same label, where each N_i ($i \geq 1$) is fully self-embedded (with something to the left and something to the right) in the subtree dominated by N_{i-1} ; that is to say, the terminal string of Q can be written in the form

$$xy_0y_1 \dots y_{m-1}zv_{m-1} \dots v_1v_0w, \quad (33)$$

where N_m dominates z , and for each $i < m$, N_i dominates

$$y_i \dots y_{m-1}zv_{m-1} \dots v_i, \quad (34)$$

and none of the strings $y_0, \dots, y_{m-1}, v_0, \dots, v_{m-1}$ is null. Thus, for example, in Fig. 5 the degree of self-embedding is two.

In Sec. 4.6 of Chapter 12 we presented a mechanical procedure Ψ that can be regarded as having the following effect: given a grammar G and an integer m , $\Psi(G, m)$ is a finite transducer M that takes a sentence x as input and gives as output a structural description $F(x)$ (which is, furthermore, a structural description assigned to x by G) wherever $F(x)$ has a degree of self-embedding of no more than m ; that is to say, where m is a measure of the computing space available to a perceptual model M , which incorporates the grammar G , M will partially understand sentences in the manner of G just to the extent that the degree of self-embedding of their structural descriptions is not too great. As the amount of computing

space available to the device M increases, M will understand more deeply embedded structures in the manner of G . For any given sentence x there is an m sufficiently large so that the device M with computing space determined by m [i.e., the device $\Psi(G, m)$] will be capable of understanding x in the manner of G ; M does not have to be redesigned to extend its capacities in this way. Furthermore, this is the best result that can be achieved, since self-embedding is, as was proved in Chapter 12, precisely the property that distinguishes context-free languages from the regular languages that can be generated (accepted) by finite automata.

In Chapter 12 this result was stated only for a certain class K of context-free grammars. We pointed out that the class K contains a grammar for every context-free language and that it is a straightforward matter to drop many, if not all, of the restrictions that define K . Extension to context-sensitive grammars is another matter, however, and the problem of finding an optimal finite transducer that understands the sentences of G as well as possible, for any context-sensitive G , has not been investigated at all. Certain approaches to this question are suggested by the results of Matthews', discussed in Chapter 12, Sec. 4.2, on asymmetrical context-sensitive grammars and PDS automata, but these have not yet been pursued.

These restrictions aside, the procedure Ψ of Sec. 4.6, Chapter 12, provides an optimal perceptual model (i.e., an optimal finite recognition routine) that incorporates a context-free grammar G . Given G , we can immediately construct such a device in a mechanical way, and we know that it will do as well as can be done by any device with bounded memory in understanding sentences in the manner of G . As the amount of memory increases, its capacity to understand sentences of G increases without limit. Only self-embedding beyond a certain degree causes it to fail when memory is fixed. We can, in fact, rephrase the construction so that the procedure Ψ determines a transducer $\Psi(G)$ which understands all sentences in the manner of G , where $\Psi(G)$ is a "single-pass" device with only push-down storage, as shown in Sec. 4.2, Chapter 12.

Observe that the optimal perceptual model $M = \Psi(G, m)$, where m is fixed, may fail to understand sentences in the manner of G even when

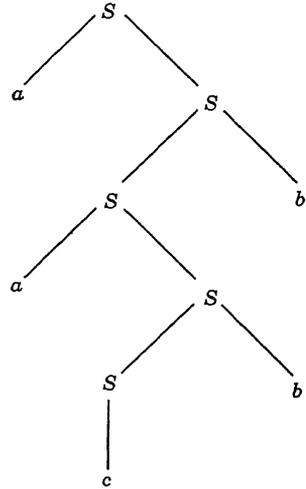


Fig. 5. Phrase marker with a degree of self-embedding equal to two.

the language L generated by G might have been generated by a one-sided linear grammar (finite automaton). For example, the context-free grammar G that gives the structural description in Fig. 5 might be the following:

$$S \rightarrow aS, \quad S \rightarrow Sb, \quad S \rightarrow c. \quad (35)$$

(It is a straightforward matter to extend Ψ to deal with rules of the kind in Example 35.) The generated language is the set of all strings $a^i cb^j$ and is clearly a regular language. Nevertheless, with $m = 1$, $\Psi(G, m)$ will not be capable of understanding the sentence $aacbb$ generated in Fig. 5 *in the manner of G* , since this derivation has a degree of self-embedding equal to two. The point is that although a finite automaton can be found to accept the sentences of this language it is not possible to find a finite device that understands all of its sentences in the manner of the particular generative process G represented in Example 35.

Observe also that the perceptual device $\Psi(G, m)$ is nondeterministic. As a perceptual model it has the following defect. Suppose that G assigns to x a structural description D with degree of self-embedding not exceeding m . Then, as we have indicated, the device $\Psi(G, m)$ will be capable of computing in such a way that it will map x into D , thus interpreting x in the manner of G . Being nondeterministic, however, it may also, given x , compute in such a way that it will fail to map x into a structural description at all. If $\Psi(G, m)$ fails to interpret x in the manner of G on a particular computation, we can conclude nothing about the status of x with respect to the grammar G , although if $\Psi(G, m)$ does map x into a structural description D we can conclude that G assigns D to x . We might investigate the problem of constructing a deterministic perceptual model that partially understands the output of a context-free grammar, or a model with nondeterminacy matching the ambiguity of the underlying grammar—that is, a model that may block on a computation with a particular string only if this string is either not generated by the grammar from which the model is constructed or is generated only by a derivation that is too deeply self-embedded for the device in question—but this matter has not yet been carefully investigated. It is clear, however, that such devices unlike $\Psi(G, m)$, would involve a restriction on the right-recursive elements in the structural descriptions (i.e., on right branchings). See, in this connection, the example on p. 473.

Self-embedding is the fundamental property that takes a system outside of the generative capacity of a finite device, and self-embedding will ultimately result from nesting of dependencies, since the nonterminal vocabulary is finite. However, the nesting of dependencies, even short of self-embedding, causes the number of states needed in the device $\Psi(G, m)$ to

increase quite rapidly with the length of the input string that it is to understand. Consequently, we would expect that nested constructions should become difficult to understand even when they are, in principle, within the capacity of a finite device, since available memory (i.e., number of states) is clearly quite limited for real-time analytic operations, a fact to which we return in Sec. 2.2. Indeed, as we observed in Chapter 11 (cf. Example 11 in Sec. 3), nested structures even without self-embedding quickly become difficult or impossible to understand.

From these observations we are led to conclude that sentences of natural languages containing nested dependencies or self-embedding beyond a certain point should be impossible for (unaided) native speakers to understand. This is indeed the case, as we have already pointed out. There are many syntactic devices available in English—and in every other language that has been studied from this point of view—for the construction of sentences with nested dependencies. These devices, if permitted to operate freely, will quickly generate sentences that exceed the perceptual capacities (i.e., in this case, the short-term memory) of the native speakers of the language. This possibility causes no difficulties for communication, however. These sentences, being equally difficult for speaker and hearer, simply are not used, just as many other proliferations of syntactic devices that produce well-formed sentences will never actually be found.

There would be no reason to expect that these devices (which are, of course, continually used when nesting is kept within the bounds of memory restriction) should disappear as the language evolves; and, in fact, they do not disappear, as we have observed. It would be reasonable to expect, however, that a natural language might develop techniques to paraphrase complex nested sentences as sentences with either left-recursive or right-recursive elements, so that sentences of the same content could be produced with less strain on memory. That expectation, formulated by Yngve (1960, 1961) in a rather different way, to which we return, is well confirmed. Alongside such self-embedding English sentences as *if, whenever X then Y, then Z*, we can have the basically right-branching structure *Z if whenever X, then Y*, and so on in many other cases. In particular, many singulary grammatical transformations in English seem to be primarily stylistic; they convert one sentence into another with much the same content but with less self-embedding. Alongside the sentence *that the fact that he left was unfortunate is obvious*, which doubly embeds *S*, we have the more intelligible and primarily right-recursive structure *it is obvious that it was unfortunate that he left*. Similarly, we have a transformation that converts *the cover that the book that John has has* to *John's book's cover*, which is left-branching rather than self-embedding. (It should also be noted, however, that some of these so-called stylistic transformations can increase

structural complexity, e.g., those that give “cleft-sentences”—from *I read the book that you told me about* we can form *it was the book that you told me about that I read*, etc.)

Now to recapitulate: from the fact that human memory is finite we can conclude only that some self-embedded structures should not be understandable; from the further assumption that memory is small, we can predict difficulties even with nested constructions. Although sentences are accepted (heard and spoken) in a single pass from left to right, we cannot conclude that there should be any left-right asymmetry in the understandable structures. Nor is there any evidence presently available for such asymmetry. We have little difficulty in understanding such right-branching constructions as *he watched the boy catch the ball that dropped from the tower near the lake* or such left-branching constructions as *all of the men whom I told you about who were exposed to radiation who worked half-time are still healthy, but the ones who worked full time are not or many more than half of the rather obviously much too easily solved problems were dropped last year*. Similarly, no conclusion can be drawn from our present knowledge of the distribution of left-recursive and right-recursive elements in language. Thus, in English, right-branching constructions predominate; in other languages—Japanese, Turkish—the opposite is the case. In fact, in every known language we find right-recursive, left-recursive, and self-embedding elements (and, furthermore, we find coordinate constructions that exceed the capacity of rewriting systems entirely, a fact to which we return directly).

We have so far made only the following assumptions about the model M for the user:

1. M is finite;
2. M accepts (or produces) sentences from left-to-right in a single pass;
3. M incorporates a context-free grammar as a representation of its competence in and knowledge of the language.

Of these, (3) is surely false, but the conclusions concerning recursive elements that we have drawn from it would undoubtedly remain true under a wide class of more general assumptions. Obviously, (1) is beyond question; (2) is an extremely weak assumption that also cannot be questioned, either for the speaker or hearer—note that many different kinds of internal organization of M are compatible with (2), for example, the assumption that M stores a finite string before deciding on the analysis of its first element or that M stores a finite number of alternative assumptions about the first element which are resolved only at an indefinitely later time.

If we add further assumptions beyond these three, we can derive additional conclusions about the ability of the device to produce or understand sentences in the manner of the incorporated grammar. Consider the two extreme assumptions:

4. M produces P -markers strictly “from the top down,” or from trunk to branch, in the tree graph of the P -marker.

5. M produces P -markers strictly “from the bottom up,” or from branch to trunk, in the tree graph of the P -marker.

In accordance with (4), the device M will interpret a rule $A \rightarrow \phi$ of the incorporated grammar as the instruction “rewrite A as ϕ ”—that is to say, as the instruction that, in constructing a derivation, a line of the form $\psi_1 A \psi_2$ can be followed by the line $\psi_1 \phi \psi_2$. Assumption 5 requires the device M to interpret each rule $A \rightarrow \phi$ of the grammar as the instruction “replace ϕ by A ”—that is to say, in constructing an inverted derivation with S as its last line and a terminal string as its first line, a line of the form $\psi_1 \phi \psi_2$ can be followed by the line $\psi_1 A \psi_2$.

From Assumption 4 we can conclude that only a bounded number of successive left-branchings can, in general, be tolerated by M . Thus suppose that M is based on a grammar containing the rule $S \rightarrow SA$. After n applications of this left-branching rule the memory of a device meeting Assumptions 2 and 4 (under the natural interpretation) would have to store n occurrences of A for later rewriting and would thus eventually have to violate Assumption 1. On the other hand, from Assumption 5 we can conclude that only a bounded number of successive right-branchings can in general be tolerated. For example, suppose the underlying grammar contains right-branching rules: $A \rightarrow cA$, $B \rightarrow cB$, $A \rightarrow a$, and $B \rightarrow b$. In this case the device will be presented with strings $c^n a$ or $c^n b$. Now, although Assumption 2 still calls for resolution from left to right, Assumption 5 implies that no node in the P -marker can be replaced until all that it dominates is known, so that resolution must be postponed until the final symbol in the string is received. Thus the device would have to store n occurrences of c for later rewriting and, again, Assumption 1 must eventually be violated. Left-branching causes no difficulty under Assumption 5, of course, just as right-branching causes no difficulty in the case of Assumption 4. Thus Assumptions 4 and 5 impose left-right asymmetries (in opposite ways) on the set of structures that can be accepted or produced by M . Observe that the devices $\Psi(G, m)$, given by the procedure Ψ of Chapter 12, Sec. 4.6, need not meet either of the restrictions in Assumption 4 or 5; in constructing a particular P -marker, they may move up or down or both ways indefinitely often, just as long as self-embedding is restricted.

Assumption 4 might be interpreted as a condition to be met by the speaker; Assumption 5, as a condition to be met by the hearer. (Of course, if we design a model of the speaker to meet Assumption 4 and a model of the hearer to meet Assumption 5 simultaneously, we will severely restrict the possibility of communication between them.) If Assumption 4 described the speaker, we would expect him to have difficulty with left-branching constructions; if Assumption 5 described the listener, we would expect him to have difficulty with right-branching constructions. Neither assumption seems particularly plausible. There is no reason to think that a speaker must always select his major phrase types before the minor subphrases or his word categories before his words (Assumption 4). Similarly, although a listener obviously receives terminal symbols and constructs phrase types, there is no reason to assume that decisions concerning minor phrase types must uniformly precede those concerning major structural features of the sentence. Assumptions 4 and 5 are but two of a large set of possible assumptions that might be considered in specifying models of the user more fully. Thus we might introduce an assumption that there is a bound on the length of the string that must be received before a construction can be uniquely identified by a left-to-right perceptual model—and so on, in many other ways.

There has been some discussion of hypotheses such as Assumptions 4 and 5. For example, Skinner's (1957) proposal that "verbal operant responses" to situations (e.g., the primary nouns, verbs, adjectives) form the raw materials of which sentences are constructed by higher level "autoclitic" responses (grammatical devices, ordering, selecting, etc.) might be loosely interpreted as a variant of Assumption 5, regarded as an assumption about the speaker. Yngve (1960, 1961) has proposed a variant of (4) as an assumption about the speaker; his proposal is explicitly directed toward our present topic and so demands a somewhat fuller discussion.

Yngve describes a process by which a device that contains a grammar rather similar to a context-free grammar produces derivations of utterances, always rewriting the leftmost nonterminal symbol in the last line of the already constructed derivation and postponing any nonterminal symbols to the right of it. Each postponed symbol, therefore, is a promise that must be remembered until the time comes to develop it; as the number of these promises grows, the load on memory also grows. Thus Yngve defines a measure of *depth* in terms of the number of postponed symbols, so that left-branching, self-embedding, and multiple-branching all contribute to depth, whereas right-branching does not. (Note that the depth of postponed symbols and the degree of embedding are quite distinct measures.) Yngve observes that a model so constructed for the speaker

will be able with a limited memory to produce structures that do not exceed a certain depth. He offers the hypothesis that Assumption 4, so interpreted, is a correct characterization of the speaker and that natural languages have developed in such a way as to ease the speaker's task by limiting the necessity for left-branching.

The arguments in support of this hypothesis, however, seem inconclusive. It is difficult to see why any language should be designed for the ease of the speaker rather than the hearer, and Assumption 4 in any form seems totally unmotivated as a requirement for the hearer; on the contrary, the opposite assumption, as we have noted, seems the better motivated of the two. Nor does (4) seem to be a particularly plausible assumption concerning the speaker, for reasons we have already stated. It is possible, of course, to construct sentences that have a great depth and that are quite unintelligible, but they characteristically involve nesting or self-embedding and thus serve merely to show that the speaker and hearer have finite memories—that is to say, they support only the obvious and unquestionable Assumptions 1 and 2, not the additional Assumption 4. In order to support Yngve's hypothesis, we would have to find unintelligible sentences whose difficulty was attributable entirely to left-branching and multiple-branching. Such examples are not readily produced. In order to explain why multiple-branching, which contributes to the measure of depth, does not cause more difficulty, Yngve treats coordinate constructions (e.g., conjunctions) as right-branching, which does not contribute to the number of postponed symbols. But this is perfectly arbitrary; they could just as well be treated as left-branching. The only correct interpretation for such constructions is in terms of multiple-branching from a single node—this is exactly the formal feature that distinguishes true coordinate constructions, with no internal structure, from others. As we have observed in Chapter 11, Sec. 5, such constructions are beyond the limits of systems of rewriting rules altogether. Hence the relative ease with which such sentences as Examples 18 and 20 of Chapter 11 can be understood contradicts not only Assumption 4 but even the underlying Assumption 3, of which 4 is an elaboration.

In short, there seems to be little that we can say about the speaker and the hearer beyond the obvious fact that they are limited finite devices that relate sentences and structural descriptions and that they are subject to the constraint that time is linear. From this, all that we can conclude is that self-embedding (and, more generally, nesting of dependencies) should cause difficulty, as indeed it does. It is also not without interest that self-embedding seems to impose a greater burden than an equivalent amount of nesting without self-embedding. Further speculations are, at the present time, quite unsupported.

2.2 Models Incorporating Transformational Grammars

There are surprising limitations on the amount of short-term memory available for human data processing, although the amount of long-term memory is clearly great (cf. Miller, 1956). This fact suggests that it might be useful to look into the properties of a perceptual model M with two basic components, M_1 and M_2 , operating as follows: M_1 contains a small, short-term memory. It performs computations on an input string x as it is received symbol by symbol and transmits the result of these computations to M_2 . M_2 contains a large long-term memory in which is stored a generative grammar G ; the task of M_2 is to determine the deeper structure of the input string x , using as its information the output transmitted to it by M_1 . (Sentence-analyzing procedures of this sort have been investigated by Matthews, 1961.)

The details of the operation of M_2 would be complicated, of course; probably the best way to get an appreciation of the functions it would have to perform is to consider an example in some detail. Suppose, therefore, that a device M , so constructed, attempts to analyze such sentences as

John is easy to please. (36)

John is eager to please. (37)

To these, M_1 might assign preliminary analyses, as in Fig. 6, in which inessentials are omitted. Clearly, however, this is not the whole story. In order to account for the way in which we understand these sentences, it is necessary for the component M_2 , accepting the analysis shown in Fig. 6 as input, to give as output structural descriptions that indicate that in

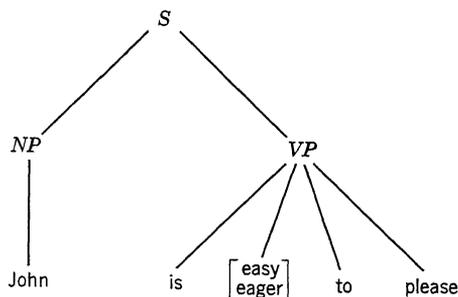


Fig. 6. Preliminary analysis of Sentences 36 and 37.

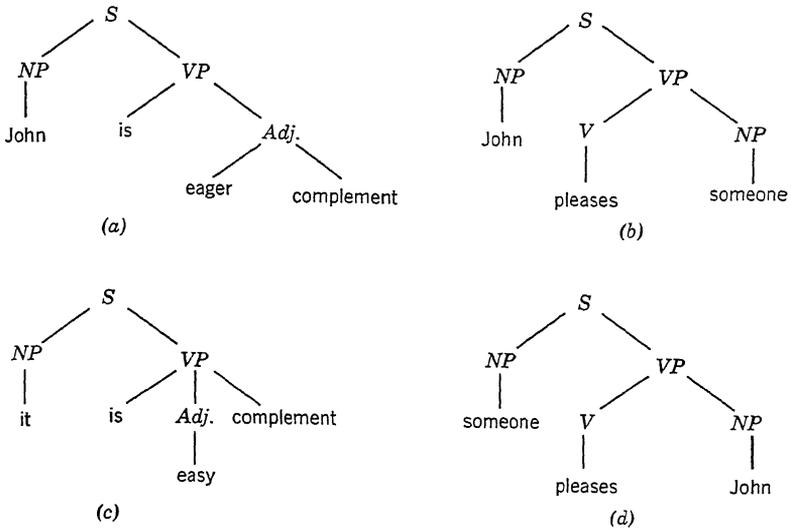


Fig. 7. Some *P*-markers that would be generated by the rewriting rules of the grammar and to which the transformation rules would apply.

Example 36 *John* is the direct object of *please*, whereas in Example 37 it is the logical subject of *please*.

Before we can attempt to provide a description of the device M_2 we must ask how structural information of this deeper kind can be represented. Clearly, it cannot be conveyed in the labeled tree (*P*-marker) associated with the sentence as it stands. No elaboration of the analysis shown in Fig. 6, with more elaborate subcategorization, etc., will remedy the fundamental inability of this form of representation to mirror grammatical relations properly. We are, of course, facing now precisely the kind of difficulty that was discussed in Chapter 11, Sec. 5, and that led to the development of a theory of transformational generative grammar. In a transformational grammar for English the rewriting rules would not be required to provide Examples 36 and 37 directly; the rewriting rules would be limited to the generation of such *P*-markers as those shown in Fig. 7 (where inessentials are again omitted). In addition, the grammar will contain such transformations as

- T_1 : replaces *complement* by "for x to y ," where x is an *NP* and y is a *VP* in the already generated sentence xy ;
- T_2 : deletes the second occurrence of two identical *NP*'s (with whatever is affixed to them);
- T_3 : deletes direct objects of certain verbs;

T_4 : deletes "for someone" in certain contexts;

T_5 : converts a string analyzable as

$$NP - \text{is} - Adj - (\text{for} - NP_1) - \text{to} - V - NP_2$$

to the corresponding string of the form

$$NP_2 - \text{is} - Adj - (\text{for} - NP_1) - \text{to} - V.$$

Each of these can be generalized and put in the form specified in Chapter 11. When appropriately generalized, they are each independently motivated by examples of many other kinds. Note, for example, the range of sentences that are similar in their underlying structural features to Examples 36 and 37; we have such sentences as *John is an easy person to please*, *John is a person who (it) is easy to please*, *this room is not easy to work in (to do decent work in)*, *he is easy to do business with*, *he is not easy to get information from*, *such claims are very easy to be fooled by*, and many others all of which are generated in essentially the same way.

Applying T_1 to the pair of structures in Figs. 7c and 7d, we derive the sentence *It is easy for someone to please John*, with its derived P -marker. Applying T_4 to this, we derive *It is easy to please John*, which is converted to Example 36 by T_5 . Had we applied T_5 without T_4 , we could have derived, for example, *John is easy for us to please* (with *we* chosen in place of *someone* in Fig. 7d—we leave unstated obvious obligatory rules). Applying T_1 to the pair of structures in Figs. 7a and 7b, we derive *John is eager for John to please someone*, which is converted by T_2 to *John is eager to please someone*. Had we applied T_3 to Fig. 7b before applying T_1 , we would, in the same way, have derived Example 37.

At this point we should comment briefly on several features of such an analysis. Notice that *I am eager for you to please*, *you are eager for me to please*, etc., are all well-formed sentences; but *I am eager for me to please*, *you are eager for you to please*, etc., are impossible and are reduced to *I am eager to please*, *you are eager to please* obligatorily by T_2 . This same transformation gives *I expected to come*, *you expected to come*, etc., from *I expected me to come*, *you expected you to come*, which are formed in the same way as *you expected me to come*, *I expected you to come*. Thus this grammar does actually regard *John* in Example 37 as identical with the deleted subject of *please*. Note, in fact, that in the sentence *John expected John to please*, in which T_2 has not applied, the two occurrences of *John* must have different reference. In Example 36, on the other hand, *John* is actually the direct object of *please*, assuming grammatical relations to be preserved under transformation (assuming, in other words, that the P -marker represented in Fig. 7d is part of the structural description of

Example 36). Note, incidentally, that T_5 does not produce such non-sentences as *John is easy to come*, since there is no *NP comes John*, though we have *John is eager to come* by T_1, T_2 . T_5 would not apply to any sentence of the form

$$NP - \text{is} - \text{eager} - (\text{for} - NP_1) - \text{to} - V - NP_2$$

to give

$$NP_2 - \text{is} - \text{eager} - (\text{for} - NP_1) - \text{to} - V$$

(for example, *Bill is eager for us to meet* from *John is eager for us to meet Bill*; *these crooks are eager for us to vote out* from *John is eager for us to vote out these crooks*), since *eager complement*, but not *eager*, is an *Adj* (whereas, *easy*, but not *easy complement*, is an *Adj*). Supporting this analysis is the fact that the general rule that nominalizes sentences of the form *NP - is - Adj* (giving, for example, *John's cleverness* from *John is clever*), converts *John is eager (for us) to come* (which comes from Fig. 7a and *we come* by T_1) to *John's eagerness for us to come*; but it does not convert Example 36 to *John's easiness to please*. Furthermore, the general transformational process that converts phrases of the form

$$\text{the} - \text{Noun} - \text{who (which)} - \text{is} - \text{Adj}$$

to

$$\text{the} - \text{Adj} - \text{Noun}$$

(for example, *the man who is old to the old man*) does convert *a fellow who is easy to please* to *an easy fellow to please* (since *easy* is an *Adj*) but does not convert *a fellow who is eager to please* to *an eager fellow to please* (since *eager* is not, in this case, an *Adj*). In brief, when these rules are stated carefully, we find that a large variety of structures is generated by quite general, simple, and independently motivated rules, whereas other superficially similar structures are correctly excluded. It would not be possible to achieve the same degree of generalization and descriptive adequacy with a grammar that operates in the manner of a rewriting system, assigning just a single *P*-marker to a sentence as its structural description.

Returning now to our main theme, we see that the grammatical relations of *John to please* in Examples 36 and 37 are represented in the intuitively correct way in the structural descriptions provided by a transformational grammar. The structural description of Example 36 consists of the two underlying *P*-markers in Figs. 7c and 7d and the derived *P*-marker in Fig. 6 (as well as a record of the transformational history, i.e., T_1, T_4, T_5). The structural description of Example 37 consists of the underlying *P*-markers in Figs. 7a and 7b and the derived *P*-marker in Fig. 6 (along with the transformational history T_1, T_2, T_3). Thus the structural description

of Example 36 contains the information that *John* in Example 36 is the object of *please* in the underlying *P*-marker of Fig. 7d; and the structural description of Example 37 contains the information that *John* in Example 37 is the subject of *please* in the underlying *P*-marker in Fig. 7b. Note that, when the appropriately generalized form of T_5 applies to *it is easy to do business with John* to yield *John is easy to do business with*, we again have in the underlying *P*-markers a correct account of the grammatical relations in the transform, although in this case the grammatical subject *John* is no longer the object of the verb of the complement, as it is in Example 3b. Notice also that it is the underlying *P*-markers, rather than the derived *P*-marker, that represent the semantically relevant information in this case. In this respect, these examples are quite typical of what is found in more extensive grammars.

These observations suggest that the transformational grammar be stored and utilized only by the component M_2 of the perceptual model. M_1 will take a sentence as input and give us as output a relatively superficial analysis of it (perhaps a derived *P*-marker such as that in Fig. 6). M_2 will utilize the full resources of the transformational grammar to provide a structural description, consisting of a set of *P*-markers and a transformational history, in which deeper grammatical relations and other structural information are represented. The output of $M = (M_1, M_2)$ will be the complete structural description assigned to the input sentence by the grammar that it stores; but the analysis that is provided by the initial, short-term memory component M_1 may be extremely limited.

If the memory limitations on M_1 are severe, we can expect to find that structurally complex sentences are beyond its analytic power even when they lack the property (i.e., repeated self-embedding) that takes them completely beyond the range of any finite device. It might be useful, therefore, to develop measures of various sorts to be correlated with understandability. One rough measure of structural complexity that we might use, along with degree of nesting and self-embedding, is the node-to-terminal-node ratio $N(Q)$ in the *P*-marker Q of the terminal string $t(Q)$. This number measures roughly the amount of computation per input symbol that must be performed by the listener. Hence an increase in $N(Q)$ should cause a correlated difficulty in interpreting $t(Q)$ for a real-time device with a small memory. Clearly $N(Q)$ grows as the amount of branching per node decreases. Thus $N(Q)$ is higher for a binary *P*-marker such as that shown in Fig. 8a than for the *P*-marker in Fig. 8b that represents a coordinate construction with the same number of terminals. Combined with our earlier speculations concerning the perceptual model M , this observation would lead us to suspect that $N(Q)$ should in general be higher for the derived *P*-marker that must be provided by the limited

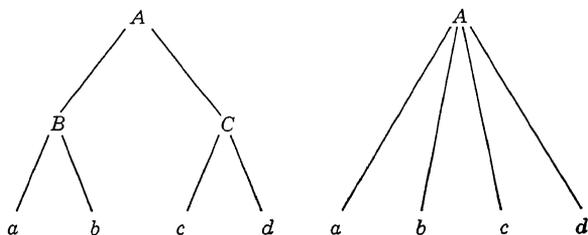


Fig. 8. Illustrating a measure of structural complexity. $N(Q)$ for the P -marker (a) is $7/4$; for (b), $N(Q) = 5/4$.

component M_1 than it would be for underlying P -markers. In other words, the general effect of transformations should be to decrease the total amount of structure in the associated P -marker. This expectation is fully borne out. The underlying P -markers have limited, generally binary branching. But, as we have already observed in Chapter 11 (particularly p. 305), binary branching is not a general characteristic of the derived P -markers associated with actual sentences; in fact, the actual set of derived P -markers is beyond the generative capacity of rewriting systems altogether, since there is no bound on the amount of branching from a single node (that is to say, on the length of a coordinate construction).

The psychological plausibility of a transformational model of the language user would be strengthened, of course, if it could be shown that our performance on tasks requiring an appreciation of the structure of transformed sentences is some function of the nature, number, and complexity of the grammatical transformations involved.

One source of psychological evidence concerns the grammatical transformation that negates an affirmative sentence. It is a well-established fact that people in concept-attainment experiments find it difficult to use negative instances (Smoke, 1933). Hovland and Weiss (1953) established that this difficulty persists even when the amount of information conveyed by the negative instances is carefully equated to the amount conveyed by positive instances. Moreover, Wason (1959, 1961) has shown that the grammatical difference between affirmative and negative English sentences causes more difficulty for subjects than the logical difference between true and false; that is to say, if people are asked to verify or to construct simple sentences (about whether digits in the range 2 to 9 are even or odd), they will take longer and make more errors on the true negative and false negative sentences than on the true affirmative and false affirmative sentences. Thus there is some reason to think that there may be a grammatical explanation for some of the difficulty we have in using negative information; moreover, this speculation has received some support from

Eifermann (1961), who found that negation in Hebrew has a somewhat different effect on thinking than it has in English.

A different approach can be illustrated by sentence-matching tests (Miller, 1962*b*). One study used a set of 18 elementary strings (for example, those formed by taking *Jane*, *Joe*, or *John* as the first constituent, *liked* or *warned* as the second, and *the old woman*, *the small boy*, or *the young man* as the last), along with the corresponding sets of sentences that could be formed from those by passive, negative, or passive-and-negative transformations. These sets were taken two at a time, and subjects were asked to match the sentences in one set with the corresponding sentences in the other. The rate at which they worked was recorded and from that it was possible to obtain an estimate of the time required to perform the necessary transformations. If we assume that these four types of sentence are coordinate and independently learned, then there is little reason to believe that finding correspondences between any two of them will necessarily be more difficult than between any other two. On the other hand, if we assume that the four types of sentence are related to one another by two grammatical transformations (and their inverses), then we would expect some of the tests to be much easier than others. The data supported a transformational position: the negative transformation was performed most rapidly, the more complicated passive transformation took slightly longer, and tests requiring both transformations (kernel to passive-negative or negative to passive) took as much time as the two single transformations did added together. For example, in order to perform the transformations necessary to match such pairs as *Jane didn't warn the small boy* and *The small boy was warned by Jane*, subjects required on the average more than three seconds, under the conditions of the test.

Still another way to explore these matters is to require subjects to memorize a set of sentences having various syntactic structures (J. Mehler, personal communication). Suppose, for example, that a person reads at a rapid but steady rate the following string of eight sentences formed by applying passive, negative, and interrogative transformations: *Has the train hit the car? The passenger hasn't been carried by the airplane. The photograph has been made by the boy. Hasn't the girl worn the jewel? The student hasn't written the essay. The typist has copied the paper. Hasn't the house been bought by the man? Has the discovery been made by the biologist?* When he finishes, he attempts to write down as many as he can recall. Then the list (in scrambled order) is read again, and again he tries to recall, and so on through a series of trials. Under those conditions many syntactic confusions occur, but most of them involve only a single transformational step. It is as if the person recoded the original sentences

into something resembling a kernel string plus some correction terms for the transformations that indicate how to reconstruct the correct sentence when he is called on to recite. During recall he may remember the kernel, but become confused about which transformations to apply.

Preliminary evidence from these and similar studies seems to support the notion that kernel sentences play a central role, not only linguistically, but psychologically as well. It also seems likely that evidence bearing on the psychological reality of transformational grammar will come from careful studies of the genesis of language in infants, but we shall not attempt to survey that possibility here.

It should be obvious that the topics considered in this section have barely been opened for discussion. The problem can clearly profit from abstract study of various kinds of perceptual models that incorporate generative processes as a fundamental component. It would be instructive to study more carefully the kinds of structures that are actually found in natural languages and the formal features of those structures that make understanding and production of speech difficult. In this area the empirical study of language and the formal study of mathematical models may bear directly on questions of immediate psychological interest in what could turn out to be a highly fruitful and stimulating way.

3. TOWARD A THEORY OF COMPLICATED BEHAVIOR

It should by now be apparent that only a complicated organism can exploit the advantages of symbolic organization. Subjectively, we seem to grasp meanings as integrated wholes, yet it is not often that we can express a whole thought by a single sound or a single word. Before they can be communicated, ideas must be analyzed and represented by sequences of symbols. To map the simultaneous complexities of thought into a sequential flow of language requires an organism with considerable power and subtlety to symbolize and process information. These complexities make linguistic theory a difficult subject. But there is an extra reward to be gained from working it through. If we are able to understand something about the nature of human language, the same concepts and methods should help us to understand other kinds of complicated behavior as well.

Let us accept as an instance of complicated behavior any performance in which the behavioral sequence must be internally organized and guided by some hierarchical structure that plays the same role, more or less, as a *P*-marker plays in the organization of a grammatical sentence. It is not

immediately obvious, of course, how we are to decide whether some particular nonlinguistic performance is complicated or simple; one natural criterion might be the ability to interrupt one part of the performance until some other part had been completed.

The necessity for analyzing a complex idea into its component parts has long been obvious. Less obvious, however, is the implication that any complicated activity obliges us to analyze and to postpone some parts while others are being performed. A task, X , say, is analyzed into the parts Y_1 , Y_2 , Y_3 , which should, let us assume, be performed in that order. So Y_1 is singled out for attention while Y_2 and Y_3 are postponed. In order to accomplish Y_1 , however, we find that we must analyze it into Z_1 and Z_2 , and those in turn must be analyzed into still more detailed parts. This general situation can be expressed in various ways—by an outline or by a list structure (Newell, Shaw, & Simon, 1959) or by a tree graph similar to those used to summarize the structural description of individual sentences. While one part of a total enterprise is being accomplished, other parts may remain implicit and still largely unformulated. The ability to remember the postponed parts and to return to them in an appropriate order is necessarily reserved for organisms capable of complicated information processing. Thus the kind of theorizing we have been doing for sentences can easily be generalized to even larger units of behavior. Restricted-infinite automata in general, and PDS systems in particular, seem especially appropriate for the characterization of many different forms of complicated behavior.

The spectrum of complicated behavior extends from the simplest responses at one extreme to our most intricate symbolic processes at the other. In gross terms it is apparent that there is some scale of possibilities between these extremes, but exactly how we should measure it is a difficult problem. If we are willing to borrow from our linguistic analysis, there are several measures already available. We can list them briefly:

INFORMATION AND REDUNDANCY. The variety and stereotypy of the behavior sequences available to an organism are an obvious parameter to estimate in considering the complexity of its behavior (cf. Miller & Frick, 1949; Frick & Miller, 1951).

DEGREE OF SELF-EMBEDDING. This measure assumes a degree of complication that may seldom occur outside the realm of language and language-mediated behaviors. Self-embedding is of such great theoretical significance, however, that we should certainly look for occurrences of it in nonlinguistic contexts.

DEPTH OF POSTPONEMENT. This measure of memory load, proposed by Yngve, may be of particular importance in estimating a person's

capacity to carry out complicated instructions or consciously to devise complicated plans for himself.

STRUCTURAL COMPLEXITY. The ratio of the total number of nodes in the hierarchy to the number of terminal nodes provides an estimate of complexity that, unlike the depth measure, is not asymmetrical toward the future.

TRANSFORMATIONAL COMPLEXITY. A hierarchical organization of behavior to meet some new situation may be constructed by transforming an organization previously developed in some more familiar situation. The number of transformations involved would provide an obvious measure of the complexity of the transfer from the old to the new situation.

These are some of the measures that we can adapt in analogy to the linguistic studies; no doubt many others of a similar nature could be developed.

Clearly, no one can look at a single instance of some performance and immediately assign values to it for any of those measures. As in the case of probability measures, repeated observations under many different conditions are required before a meaningful estimate is available.

Many psychologists, of course, prefer to avoid complicated behavior in their experimental studies; as long as there was no adequate way to cope with it, the experimentalist had little other alternative. Since about 1945, however, this situation has been changing rapidly. From mathematics and logic have come theoretical studies that are increasingly suggestive, and the development of high-speed digital computers has supplied a tool for exploring hypotheses that would have seemed fantastic only a generation ago. Today, for example, it is becoming increasingly common for experimental psychologists to phrase their theories in terms of a computer program for simulating behavior (cf. Chapter 7). Once a theory is expressed in that form, of course, it is perfectly reasonable to try to apply to it some of the indices of complexity.

Miller, Galanter, and Pribram (1960) have discussed the organization of complicated behavior in terms of a hierarchy of *tote* units. A *tote unit* consists of two parts: a *test* to see if some situation matches an internally generated criterion and an *operation* that is intended to reduce any differences between the external situation and some internal criterion. The criterion may derive from a model or hypothesis about what will be perceived or what would constitute a satisfactory state of affairs. The operations can either revise the criterion in the light of new evidence received or they can lead to actions that change the organism's internal and/or external environment. The test and its associated operations are actively linked in a feedback loop to permit iterated adjustments until the criterion

is reached. A tote (test-operate-test-exit) unit is shown in the form of a flow-chart in Fig. 9. A hierarchy of tote units can be created by analyzing the operational phase into a sequence of tote units; then the operational phase of each is analyzed in turn. There should be no implication, however, that the hierarchy must be constructed exclusively from strategy to tactics or exclusively from tactics to strategy—both undoubtedly occur. An example of the kind of structures produced in this way is shown in the flowchart in Fig. 10.

These serial flowcharts are simply the finite automata we considered in Chapter 12, and it is convenient to replace them by oriented graphs (cf.

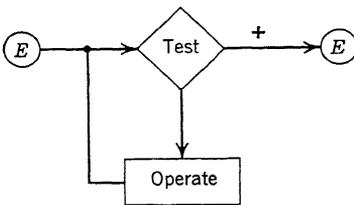


Fig. 9. A simple tote unit.

Karp, 1960). Wherever an initial or terminal element or operation occurs in the flowchart, replace it by a node with one labeled arrow exiting from the node; wherever a test occurs, replace it by a node with two labeled exits. Next, replace every nonbranching sequence of arrows by a single arrow bearing a compound label. The graph corresponding to the flow-chart of Fig. 10 is

shown in Fig. 11. From such oriented graphs as these it is a simple matter to read off the set of triples that define a finite automaton.

A tote hierarchy is just a general form of finite automaton in the sense of Chapter 12. We know from Theorem 2 of Chapter 12 that for any finite automaton there is an equivalent automaton that can be represented by a finite number of finite notations of the form $A_1(A_2, \dots, A_m)^*A_{m+1}$, where the elements A_2, \dots, A_m can themselves be notations of the same form, and so on, until the full hierarchy is represented. For any finite state model that may be proposed, therefore, there is an equivalent model in terms of a (generalized) tote hierarchy.

Since a tote hierarchy is analogous to a program of instructions for a serial computer, it has been referred to as a *plan* that the system is trying to execute. Any postponed parts of the plan constitute the system's *intentions* at any given moment. Viewed in this way, therefore, the finite devices discussed in these chapters are clearly applicable to an even broader range of behavioral processes than language and communication. Some implications of this line of argument for nonlinguistic phenomena have been discussed informally by Miller, Galanter, and Pribram.

A central concern for this type of theory is to understand where new plans come from. Presumably, our richest source of new plans is our old plans, transformed to meet new situations. Although we know little about it, we must have ways to treat plans as objects that can be formed and transformed according to definite rules. The consideration of transformational grammars gives some indication of how we might combine

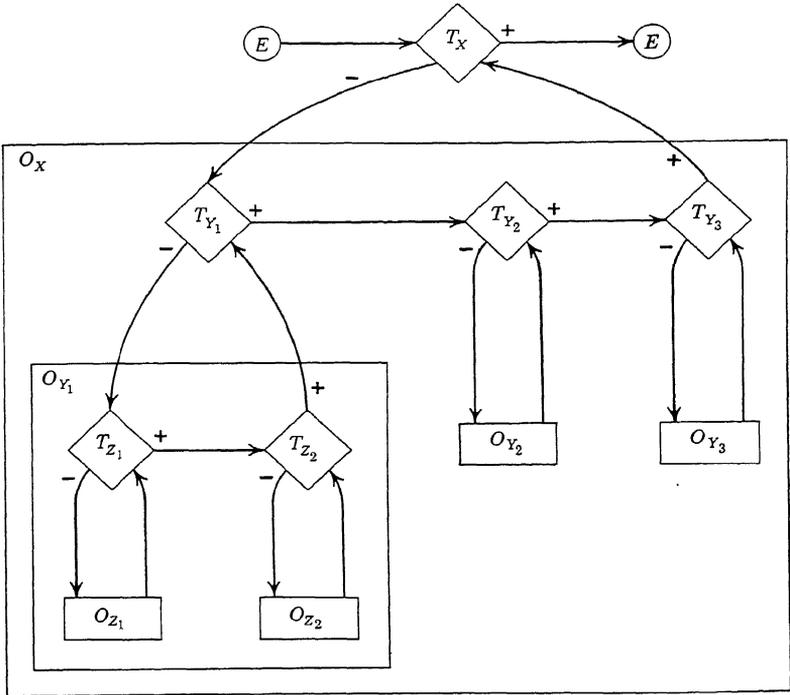


Fig. 10. A hierarchical system of tote units.

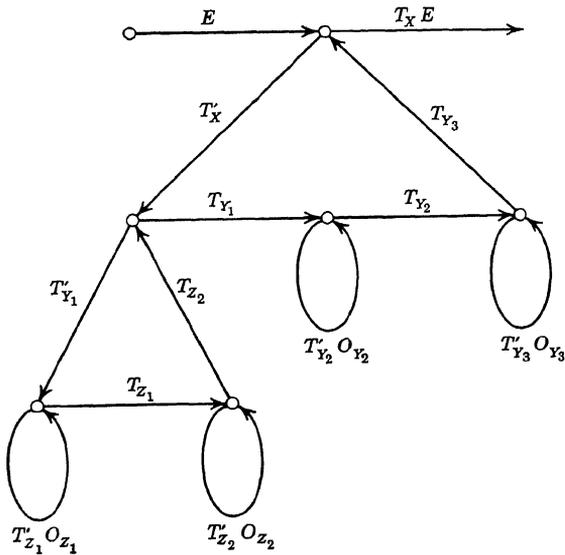


Fig. 11. Graph of flowchart in Fig. 10.

and rearrange plans, which are, of course, so closely analogous to *P*-markers. As in the case of grammatical transformations, the truly productive behavioral transformations are undoubtedly those that combine two or more simpler plans into one. These three chapters make it perfectly plain, however, how difficult it is to formulate a transformational system to achieve the twin goals of empirical adequacy and feasibility of abstract study.

When we ask about the source of our plans, however, we also raise the closely related question of what it might be that stands in the same relation to a plan as a grammar stands to a *P*-marker or as a programming language stands to a particular program. In what form are the rules stored whereby we construct, evaluate, and transform new plans? Probably there are many diverse sets of rules that govern our planning in different enterprises, and only patient observation and analysis of each behavioral system will enable us to describe the rules that govern them.

It is probably no accident that a theory of grammatical structure can be so readily and naturally generalized as a schema for theories of other kinds of complicated human behavior. An organism that is intricate and highly structured enough to perform the operations that we have seen to be involved in linguistic communication does not suddenly lose its intricacy and structure when it turns to nonlinguistic activities. In particular, such an organism can form verbal plans to guide many of its nonverbal acts. The verbal machinery turns out sentences—and, for civilized men, sentences have a compelling power to control both thought and action. Thus the present chapters, even though they have gone well beyond the usual bounds of psychology, raise issues that must be resolved eventually by any satisfactory psychological theory of complicated human behavior.

References

- Attneave, F. *Applications of information theory to psychology*. New York: Holt-Dryden, 1959.
- Burton, N. G., & Licklider, J. C. R. Long-range constraints in the statistical structure of printed English. *Amer. J. Psychol.*, 1955, **68**, 650–653.
- Carnap, R., & Bar-Hillel, Y. *An outline of a theory of semantic information*. Res. Lab. Electronics, Cambridge: Mass. Inst. Tech. Tech. Rept. 247, 1952.
- Chapanis, A. The reconstruction of abbreviated printed messages. *J. exp. Psychol.*, 1954, **48**, 496–510.
- Cherry, C. *On human communication*. New York: Technology Press and Wiley, 1957.
- Chomsky, N. *Logical structure of linguistic theory*. Microfilm. Mass. Inst. Tech. Libraries, 1955.
- Condon, E. V. Statistics of vocabulary. *Science*, 1928, **67**, 300.

- Cronbach, L. J. On the non-rational application of information measures in psychology. In H. Quastler (Ed.), *Information theory in psychology*. Glencoe, Ill.: Free Press, 1955. Pp. 14-26.
- Eiffmann, R. R. Negation: a linguistic variable. *Acta Psychol.*, 1961, **18**, 258-273.
- Estoup, J. B. *Gammes sténographique*. (4th ed.) Paris: 1916.
- Fano, R. M. *The transmission of information*. Res. Lab. Electronics, Cambridge: Mass. Inst. Tech. Tech. Rept. 65, 1949.
- Fano, R. M. *The transmission of information*. New York: Wiley, 1961.
- Feinstein, A. *Foundations of information theory*. New York: McGraw-Hill, 1958.
- Feller, W. *An introduction to probability theory and its applications*. (2nd ed.) New York: Wiley, 1957.
- Fletcher, H. *Speech and hearing in communication*. (2nd ed.). New York: Van Nostrand, 1953.
- Frick, F. C., & Miller, G. A. A statistical description of operant conditioning. *Amer. J. Psychol.*, 1951, **64**, 20-36.
- Frick, F. C., & Sumbly, W. H. Control tower language. *J. acoust. Soc. Amer.*, 1952, **24**, 595-597.
- Fritz, E. L., & Grier, G. W., Jr. Pragmatic communications: A study of information flow in air traffic control. In H. Quastler (Ed.), *Information theory in psychology*. Glencoe, Ill.: Free Press, 1955. Pp. 232-243.
- Garner, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.
- Gnedenko, B. V., & Kolmogorov, A. N. *Limit distributions for sums of independent random variables*. Translated by K. L. Chung. Cambridge, Mass.: Addison-Wesley, 1954.
- Halle, M., & Stevens, K. N. Analysis by synthesis. In *Proc. Seminar on Speech Compression and Production*, AFCRC-TR-59-198, 1959.
- Halle, M., & Stevens, K. N. Speech recognition: A model and a program for research. *IRE Trans. on Inform. Theory*, 1962, **II-8**, 155-159.
- Hardy, G. H., Littlewood, J. E., & Pólya, G. *Inequalities*. (2nd ed.). Cambridge: Cambridge Univ. Press, 1952.
- Hartley, R. V. The transmission of information. *Bell System Tech. J.*, 1928, **17**, 535-550.
- Hovland, C. I., & Weiss, W. Transmission of information concerning concepts through positive and negative instances. *J. exp. Psychol.*, 1953, **45**, 175-182.
- Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proc. IRE*, 1952, **40**, 1098-1101.
- Karp, R. M. A note on the application of graph theory to digital computer programming. *Information and Control*, 1960, **3**, 179-190.
- Katz, J., & Fodor, J. *The structure of a semantic theory*. To appear in *Language*. Reprinted in J. Katz & J. Fodor. *Readings in the philosophy of language*. New York: Prentice-Hall, 1963.
- Khinchin, A. I. *Mathematical foundations of information theory*. Translated by R. A. Silverman and M. D. Friedman. New York: Dover, 1957.
- Luce, R. D. *Individual choice behavior*. New York: Wiley, 1959.
- Luce, R. D. (Ed.) *Developments in mathematical psychology*. Glencoe, Ill.: Free Press, 1960.
- Mandelbrot, B. An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. In W. Jackson, (Ed.), *Proc. symp. on applications of communication theory*. London: Butterworth, 1953.

- Mandelbrot, B. Linguistique statistique macroscopique. In L. Apostel, B. Mandelbrot, & A. Morf. *Logique, langage and théorie de l'information*. Paris: Universitaires de France, 1957. Pp. 1-78.
- Mandelbrot, B. Les lois statistique macroscopiques du comportement. *Psychol. Française*, 1958, 3, 237-249.
- Mandelbrot, B. A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon. *Information and Control*, 1959, 2, 90-99.
- Mandelbrot, B. On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson (Ed.), *Structure of language in its mathematical aspect. Proc. 12th Symp. in App. Math.* Providence, R. I.: American Mathematical Society, 1961. Pp. 190-219.
- Markov, A. A. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin," *Bull acad. imper. Sci., St. Petersburg*, 1913, 7.
- Marschak, J. Remarks on the economics of information. In *Contributions to Scientific Research in Management*. Berkeley, Calif.: Univer. of California Press, 1960. Pp. 79-98.
- Matthews, G. H. Analysis by synthesis of sentences of natural languages. In *Proc. 1st Int. Cong. on Machine Translation of Languages and Applied Language Analysis, 1961*. Teddington, England: National Physical Laboratory, (in press).
- McMillan, B. The basic theorems of information theory. *Ann. math. Stat.*, 1953, 24, 196-219.
- Miller, G. A. *Language and communication*. New York: McGraw-Hill, 1951.
- Miller, G. A. What is information measurement? *Amer. Psychologist*, 1953, 8, 3-11.
- Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 1956, 63, 81-97.
- Miller, G. A. Some effects of intermittent silence. *Amer. J. Psychol.*, 1957, 70, 311-313.
- Miller, G. A. Decision units in the perception of speech. *IRE Trans. Inform. Theory*, 1962, II-8, No. 2, 81-83. (a)
- Miller, G. A. Some psychological studies of grammar. *Amer. Psychologist*, 1962 17, 748-762. (b)
- Miller, G. A., & Frick, F. C. Statistical behavioristics and sequences of responses. *Psychol. Rev.*, 1949, 56, 311-324.
- Miller, G. A., & Friedman, E. A. The reconstruction of mutilated English texts. *Information and Control*, 1957, 1, 38-55.
- Miller, G. A., Galanter, E., & Pribram, K. *Plans and the structure of behavior*. New York: Holt, 1960.
- Miller, G. A., Heise, G. A., & Lichten, W. The intelligibility of speech as a function of the context of the test materials. *J. exp. Psychol.*, 1951, 41, 329-335.
- Miller, G. A., & Newman, E. B. Tests of a statistical explanation of the rank-frequency relation for words in written English. *Amer. J. Psychol.*, 1958, 71, 209-258.
- Miller, G. A., Newman, E. B., & Friedman, E. A. Length-frequency statistics for written English. *Information and Control*, 1958, 1, 370-398.
- Miller, G. A., & Selfridge, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, 63, 176-185.
- Newell, A., Shaw, J. C., & Simon, H. A. Report on a general problem-solving program. In *Information Processing. Proc. International Conference on Information Processing, UNESCO, Paris, June 1959*. Pp. 256-264.
- Newman, E. B. The pattern of vowels and consonants in various languages. *Amer. J. Psychol.*, 1951, 64, 369-379.
- Pareto, V. *Cours d'economie politique*. Paris: 1897.

- Quastler, H. (Ed.). *Information theory in psychology*. Glencoe, Ill.: Free Press, 1955.
- Shannon, C. E. A mathematical theory of communication. *Bell System Tech. J.*, 1948, **27**, 379-423.
- Shannon, C. E. Prediction and entropy of printed English. *Bell Syst. tech. J.*, 1951, **30**, 50-64.
- Skinner, B. F. *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.
- Smoke, K. L. Negative instances in concept learning. *J. exp. Psychol.*, 1933, **16**, 583-588.
- Somers, H. H. The measurement of grammatical constraints. *Language and Speech*, 1961, **4**, 150-156.
- Thorndike, E. L., & Lorge, I. *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teachers College, Columbia University, 1944.
- Toda, M. Information-receiving behavior in man. *Psychol. Rev.*, 1956, **63**, 204-212.
- Wason, P. C. The processing of positive and negative information. *Quart. J. exp. Psychol.*, 1959, **11**, 92-107.
- Wason, P. C. Response to affirmative and negative binary statements. *Brit. J. Psychol.*, 1961, **52**, 133-142.
- Wiener, N. *Cybernetics*. New York: Wiley, 1948.
- Willis, J. C. *Age and area*. Cambridge: Cambridge Univer. Press, 1922.
- Yngve, V. H. A model and an hypothesis for language structure. *Proc. Am. Phil. Soc.*, 1960, **104**, 444-466.
- Yngve, V. H. The depth hypothesis. In R. Jakobson (Ed.), *Structure of language and its mathematical aspect*. *Proc. 12th Symp. in App. Math.* Providence, R. I.: American Mathematical Society, 1961. Pp. 130-138.
- Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, FRS. *Phil. Trans. Roy. Soc. (London)*, 1924, **B 213**, 21-87.
- Yule, G. U. *The statistical study of literary vocabulary*. London: Cambridge Univer. Press, 1944.
- Ziff, P. *Semantic analysis*. Ithaca: Cornell Univ. Press, 1960.
- Zipf, G. K. *The psychobiology of language*. Boston: Houghton-Mifflin, 1935.
- Zipf, G. K. *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley, 1949.

I 4

*Mathematical Models of
Social Interaction*

Anatol Rapoport
University of Michigan

Contents

| | |
|---|-----|
| 1. Interaction in Large Well-Mixed Populations | 497 |
| 1.1. The general equation, | 498 |
| 1.2. The linear model, | 499 |
| 1.3. The logistic model, | 504 |
| 1.4. Time-dependent contagion, | 505 |
| 1.5. Contagion with diffusion, | 508 |
| 1.6. Nonconservative, nonlinear interaction models, | 509 |
| 2. Statistical Aspects of Net Structures | 512 |
| 2.1. The random net, | 513 |
| 2.2. Biased nets, | 515 |
| 2.3. Application of the overlapping clique model to an information-spread process, | 519 |
| 2.4. Application of the biased net model to a large sociogram, | 522 |
| 3. Structure of Small Groups | 529 |
| 3.1. Descriptive theory of small group structures, | 531 |
| 3.2. The detection of cliques and other structural characteristics of small groups, | 536 |
| 3.3. The theory of structural balance, | 539 |
| 3.4. Dominance structures, | 541 |
| 4. Psychoeconomics | 546 |
| 4.1. A mathematical model of parasitism and symbiosis, | 546 |
| 4.2. Bargaining, | 549 |
| 4.3. Bilateral monopoly, | 551 |
| 4.4. Formal experimental games, | 556 |
| 5. Group Dynamics | 562 |
| 5.1. A "classical" model of group dynamics, | 563 |
| 5.2. A semiquantitative model, | 565 |
| 5.3. Markov chain models, | 567 |
| References | 576 |

Mathematical Models of Social Interaction

In order to apply mathematical methods to the study of social interactions, it is obviously necessary to single out entities among which specifiable functional relations are assumed to exist. Attempts to do so have proceeded along two distinct paths.

One path parallels the methodological road of mathematical physics, where attention is focused on numerically measurable quantities and their rates of change. The mathematical tools appropriate to this method are those of classical analysis. The area to which such methods seem pertinent is that of large-scale social phenomena. When dealing with large masses of entities, it is natural to disregard the determinants governing the behavior of the individual elements in favor of gross statistically determined effects. Such gross effects are represented by functional relations among a few essentially continuous variables and usually their time derivatives, typically as differential equations.

The other path, departing from the methods of classical analysis, is directed toward the consideration of sets of discrete entities and structural relations among them. This is the path followed by those interested in interactions among a small number of individuals, especially those based on relations in which the individuals stand to each other, or in interactions based on decisions governed by complex logical considerations. In such situations rather detailed structural descriptions are required. Here some of the "modern" developments of mathematics play an important part. Among examples of the tools used in this approach are the theory of linear graphs, in which binary relations instead of numerical variables are fundamental; the associated matrix representations of the relational structures; set theory, in which the fundamental entities are subsets of a given set, instead of the individual elements composing it; stochastic processes, in which probability distributions instead of just probabilistically determined expected values are the objects of attention; and the mathematical theory of games, which studies in complete detail the logical structure of conflict situations.

In this chapter I have selected a number of developments in mathematical theories of social interaction of both types, which seem to me either (1) typical, (2) interesting, or (3) representative of an attempt to link a mathematical theory to observation or experiment.

Typical methods are worth studying because they indicate some unifying theoretical principle. Therefore, regardless of whether the corresponding mathematical models have been found to be applicable in some specific instance, there is reason to suppose that similar models will be applicable in some situations sooner or later. In these models a great many idealized situations appear logically isomorphic. "Interesting" models have been included for their inspirational value. The inclusion of models in which some link has already been made between theory and observation corroborates, at least partly, the feasibility of using mathematical methods in constructing a science or social interaction.

Our models can also be classified in a triple dichotomy, namely (1) static versus dynamic, (2) pertaining to large-scale versus small-scale social events, and (3) deterministic versus stochastic. We have confined ourselves to those theoretical developments that can be treated in the language of undergraduate mathematics, not going beyond the elements of probability theory and elementary differential equations. The mathematics of game theory, being a vast field of research in its own right, has not been included, although a mathematical treatment of some instructive gamelike situations has been presented where the underlying principles seemed central to our topic.

Finally, attempts have been made, wherever possible, to point out the links that exist between the various approaches. For example, the classical mathematical approach to social interaction via systems of differential equations leads to the consideration of the phase space (cf. Sec. 1.6) which, in turn, leads to questions of stability of certain steady states. These questions are seen to have a relation to certain game-theoretical questions. Thus a transition between the distinctly "classical" and the characteristically "modern" approaches can be discerned. Nor are the distinctions between the "dichotomies" as sharp as they appear on being named. The abstract mathematical model does not distinguish between the replication of an event in "space" or in "time," and so the same framework may at times fit a large population of persons or a large population of responses in a small group (cf. Secs. 1.2 and 5.1). The distinction between the large-scale static sociometric model and the large-scale dynamic contagion process (cf. Secs. 2.3 and 2.4) is likewise obscured by the similarity of the recursive formulas used in their respective treatments. These fusions are not surprising in view of the ubiquitous presence of logical isomorphisms among the conceptual models used in our day. Whether this frequent occurrence of similar formulations bespeaks an underlying logical similarity of events or a comparative paucity of ideas remains for future generations to decide.

1. INTERACTION IN LARGE WELL-MIXED POPULATIONS

Intuitive observations and more systematic but hitherto, for the most part, purely empirical studies of mass behavior indicate that information and behavioral patterns often spread through populations by a contagion-like process. The occasional explosive spreads of rumors, fads, and panics attest to the underlying similarity between social diffusion and other diffusion and chain-reaction processes, such as epidemics, the spread of solvents through solutes, crystallization, dissemination of genes through an interbreeding population, etc. Accordingly, the mathematical behavioral scientist is motivated to seek mathematical models that can be supposed to underlie whole classes of processes of diverse content but of similar mathematical type.

In all cases the model must postulate a *population* and a set of *states* in which each member of the population may find himself. If the number of individuals is large, the fraction or density of individuals in a given state can be taken as a continuous variable. In some situations passing from one state to another means suffering an increment or a decrement of some quantity; for example, biomass or displacement in space or assets. If this quantity is also continuous, the states themselves form a continuum. Otherwise there is only a finite (perhaps denumerably infinite) number of states—in the simplest case, just two.

The number of individuals in the population may be constant or not. If it is not, we account for “sources” and “sinks” in the population, which allow increments (or decrements) of individuals in a certain state without compensating decrements (or increments) of individuals in other states.

The states may be reversible or irreversible. For example, if the interaction is contagion, in which a disease (or a piece of information) passes from one individual to another, the passage of individuals from the non-infected (or not-knowing) to the infected (or knowing) state may be assumed irreversible if the individuals are not expected to recover (or forget) during the time under consideration. If recovery without immunity does occur or if the contagion results in the spread of an attitude or a form of behavior which can also be abandoned, we are dealing with a reversible process. The analogy with chemical reactions is obvious.

If there are more than two states, the concept of reversibility has a more general analogue in the concept of two-way connectedness among sets of states; that is, the counterpart of reversibility in a multistate situation is the possibility of passing from any state to any other state (in general via other

states). The analogue of irreversibility is the existence of a group of states into which it is possible to pass but from which it is impossible to escape, although it may be possible to pass from state to state among them. If there are subsets of states that are not connected to each other in either direction, we have essentially several independent processes, which can be treated separately. In this case there is no need to consider the entire process within the framework of a single system.

As an example of a multivariate contagion system, consider a disease with time-limited infectiousness, in which the following states are distinguished; uninfected, infected and contagious, infected noncontagious, recovered without immunity, recovered with immunity, dead. Some of these may be "absorbing states," in which the individuals once having entered will persist for the duration of the process, for example, "dead" or possibly "recovered with immunity." Therefore this process contains some irreversible "reactions."

To construct a general model of a contagion process, it is necessary to list all the relevant states in which the members of the population may be and also to indicate the transition probabilities from one state to another. The event contributing to the probability of such a transition, typical for a contagion process, is contact between two individuals as a result of which one or both individuals pass into another state. However, it is possible to imagine also "spontaneous" changes of state, for example, from one stage of a disease to the next. Also when two individuals come into contact this may contribute to an increment of a state to which neither of the individuals belongs.

Thus Richardson (1948), in his discussion of war moods, differentiates among several "psychological states" associated with people in peace and war times, such as "friendly," "hostile," "war-weary," "dead," and combinations of some of these. On the battlefield, contacts between two "hostile" individuals contribute to an increment of dead individuals, an example of contact between individuals in one state contributing to an increment of individuals in another. Likewise, we can imagine various degrees of two conflicting political opinions as the states. It is conceivable that contacts between two individuals of mild but opposite opinions contribute to increments of individuals with stronger opposite opinions because of mutual irritation or, on the contrary, to increments of individuals with intermediate opinions because of mutual influence.

1.1 The General Equation

To account for a social interaction process of n states, in which time rates of increments to each state depend on (1) independent sources or

sinks, (2) spontaneous changes from state to state, and (3) changes of state occasioned by contacts, our model would have to be a system of nonhomogeneous, first-order, second-degree differential equations of the following type:

$$\frac{dx_i}{dt} = \sum_{j=1}^n a_{ij}x_j + \sum_{k=1}^n \sum_{j=1}^n b_{jk}^{(i)}x_jx_k + c_i, \quad (i = 1, 2, \dots, n). \quad (1)$$

Here x_i represents the number (or fraction or density) of individuals in the i th state. The a_{ij} ($i \neq j$) represent the rates of net absolute flow into or from the i th state from or to other states (due to concentrations of other states); a_{ii} represents the net reproduction (or dissipation) rate of the individuals in the i th state; $b_{jk}^{(i)}$ represents the rates of conversion to or from the i th state due to contacts between pairs of individuals; and c_i represents the sources or sinks.

Note that the increments due to contacts, as given by Eq. 1, depend only on the total numbers (or concentrations) of individuals in the different states; that is to say, the probability of contact between any two individuals from a pair of specified states is assumed to be the same for any pair of individuals in those states. This is the assumption of well-mixedness. Obviously this assumption cannot be made if the mobility of the population is restricted. In a real contagion, for example, the focus of contagion is at least temporarily geographically circumscribed so that only those uninfected who are near the focus can be expected to become infected at that time. Hence the probability of new infections near the focus will depend on the concentration of individuals near the focus of infection and not on the over-all concentration. These complications are deliberately bypassed when the assumption of well-mixedness is made. In our discussion of contagion models we shall for the most part assume well-mixedness. Later we shall drop this assumption, and this will carry us to the consideration of some structural properties of social space (cf. Sec. 2.2). For the present we shall examine some important special cases of Eq. 1, which underlie various proposed models of social interaction.

1.2 The Linear Model

If all the $b_{jk}^{(i)}$ in the system described by Eq. 1 are zero, the system reduces to a linear one:

$$\frac{dx_i}{dt} = \sum_{j=1}^n a_{ij}x_j + c_i, \quad (i = 1, 2, \dots, n). \quad (2)$$

The general solution of such systems is known. The special cases, which result when certain restraints are imposed on the coefficients, can be described in qualitative terms. For example, under certain conditions,

some of the x_i will be periodic (oscillatory) functions of time. These conditions are usually too special to be of interest in sociological applications. If the system is nonhomogeneous and nonsingular (i.e., if not all the c_i are zero and the determinant of the coefficients a_{ij} does not vanish), setting dx_i/dt equal to zero and solving the resulting system of linear algebraic equations yields an equilibrium point in the n -dimensional space, $x_i = x_i^*$ ($i = 1, 2, \dots, n$). An important question then arises concerning the stability of the equilibrium. If it is stable, then an accidental fluctuation from it in the values of the x_i tends to be "corrected," that is, the system returns to the equilibrium state. If the equilibrium is unstable, an accidental increment in some of the x_i will tend to be magnified, carrying some of the variables to infinity of either sign (or to zero if only positive values are meaningful).

If there are only two variables, Eq. 2 reduces to

$$\begin{aligned}\frac{dx}{dt} &= a_{11}x + a_{12}y + c_1, \\ \frac{dy}{dt} &= a_{21}x + a_{22}y + c_2.\end{aligned}\tag{3}$$

This model has been used by Richardson (1939) to represent an idealized arms race between two states or alliances and by Rashevsky (1939) to represent the simplest case of mass behavior based on mutual imitation.

In Richardson's model x and y represent, respectively, armament expenditures of two rival states. He assumes that increases in the armament expenditures of each state are stimulated by the armament expenditures of the rival. Hence $a_{12} > 0$, $a_{21} > 0$. He further assumes that the rate of increase of armament expenditures is *inhibited* by the level of one's own expenditures (as the burden increases). Hence $a_{11} < 0$, $a_{22} < 0$. Finally, the constant terms represent positive or negative stimulation to armament expenditures independent of the expenditure levels. These are the "grievances," if positive, or reservoirs of "good will" if negative. We can therefore write

$$\begin{aligned}\frac{dx}{dt} &= -ax + my + g, \\ \frac{dy}{dt} &= nx - by + h.\end{aligned}\tag{4}$$

Here a , b , m , and n are positive. The equilibrium point (assuming $ab - mn \neq 0$) is obtained by setting $dx/dt = dy/dt = 0$ and solving the resulting algebraic equations. The position of the equilibrium therefore depends on all the coefficients. The stability of the equilibrium, on the other hand, depends only on the sign of $ab - mn$. It is easy to show that

the equilibrium is stable if and only if $ab - mn > 0$, that is, if the product of the self-restraint coefficients is greater than that of the mutual-stimulation coefficients. If the equilibrium is unstable, the armament expenditures will either increase without bound (a runaway arms race or a war in Richardson's interpretation) or, if the expenditures are sufficiently low initially, tend to be reduced still further until complete disarmament is achieved.

Obviously the model is much too simple-minded to be of use in the analysis of actual international behavior. Still Richardson ventured to apply it to the description of some arms races, notably to that preceding World War I (1909-1914). Giving the coefficients certain values and solving the system of differential equations, he obtained the time course of the combined armament expenditures of the two rival camps, which fitted the data very well. The coefficients were such that the system was inherently unstable. Hence its fate was determined by the initial conditions. In Richardson's interpretation these initial conditions had to do with the difference between armament expenditures and the volume of trade between the two blocks of states. It appears that, had the volume of trade been greater by £5 million or the level of armament expenditures correspondingly lower, the "reaction" would have gone in the opposite direction, that is, toward disarmament and increasing cooperation (trade).

These conclusions are not easy to take seriously. Nor is the agreement between theoretically derived and observed armament expenditures impressive, in view of the small number of points fitted by the equations. Still the approach is noteworthy as an early example of a method for dealing with large-scale social interactions, which may, under certain conditions, find application. Moreover, the method has unquestionable heuristic value in that it serves as a framework in which more sophisticated approaches can be developed.

Rashevsky's model of mass behavior (1939) is based on influences mutually exerted by two classes of individuals, X and Y , representing, respectively, two different patterns of behavior or attitudes R_1 and R_2 , for example, allegiance to two different political parties. In each class there is a fixed number of "actives" (x_0, y_0) who are immune to change. The remaining individuals (x, y) are "passives," subject to influence by both the actives and the passives. Accordingly, if N is the total population, $x + y = \theta$, a constant, and $x_0 + y_0 = N - \theta$, Rashevsky's linear model becomes

$$\begin{aligned}\frac{dx}{dt} &= ax - my + g, \\ \frac{dy}{dt} &= -nx + by + h,\end{aligned}\tag{5}$$

analogous to Richardson's except that the signs of a , b , m , and n are reversed. The constants g and h , as in Richardson's model, can have either sign. They represent the net (constant) influence of the "actives." Substituting $\theta - x$ for y , we get

$$\frac{dx}{dt} = (a + m)x - m\theta + g, \quad (6)$$

whose solution in terms of the initial condition $x(0)$ is

$$x(t) = \frac{[(a + m)x(0) - r]e^{(a+m)t} + r}{a + m}, \quad (7)$$

where $r = m\theta - g$.

The fate of x , therefore, depends crucially on $x(0)$. If $x(0) > r/(a + m)$, $x(t)$ will increase until the whole population will turn to R_1 . If the inequality is reversed, the opposite will happen. The equilibrium at $x = r/(a + m)$; $y = \theta - x = [(a + m)\theta - r]/(a + m)$ is unstable.

In a somewhat more involved model Rashevsky (1951) assumed that each individual possesses some inherent tendency to behave one way or the other. The magnitude of this tendency is denoted by a quantity ϕ , which is positive if the individual prefers R_1 and negative, otherwise. We have, then a distribution of ϕ , $N(\phi)$ in the population, so that $N(\phi) d\phi$ denotes the number of individuals characterized by the value of ϕ between ϕ and $\phi + d\phi$. Rashevsky assumed that $N(\phi)$ is Laplacian, that is,

$$N(\phi) = \frac{1}{2} N_0 \sigma e^{-\sigma|\phi|}. \quad (8)$$

He assumed also that ϕ fluctuates randomly in an individual and that the time distribution of its value is also Laplacian but with a different dispersion constant k instead of σ .

The probability that an individual will perform R_1 at a given time depends on the magnitude of ϕ at that time and also on a magnitude of another "propensity," ψ , contributed by the tendency to imitate others. Specifically, assuming $\psi > 0$ (i.e., the net imitation influence is toward R_1) the probability that R_1 will be performed is given by

$$\begin{aligned} p_1(\phi, \psi) &= 1 - \frac{1}{2} e^{-k(\phi+\psi)} & \text{if } \phi > -\psi, \\ p_1(\phi, \psi) &= \frac{1}{2} e^{k(\phi+\psi)} & \text{if } \phi < -\psi. \end{aligned} \quad (9)$$

The expressions for $\psi < 0$ are analogous. In the remaining discussion ψ is assumed to be positive.

The total numbers of individuals X and Y performing R_1 and R_2 , respectively, at a given moment are

$$\begin{aligned} X &= \int_{-\infty}^{\infty} p_1(\phi, \psi) N(\phi) d\phi, \\ Y &= \int_{-\infty}^{\infty} [1 - p_1(\phi, \psi)] N(\phi) d\phi. \end{aligned} \quad (10)$$

It remains to postulate the equation that determines ψ as a function of X and Y . Rashevsky takes this to be

$$\frac{d\psi}{dt} = A(X - Y) - a\psi, \quad (11)$$

where A and a are constants. In other words, the increase of ψ is enhanced by the excess of individuals performing R_1 , and ψ also "decays" proportionately to its own magnitude.¹

Combining Eqs. 10 and 11, we obtain

$$\frac{d\psi}{dt} = AN_0 \left(1 + \frac{k^2}{\sigma^2 - k^2} e^{-\sigma\psi} - \frac{\sigma^2}{\sigma^2 - k^2} e^{-k\psi} \right) - a\psi. \quad (12)$$

Although the equation is nonlinear, its variables can be separated and so an explicit solution can be obtained for t as a quadrature in terms of a function of ψ . From this solution X and Y can be obtained as functions of time.²

In view of the practical impossibility of making the sort of observations required to check this theory, the explicit dynamic solution is of little interest. However, certain general qualitative conclusions are suggestive. For example, Eq. 12 implies an equilibrium at $\psi = 0$. This equilibrium is stable if and only if

$$\frac{AN_0k\sigma}{\sigma + k} > a. \quad (13)$$

Now the expression $k\sigma/(\sigma + k)$ is the reciprocal of $1/\sigma + 1/k$. Hence $k\sigma/(k + \sigma)$ is greater, the greater the sum of the reciprocals of the dispersions, which refer, respectively, to the nonhomogeneity of the population with respect to the inherent preference and to the nonstereotypy of individuals characterized by a certain preference intensity in performing the preferred act. Combining these interpretations, we have the following qualitative result. The more homogeneous the population and the more stereotyped the behavior of its members, the more likely the instability at

¹ This form of equation has been used extensively by Rashevsky and his co-workers to describe the rate of increase of excitation produced by an external stimulus. A positive contribution is assumed to be proportional to the magnitude of the stimulus (in this case the size of the majority performing R_1), whereas a negative contribution results in the dissipation of the excitation at a rate proportional to its own magnitude.

² Although Eq. 12, being nonlinear, formally excludes the model just described from the class of linear models, we have included it as a variant in view of the linear dependence of $d\psi/dt$ on X and Y (cf. Eq. 11). Under nonlinear models we have understood those in which increments to subpopulations in the various states depend on *products* of the numbers of individuals in pairs of states, that is, presumably on the frequency of contacts. These models are treated in Secs. 1.3 to 1.6.

$\psi = 0$ (equal frequency of acts of both kinds), hence the easier it is to swing the population to the predominant performance of one or the other act.

Interpreting the quantity a/AN_0 , we find that it is directly proportional to the decay constant a and inversely to the imitation propensity A and to the absolute size of the population N_0 . The corresponding result with reference to a and A is obvious. The interesting result is with reference to N_0 , namely, the larger the population, the more likely it is to be swung to the predominance of acts of the one or the other kind. This is the mob effect.

For further elaborations and generalizations of the model, in particular involving asymmetrical distributions of preferences, the interested reader is referred to Rashevsky (1951, 1957).

1.3 The Logistic Model

The previous model departs from linearity but not in the fundamental sense of the general contagion equation (1). The essential feature of this equation is that the variables representing fractions of the population in the several states appear in the second degree. This means that frequencies of *contacts* among the members of the subpopulations contribute to their rates of change. To use a chemical analogy, the foregoing models are tantamount to assumptions that the various "substances" (subpopulations) are produced from all-pervading substrates. The contagion assumptions, on the other hand, imply that substances are produced or destroyed only in interactions with each other. The general Eq. 1, of course, combines both assumptions.

The simplest special case of Eq. 1 involving second-degree terms is the equation of simple contagion, the so-called logistic equation:

$$\frac{dx}{dt} = Bx(y - \theta'); \quad x + y = 1. \quad (14)$$

Here x is the fraction of the infected, y the fraction of the uninfected, and θ' the fraction of the permanently immune. The rate of increment is proportional to the frequency of contacts between the infected and the susceptible (nonimmune) uninfected. The solution of Eq. 14 is given by

$$x(t) = \frac{x_0 \theta e^{B\theta t}}{\theta - x_0(1 - e^{B\theta t})}, \quad (15)$$

where $\theta = 1 - \theta'$.

If x_0 , the initial fraction of infected, is small but finite, the time course is a sigmoid curve tending to θ ; that is, eventually all except the immune will be infected.

1.4 Time-Dependent Contagion

The assumptions underlying the logistic model are quite strong. The population is assumed to be well mixed, and all the infected are assumed to remain infected and infectious. For the time being we shall keep the assumption of well-mixedness but relax the other. The infectiousness of the infected will now be assumed to be a function of time, namely, both of the over-all duration of the process and of the time elapsed since the particular individual became infected. The first dependency reflects the changing "potency" of the process in time (e.g., the virility of the infecting organism); the second dependency reflects the well-known variation of infectiousness during the course of a disease and possibly the removal of infectious individuals by recovery, isolation, or death. The same considerations apply to the spread of information. For example, the newsworthiness of an item of information may be a function of its age, and the tendency to transmit it may be a function of how long ago it was received.

Let t represent time measured from the start of the process and τ the time measured from the moment of infection of each individual. We shall call $p(t, \tau)$ the probability that on contact at time t an individual infected at time $t - \tau$ will transmit the infection to an uninfected individual. If no one is immune, the contagion equation now assumes the following more general form:

$$\frac{dx}{dt} = A(1 - x) \left[x_0 p(t, t) + \int_0^t \frac{dx}{d\lambda} p(t, t - \lambda) d\lambda \right], \tag{16}$$

where x is the fraction of the infected. If $p(t, \tau)$ is a constant, Eq. 16 reduces to a simple logistic. Two other special cases are of interest, namely, (1) when p is a function of t alone and (2) when p is a function of τ alone. The first case can be solved in general. The solution is given by

$$x(t) = \frac{C \exp \left[A \int_0^t p(\xi) d\xi \right]}{1 + C \exp \left[A \int_0^t p(\xi) d\xi \right]}, \tag{17}$$

where $x_0 = C/(1 + C)$. Equation 17 has the same form as Eq. 15, except that the exponent $B\theta t$ of Eq. 15 now appears as an integral. If

$\int_0^\infty p(\xi) d\xi = b$, that is, finite, the ultimate fraction of the infected will be

$$x(\infty) = \frac{Ce^{Ab}}{1 + Ce^{Ab}}. \quad (18)$$

If, on the other hand, the integral diverges, the ultimate fraction will be unity, that is, everyone will succumb.

If $p(t, \tau)$ depends on τ alone, the general solution is obtainable in closed form in some special cases. The case in which $p(\tau) = e^{-k\tau}$ is of interest because it is formally identical to the case in which the infected individuals are removed from the population at random at a constant rate per infected individual. In that case Eq. 16 reduces to

$$\frac{dx}{dt} = A(1-x)e^{-kt} \left[x_0 + \int_0^t \frac{dx}{d\lambda} e^{k\lambda} d\lambda \right]. \quad (19)$$

This leads, after appropriate manipulations, to

$$\frac{dx}{(1-x)\{k \log [(1-x)/(1-x_0)] - Ax\}} = dt. \quad (20)$$

The solution gives t as a quadrature in x :

$$t = \int_{x_0}^x \frac{d\xi}{\sqrt{(1-\xi)\{k \log [(1-\xi)/(1-\xi_0)] + A\xi\}}}. \quad (21)$$

Again we are interested in the value of $x(\infty)$ as a function of the parameters x_0 , A , and k . Clearly, $x(\infty)$ is the smaller root of the denominator of the integrand in Eq. 21. By the nature of the problem this is not greater than unity. It is therefore the root of the transcendental equation

$$k \log \frac{1-x}{1-x_0} + Ax = 0. \quad (22)$$

Taking exponentials of both sides of Eq. 22, we find that the asymptotic value $x^* = x(\infty)$ must satisfy the equation

$$x^* = 1 - (1-x_0)e^{-Ax^*/k}. \quad (23)$$

If the initial number of the infected forms an infinitesimal fraction of the population, we may set $x_0 = 0$ and obtain the equation derived by Solomonoff and Rapoport (1951) and independently by Landau (1952) for the "connectivity" of a random net with axon density a . The meanings of this parameter, of the random net model, and of its generalizations are discussed in Sec. 2.1.

The importance of Eq. 23 is that it holds *no matter what the form of* $p(\tau)$ is, as long as the function governing the probability of transmission

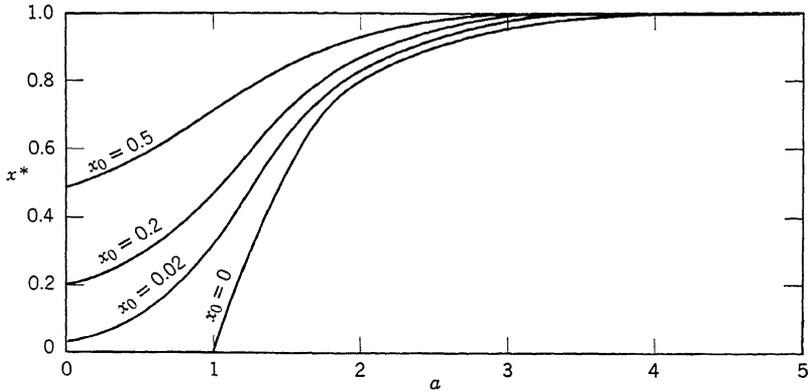


Fig. 1. Ultimate fraction of infected individuals (x^*) as a function of the average number of individuals ever infected by each individual (a) for different values of the original fraction of the infected (x_0).

depends on τ alone and not on t , that is, it depends only on how long an infected individual has been infected and not on how long the epidemic has been going on. In that case the total fraction of the population that will have succumbed depends only on the average number of individuals infected by each infected individual *ever* and not on the times when he has transmitted his infection. The dependence of x^* on $a = A/k$ and on x_0 is shown in Fig. 1. The curve corresponding to $x_0 = 0$ must, of course, be interpreted as the limiting curve of a sequence in which x_0 tends to zero. (Obviously, there will be no epidemic if no one has ever been infected.) It is interesting to observe that if $a = 2$, that is, if on the average each infected individual can infect two others, then for an arbitrarily small x_0 eventually about 80% of the population will succumb.

Equation 23 is also closely related to the result obtained by D. G. Kendall for an epidemic spreading over a geographical area with constant relative rate of removal of the infected (Bailey, 1957). The ultimate fraction *removed*, γ , satisfies the equation

$$\gamma = 1 - e^{-(\sigma/\rho)\gamma}. \quad (24)$$

Here σ is an infection rate constant dependent on the population density and ρ is the ratio of the removal rate (per infected individual) to the infection rate (per contact per individual). Equation 24 is formally identical to Eq. 23 if x_0 in (23) is neglected. The equation has a root at zero and one other positive root. It implies the following conclusions pertinent to Kendall's theory of pandemics:

1. A pandemic will occur if and only if $\rho < \sigma$.
2. If a pandemic does occur, a fraction at least as great as γ (which depends on σ/ρ) will be ultimately infected arbitrarily far from the focal point of the epidemic.

1.5 Contagion with Diffusion

In all the interaction models so far (except Kendall's pandemic, whose derivation we have not discussed) a "well-mixed" population was always assumed. If this well-mixedness assumption is dropped, the interaction problem becomes much more difficult. For example, in contagion models, we must take into consideration, in addition to the spread of state due to contacts between individuals, the *diffusion* of the infected individuals through the population. The assumption of well-mixedness means that there is no restraint on mobility, hence that the diffusion is instantaneous. It is as if the infected individuals became so rapidly mixed throughout the population that their density was always constant everywhere. This is the justification for assuming the transmission to be equally probable between any pair of individuals. In the foregoing generalization (cf. Sec. 1.4) we introduced an additional probability, namely, that the state will be transmitted *if contact does occur*, this probability being dependent on time but not on space variables. Abandoning well-mixedness, we introduce a dependence on space variables. We can write in general, assuming a three-dimensional diffusion space,

$$\frac{\partial c}{\partial t} = D \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2} \right) + Q(c, x, y, z). \quad (25)$$

Here c is the concentration of infected individuals, the first term on the right governs the diffusion of these individuals throughout space, and the second term governs the contagion process. In particular, if the contagion probability depends only on the concentration of the infected explicitly and in the elementary way of the logistic process, Eq. 25 becomes

$$\frac{\partial c}{\partial t} = D \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2} \right) + \alpha c - \beta c^2, \quad (26)$$

where c is, of course, a function of all four independent variables, and α and β are constants.

Landahl (1957) attacked this equation by approximation methods developed previously in problems involving nonconservative diffusion. The interested reader is referred to his paper for further development of this topic.

1.6 Nonconservative, Nonlinear Interaction Models

All of the interaction models so far considered except Richardson's were conservative in the sense that either the total population was constant or, in the case of an infinite population spread over an infinite area, the density of the population was constant in time. These assumptions imply that increases in numbers or in densities of individuals in some states are always compensated for by corresponding decreases in other states. These assumptions cannot be made, therefore, where sources or sinks, for example, birth and death rates, are an integral part of the interaction process.

As the simplest example of a nonlinear, nonconservative interaction process between two subpopulations X and Y , we shall consider the following pair of equations:

$$\begin{aligned}\frac{dx}{dt} &= A_1x + B_1x^2 + C_1xy, \\ \frac{dy}{dt} &= A_2y + B_2y^2 + C_2xy.\end{aligned}\tag{27}$$

The coefficients are to be interpreted as follows. The A 's are net "birth" or "death" rates; the B 's are (positive or negative) contributions due to contacts between individuals in the same state; the C 's are contributions (positive or negative) due to contacts between individuals in opposite states.

By giving different signs to these coefficients, we can describe various models qualitatively different from one another. For example, if $A_1 < 0$, $B_1 > 0$, $C_2 > 0$, $C_1 = -C_2$, we see that members of the X -population tend to disappear if left to themselves, to generate more of their own kind after contacts with their own kind, and to be "converted" to the Y -population by contact with it.

Any combination of the six coefficients can be interpreted in a similar manner. Of particular interest are certain special cases which have been given a biological interpretation. For example, let the X -population consist of predators that feed on the Y -population, which, in turn, feeds on an unlimited food supply. In that case we must have $A_1 < 0$ (the predators cannot survive in the absence of their prey); $C_1 > 0$ (the biomass of the predators increases with the number of contacts with the prey); $A_2 > 0$ (the prey multiplies in the absence of the predators); $C_2 < 0$ (the biomass of the prey decreases on contact with predators, as the prey is eaten by them). The signs of B_1 and B_2 depend on whether contacts between members of the same species enhance or inhibit the growth of the respective

biomasses or populations. Such equations have been treated by Volterra (1931), Kostitzin (1937), and other biomathematicians.

Another interesting case represents a competition between two species, each of which could exist in a given environment without the other. Here the A 's are positive, and all the other coefficients are negative. The B 's represent inhibition to population growth due to intraspecies crowding, whereas the C 's represent inhibition due to interspecies crowding.³

Treatments of such nonlinear systems are often confined to the examination of their statics, that is, the equilibrium conditions. Both derivatives of Eq. 27 are set equal to zero, and the expressions on the right are factored. Assigning signs to the coefficients appropriate to the competition model, we now have

$$\begin{aligned} x(A_1 - B_1x - C_1y) &= 0, \\ y(A_2 - B_2y - C_2x) &= 0. \end{aligned} \tag{28}$$

We see that $x = y = 0$ is a trivial equilibrium point. A nontrivial equilibrium is found at the intersection of the two straight lines, whose equations are the expressions in parentheses of Eq. 28 set equal to zero. The equilibrium is biologically meaningful if the intersection is in the first quadrant. It is stable if certain additional conditions are satisfied by the coefficients.

Carrying the analysis through, we have the following results on the statics of the system representing competition between two species.

1. The equilibrium point will be in the first quadrant (i.e., biologically meaningful if and only if either $B_2/C_1 > A_2/A_1 > C_2/B_1$ or $C_2/B_1 > A_2/A_1 > B_2/C_1$).

2. If a biologically meaningful equilibrium exists, it will be stable if and only if $B_1B_2 > C_1C_2$, that is, if the product of self-restraint coefficients is greater than the product of the other's restraint coefficients.⁴

If the two straight lines determined by Eq. 28 do not intersect in the first quadrant, there will be no meaningful equilibrium. One of the lines will lie entirely above the other in the first quadrant, and the species whose line is farthest from the origin will be the sole survivor. The conditions for the survival of X are $A_1/C_2 > A_2/B_2$ and $A_1/B_1 > A_2/C_2$, and, of course, the conditions for the survival of Y are just the reverse. Figure 2 represents the entire situation graphically.

Whatever biological or social interpretation is made of similar models,

³ "Crowding" is measured by frequency of contacts, which is proportional to the products (or squares) of the densities.

⁴ Note the similarity of this condition to that in Richardson's arms-race model (Sec. 1.2).

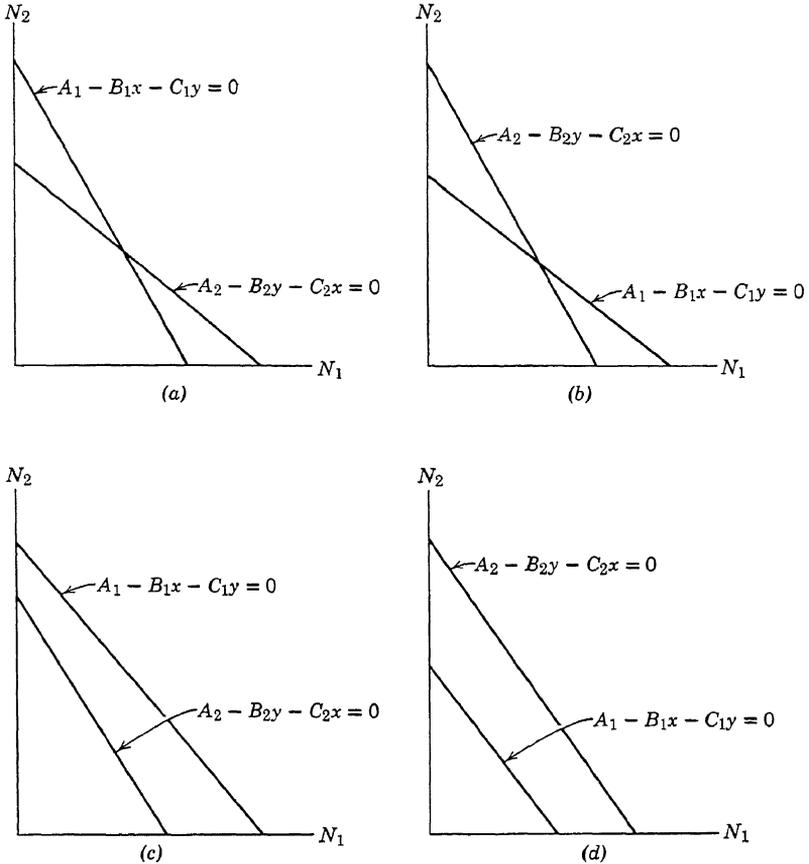


Fig. 2. Cases of interspecies competition between populations N_1 and N_2 (after Gause): (a) stable equilibrium—co-existence; (b) unstable equilibrium—only N_1 or N_2 will survive, depending on initial conditions; (c) only N_1 will survive; (d) only N_2 will survive.

the investigation of consequences always follows the same method.⁵ In the absence of explicit dynamic solutions (time courses of the dependent variables), which are difficult or impossible to obtain in the general (non-numerical) case, the investigation is confined largely to the examination of the “phase space,” that is, the space determined by the dependent variables. At each point of this phase space we can calculate dx/dt and dy/dt . These derivatives determine the direction of motion of the point (x, y) in the phase space. Moreover, the quantity $[(dx/dt)^2 + (dy/dt)^2]^{1/2}$

⁵ For further discussion of the statics of such systems see Gause (1934, 1935) and Slobodkin (1958). For extensions to stochastic models see Neyman, Park, & Scott (1955).

gives the "speed of motion" of this point. We can therefore draw a vector at each point in the phase space. Important properties of the system can be determined by investigating the nature of the resulting vector field.

In Sec. 4.1 we shall examine another such system in a different context, which will serve as a conceptual link between the mathematical models related to population dynamics and those related to mathematical economics and decision theory.

2. STATISTICAL ASPECTS OF NET STRUCTURES

In our treatment of mathematical theories of contagion, which can also serve as mathematical models of many other kinds of social interactions, we have kept the assumptions of well-mixedness throughout, with only a single reference to work in which a diffusion term has been included (Landahl, 1957). We may now ask what happens if there is no diffusion at all, that is, when everyone stays put and interacts only with his "neighbors." Evidently, in this case, the social interaction process will depend on how many neighbors each individual has. We have seen an analogue to the limited number of neighbors in the constant A/k of our contagion models described on p. 507, this constant being the number of individuals *ever* infected by an infected individual. However, aside from the finite size of this number, there was no further limitation. It was given how many "neighbors" (contacts) each individual could have but not who they were. The question "Who are the neighbors?" does not refer here to their intrinsic properties but to their relations to *their* neighbors. In a well-mixed population I have always assumed that the individuals who were my contacts' contacts would with equal probability be any individuals in the population. With the introduction of the "neighbor" concept, this assumption is no longer tenable. For one thing, it is natural to endow the relation "neighbor" with a symmetrical property: I am one of my neighbors' neighbors. But, if so, then I am certain to be found among the set of individuals who are my neighbors' neighbors, in contrast to the equiprobability assumed for my contacts' contacts in previous models.

These considerations lead us to the study of net structure. The branch of mathematics which deals with such questions rigorously is the theory of linear graphs. We shall examine some graph-theoretical treatments of social structure in Sec. 3.1. Since we are for the moment dealing with large populations, for which it is out of the question to list all the relations among the entities, graph theory will not be of much help. We shall resort instead to examining some gross statistical properties of nets.

2.1 The Random Net

Our point of departure will be to determine certain statistical properties of a so-called random net, defined as follows. Let a certain fixed number a of directed line segments issue from each node of the net. Let each of these line segments, which we shall call "axones" to bring to mind the analogy with neural nets, connect at random to any of the other nodes. To define a "connection at random," we imagine a chance device with N equiprobable states, where N is the number of nodes in the net. We take each of the aN axones in turn and determine its "target," that is, the node on which it terminates by the chance device. The resulting net we call a random net.⁶

We now seek a mathematical description of some of the properties of such a net. Our first concern is with the results of a certain "tracing procedure." Start with an arbitrary number of randomly selected nodes, this number being small compared with N . Call the set of nodes, which are targets of all of the axones of this initial set, excluding the initial set itself, targets of the first remove. Call all the nodes that are targets of all the axones from the targets of the first remove, excluding the targets of the first remove and the initial set, targets of the second remove, etc. Let $p_0, p_1, p_2, \text{ etc.}$, be the corresponding population fractions. We seek a recursion formula for the successive p_t ($t = 1, 2, \dots$).

Select an arbitrary node and consider the probability that it belongs to the targets of the $(t + 1)$ th remove. This is a product of two probabilities, namely, (1) that the node in question does not belong to any of the targets of previous removes and (2) that one of the axones from the targets of the t th remove does connect with the node in question. The product of the probabilities is justified by the fact that the two probabilities are independent, since all of the connections are determined by the chance device without any reference to the state of the tracing procedure. Moreover, the probability that the node in question does not belong to the target of all the removes before the $(t + 1)$ th is the sum of the component probabilities, since the sets defining the targets of the successive removes are mutually exclusive.

We can therefore write

$$p_{t+1} = \left(1 - \sum_{j=0}^t p_j\right) \left[1 - \left(1 - \frac{1}{N}\right)^{aNp_t}\right]. \quad (29)$$

⁶ In a more general model the number of axones per node, a , can itself be a random variable. This generalization does not modify the principal results of our model, and we shall not make it.

The factor in the brackets is well approximated, for large N , by $1 - e^{-ap_t}$, and Eq. 29 can be simplified to

$$p_{t+1} = (1 - \sum_j p_j)(1 - e^{-ap_t}), \quad (30)$$

which is the recursion formula required.

If we denote by x_t the fraction of nodes contacted by the t th remove, that is, $x_t = \sum_{j=0}^t p_j$, we can write Eq. 30 after appropriate rearrangements,

$$\text{as} \quad (1 - x_{t+1})e^{ax_t} = (1 - x_t)e^{ax_{t-1}}. \quad (31)$$

Since the equality holds for all values of t , the equated expressions must be independent of t , that is, equal to a constant. The constant can be evaluated by setting $x_0 = p_0$, and we obtain the equation, which determines $\gamma \equiv x(\infty)$ as a function of p_0 and a , namely,

$$\gamma = 1 - (1 - p_0)e^{-a\gamma}, \quad (32)$$

which is formally identical with Eq. 23.

Let us solve for a in terms of the p_j in Eq. 30. Since the p_j (the expected values of the fractions representing the magnitudes of the sets of targets of the successive removes) are completely determined by the "average" tracing procedure, it follows that the expression representing a will be a function of t alone. *Formally*, a is a function of t , although it is a constant by definition, and so must be independent of t . To indicate this (fictional) dependence on t , let us designate a by $\alpha(t)$ and call it the "apparent" axone density for reasons that will presently become clear. We have accordingly

$$\alpha(t) = \frac{1}{p_t} \log \frac{1 - x_t}{1 - x_{t+1}}. \quad (33)$$

If, in any observed tracing, the function of t , represented on the right side of Eq. 33, is indeed independent of t , we have experimental corroboration of the hypothesis that the net in question is a random net. It may very well be that $\alpha(t)$ will turn out to be a constant in some nonrandom nets. However, if $\alpha(t)$, as determined from empirical determinations of the successive p_j in the tracing of an actual net, is *not* constant, we have a refutation of the hypothesis that the net is random. Moreover, the behavior of $\alpha(t)$ may give us some indication of the nature of the non-randomness of the net we are studying.

Now it is clear why we have called $\alpha(t)$ the "apparent axone density." If the value of $\alpha(t)$ is adjusted so that every successive p_j , as calculated from Eq. 30, with $\alpha(t)$ replacing the exponent a , is equal to the observed

value of p_t , we can interpret $\alpha(t)$ as the axone density related to the t th remove, which in a *random* net would determine the observed value of p_{t+1} .

2.2 Biased Nets

Let us see how we would expect $\alpha(t)$ to behave if biases of a certain kind were operating in the structure of a net. In a real net, we may expect some sort of a distance bias to determine the actual connections. If the net is a net of social relations, say the acquaintance relation, we can imagine that the nodes are immersed in some sort of "social space." The topology of this space is by no means easy to discern, but we feel intuitively that neighborhoods can be defined in it. In particular, suppose the social space resembles geographical space, since it is certainly partially determined by it. We can then expect our neighborhoods to resemble geographical neighborhoods to some extent. Apart from geography, we expect certain symmetry biases and certain transitivity biases to operate. A symmetry bias makes itself felt in the fact that if an axone from node A terminates on node B (say A knows B) the probability that an axone from B will terminate on A (B knows A) is greater than the a priori probability. A transitivity bias operates if it is true that whenever an axone from A terminates on B and an axone from B terminates on C , the probability that another axone from A will terminate on C is greater than the a priori probability. (If A knows B and B knows C , it is likely that A knows C .) Combining these two biases, we have the circularity bias: whenever an axone terminates on B and an axone from B , on C , it is likely that an axone from C will terminate on A .

All these biases ensure that the number of targets in the *second* remove will be smaller than expected in a random net. The number of targets in the first remove is not affected by the biases because the initial nodes have presumably been chosen randomly; hence the targets will be determined randomly regardless of what biases operate in the population. The targets of the second remove are the "friends of friends" (assuming "friendship" as the net relationship) of the initial nodes. Thus axones from the first remove are more likely to "converge" on certain targets (common friends) in the second remove. There will be more "hits" per target hit, and consequently fewer targets will be hit than expected on a random basis.

It follows that if a axones are traced at each remove we shall have $\alpha(0) = a$ (since the biases do not disturb the random selection of the targets of the first remove), but $\alpha(1)$ will be smaller than a by our previous considerations. The drop in the value of α from $\alpha(0)$ to $\alpha(1)$ provides us

with a rough index of "cliquishness" or "tightness" or "compactness" of the social space in which the net under consideration is submerged. Note that the vaguely descriptive terms "cliquishness," etc., have not been exactly defined here except as the corresponding property manifests itself in the reduction of α . This is in accord with our investigation of only the gross properties of biased nets.

In particular, we may investigate the expected behavior of $\alpha(t)$ for some special kinds of social spaces.

Let us define an individual's *acquaintance circle* as a set of individuals from whom his sociometric choices are to be made or who are likely to choose him. We shall take the number in this set to be constant for all individuals and shall denote it by q ($q > a$). The question how these q acquaintances were chosen from the population now arises. If they were chosen randomly from the population, no essential modification would be introduced, even if $q \ll N$, so long as $q \gg 1$, as we shall now show.

We have, in fact, instead of Eq. 29,

$$p_{t+1} = [1 - x(t)] \left[1 - \left(1 - \frac{1}{q} \right)^{ap_t a} \right]. \quad (34)$$

If $q \gg 1$, the last parenthesis can still be approximated by e^{-ap_t} , and so Eq. 34 reduces to Eq. 30.

If q , although large compared to a , is small compared to N , we can reason along a different line, which will bring us to essentially the same result but will also provide a point of departure for introducing a bias.

Let us fix our attention on an arbitrary individual X at the tracing of the $(t + 1)$ st remove. We seek the probability that on that remove X was *not* chosen by an arbitrary but definite individual A from among his q acquaintances. This can happen in either of two mutually exclusive ways: either A himself was not chosen on the t th remove or he was chosen on the t th remove, but his own choices did not include X . The probability we seek is the sum of the probabilities of these two events, that is, $1 - p_t + p_t(1 - 1/q)^a$.

Now if all the states of the acquaintances of X are independent of one another and if q is small compared to the total population, so that sampling with replacement can be assumed for any sample of individuals not greater than q , then the probability that X was not chosen on the $(t + 1)$ st remove, that is, the probability that he was not chosen by any of his q acquaintances (the only individuals who could choose him) will be given by

$$\left[1 - p_t + p_t \left(1 - \frac{1}{q} \right)^a \right]^a. \quad (35)$$

Therefore, the probability that X was chosen on the $(t + 1)$ st remove is

$$1 - \left(1 - p_t m \right)^a, \quad (36)$$

where we have written m for $1 - (1 - 1/q)^a$. Expression 36 corresponds to $1 - e^{-ap_t}$ in Eq. 30. We therefore write for our modified expression, representing the recursion formula of the tracing,

$$p_{t+1} = (1 - x_t)[1 - (1 - p_t m)^q]. \tag{37}$$

Solving for $\alpha(t)$, as defined by Eq. 33, we have

$$\alpha(t) = \frac{-q}{p_t} \log(1 - p_t m). \tag{38}$$

If q is large and a is small, $m \ll 1$ and a fortiori $p_t m \ll 1$, so that we can approximate the logarithm by $-p_t m$ and $\alpha(t)$ by qm . For large q this is approximately a . Thus the introduction of a finite but sufficiently large acquaintance circle, which limits the set of individuals who can choose a given individual or be chosen by him, does not lead to an appreciable modification of the result. However, this approach lends itself to the imposition of a sociostructural bias, which we now discuss.

Until now the assumption underlying the whole argument has been that the probabilities $1 - p_t m$, namely, the probabilities that each of the q individuals in X 's acquaintance circle did not choose X on the $(t + 1)$ st remove, were all equal and independent. Another way of saying this is that our knowledge that the first acquaintance did not choose X did not affect our assumption regarding the state of the second acquaintance, etc. If we drop this assumption of independence, the compound probability that none of X 's q acquaintances chose X can no longer be represented by Expression 36. Instead of the q th power, we must write a q -fold product

$$\prod_{k=0}^{q-1} [1 - p_k(t)m], \tag{39}$$

where the $p_k(t)$ are conditional probabilities to be determined. Using Expression 39 instead of Expression 35 in Eq. 37, and solving for $\alpha(t)$, defined by Eq. 33, we now have

$$\alpha(t) = \frac{-1}{p_t} \sum_{k=0}^{q-1} \log [1 - p_k(t)m]. \tag{40}$$

As before, assuming large q and small a , the logarithms in Eq. 40 can be well approximated by $-p_k(t)m$, and we have the simplified form of $\alpha(t)$, namely

$$\alpha(t) = \frac{m}{p_t} \sum_{k=0}^{q-1} p_k(t). \tag{41}$$

In the special case of the completely mixed population, all the $p_k(t)$ are equal for a given t , and $\alpha(t) = qm \simeq a$, as, of course, should be the case.

If a bias operates, the $p_k(t)$ cannot be assumed to be equal. How they will vary with k depends on the nature of the bias. We assume in what follows that the acquaintance circles are "strongly overlapping." The exact meaning of this assumption will, I hope, become clear in the discussion.

Consider the state of affairs we have described in which, we recall, choices are made among fairly large but finite acquaintance circles, the latter having been "randomly recruited." Because of random recruitment, the acquaintance circles of two individuals, A and B , who are themselves acquainted, have the same expected intersection as the acquaintance circles of any two arbitrarily selected individuals. Suppose now we are given the "density" of individuals newly chosen on the t th remove in A 's acquaintance circle (i.e., the probability that an arbitrary individual in that set is among the p_t individuals of the t th remove). Call this density $p_t(A)$. According to our assumptions of arbitrarily recruited acquaintance circles, this knowledge in no way modifies our knowledge of the density of the individuals in B 's acquaintance circle, $p_t(B)$, because the two acquaintance circles are in no way related.

Suppose now that both A and B are in the acquaintance circle of X . The knowledge that A did not happen to choose X on the $(t + 1)$ st remove somewhat modifies our estimate of the $p_t(A)$ because of the way this probability can be calculated as a conditional probability, given that A did not choose X . It does not modify our estimate of $p_t(B)$ in the random case. Therefore, the probability that B did not happen to choose X , given that A did not, remains the same as the a priori probability given by $1 - p_t m$. This is the justification for Expression 35 as the probability that none of X 's acquaintances chose X on the $(t + 1)$ st remove.

The bias of strongly overlapping acquaintance circles is reflected in the assumption that knowledge of the $p_t(A)$ does modify our estimate of $p_t(B)$ because both A and B are in X 's acquaintance circle and so probably in each other's.⁷ Hence knowing that A did not choose X introduces a new *contingent* probability that B is among the p_t individuals of the t th remove. Calculating this probability by Bayes' rule, we obtain

$$p_1(t) = \frac{p_t(1 - m)}{1 - p_t m}, \quad (42)$$

⁷ Note that we cannot make the assumption of complete transitivity of the acquaintance relation, namely, that two individuals who are in the acquaintance circle of a third are also in each other's. This would imply that the entire population was partitioned into mutually exclusive cliques of q members, each with random choices within the cliques. We would then have several random nets instead of a single biased one. The word "probably" in the sentence to which this footnote refers points up the approximate and nonprecise assumption of the strongly overlapping acquaintance circles. Roughly speaking, there are leaks in the cliques.

which reduces to p_t for $m = 0$, that is, when the acquaintance circles are infinitely large. If this is not the case, we must take $p_1(t)$ as the density of t th remove individuals in B 's acquaintance circle. If B also did not happen to choose X , we get a further modification of our estimate of the density of t th remove individuals in the vicinity of A . By iteration, we get

$$p_k(t) = \frac{p_{k-1}(t)(1 - m)}{1 - p_{k-1}(t)m}, \tag{43}$$

and by induction

$$p_k(t) = \frac{p_t s^k}{1 - p_t + p_t s^k}, \quad (k = 1, 2, \dots, q). \tag{44}$$

where $s = 1 - m$.

All the $p_k(t)$ being determined, we calculate $\alpha(t)$ by Eq. 41, which for large q is approximated by

$$\alpha(t) = \frac{-1}{p_t} \log [1 - p_t(1 - e^{-a})]. \tag{45}$$

For small p_t , the right side of Eq. 45 is well approximated by $1 - e^{-a}$.⁸

We now relax the assumption of "strongly overlapping acquaintance circles" by introducing an additional parameter θ , which represents the average fraction of individuals in the acquaintance circles who are included in the overlap bias. Alternatively, we can say that in selecting acquaintances, each individual selects a fraction θ in accordance with the overlap bias and a fraction $1 - \theta$ randomly. This modification leads to the following approximate expression for $\alpha(t)$ (Rapoport, 1953):

$$\begin{aligned} \alpha(0) &= a, \\ \alpha(t) &= 1 - e^{-a\theta} + (1 - \theta)a, \quad (t \geq 1). \end{aligned} \tag{46}$$

For $\theta = 0$, $\alpha(t)$ reduces to a , and for $\theta = 1$ to $1 - e^{-a}$. The two special cases thus coincide with the random net and with the strongly overlapping acquaintance circle net, respectively. Thus θ will appear as a measure of the "tightness" of the net.

2.3 Application of the Overlapping Clique Model to an Information-Spread Process

The model described in the preceding section was put to a test in experiments on information spread conducted by the Washington Public Opinion

⁸ Details of the calculation are given by Rapoport (1953). In the empirical tests of the theory to be described subsequently, p_t seldom exceeds 0.1 and the approximations introduced throughout are well justified.

Laboratory (Dodd, Rainboth, & Nehnevajsa, 1952) at the University of Washington. Subjects were seventh grade children and college students. The experiments were designed so that the values of p_t were available. Thus $\alpha(t)$ could be computed and tested for constancy. The observed values of $\alpha(t)$ of the seventh graders are shown in Table 1. The results obtained from college students were similar.

Table 1 Values of $\alpha(t)$, the Apparent Axone Density, in an Information-Spreading Net of School Children for Successive Values of t

| t | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|-------------|------|------|------|------|------|------|------|------|------|------|------|------|
| $\alpha(t)$ | 7.23 | 1.47 | 1.63 | 1.80 | 1.98 | 2.32 | 2.23 | 2.90 | 1.76 | 3.01 | 1.54 | 1.21 |

We see from Table 1 that $\alpha(t)$ exhibits a conspicuous drop on the first remove and thereafter rises steadily for eight removes. The decline on the last two removes is probably not significant, for toward the end of the process $\alpha(t)$ becomes exceedingly sensitive to p_t , so that small fluctuation errors in the data produce very large fluctuation errors in $\alpha(t)$.

Equation 46 is able to account for the initial drop of $\alpha(t)$ from an arbitrary initial value to a lower value. However, Eq. 46 also implies that subsequently $\alpha(t)$ remains approximately constant, in contradistinction to the data shown in Table 1.

It was necessary, therefore, to make an additional hypothesis, namely, that there was in the course of the spread a progressively increasing randomization of contacts. This hypothesis is psychologically plausible. The experiment was conducted under contest conditions, the subjects being motivated to get and to pass on as much information as possible. It is reasonable to suppose that as a subject's own acquaintance circle became saturated with knowers he was more and more likely to seek random contacts. This process is tantamount to a steady decline of θ as a function of x (the cumulated fraction of "knowers"). In order to minimize the number of parameters, it was assumed that $\theta(0) = 1$, that is, were it not for the increasing randomization of contacts motivated by the contest conditions, the spread of information would be described by a net with tightly overlapping acquaintance circles. Taking θ as a decaying exponential function of x , namely, $\theta = e^{-\beta x}$, where β is a constant, we obtain, instead of Eq. 46,

$$\alpha(t) = [1 - \exp(-ae^{-\beta x_t})] + (1 - e^{-\beta x_t})a. \quad (47)$$

Thus the free parameter θ has been replaced by the free parameter β , which measures the propensity of a carrier to seek random contacts, as his

acquaintance circle becomes saturated with knowers. We note that setting $\beta = 0$ is equivalent to setting $\theta = 1$, whereas setting $\beta = \infty$ reduces $\alpha(t)$ to a , that is, it reduces this bias to zero. When Eq. 47 is substituted for a in the fundamental recursion formula 30, the modified recursion formula in terms of x_t and p_t is obtained:

$$x_{t+1} = 1 - (1 - x_t)\{1 - p_t[1 - \exp(-ae^{-\beta x_t})]\} \exp[-a(1 - e^{-\beta x_t})]p_t. \quad (48)$$

Equation 48 contains two free parameters, a and β . It was this equation that was applied to the data obtained from the seventh graders. Comparison of predicted and observed values is shown in Fig. 3.

The first point is given by the actual value of p_0 ; the second point is used for computing a , and the third for computing β . The remaining points are predicted by the resulting equation with numerical values of the two parameters substituted, namely, $a = 7.2$, $\beta = 0.22$. Similar results were obtained from the data on college students with $a = 4.4$ and $\beta = 0.3$.

If we ventured to take the fits seriously, we could, in view of the psychosociological interpretation of a and β , conclude that the children made on the average more contacts than the college students (as reflected in the larger value of a) but were less inclined to randomize their contacts as the contest proceeded (as reflected in the smaller value of β). It is a moot question, of course, whether the fit can be taken as a strong indication of the validity of our model in its present form. For one thing, the derivation of Eq. 48 is not rigorous. The so-called overlap of acquaintance circles cannot really be expressed by a single parameter because of the tremendous complexity of the actual acquaintance structure. Actually the bias resulting from "first-order transitivity" (if A knows B and B knows C , then A is also likely to know C) implies certain biases of higher transivities involving longer chains. The exact computation of parameters representing sociostructural bias is an extremely difficult matter, even in the most drastically idealized cases. Second, the time factor has been completely omitted from the theory, whereas time, as an explicit variable, certainly had an effect on the process (interaction frequencies varied radically during the course of the day) in terms of clock time and thus indirectly in terms of the removes. The striking agreement between theory and data must be attributed, as in many such cases, to the over-all smoothness of the curves, which allow a good fit with two free parameters.

To be sure, the theory outlined here gives not only a fit but also a rationale for some aspects of social diffusion. It is therefore advisable to design further experiments to test the particular rationale. Some of these experiments are discussed in the next section.

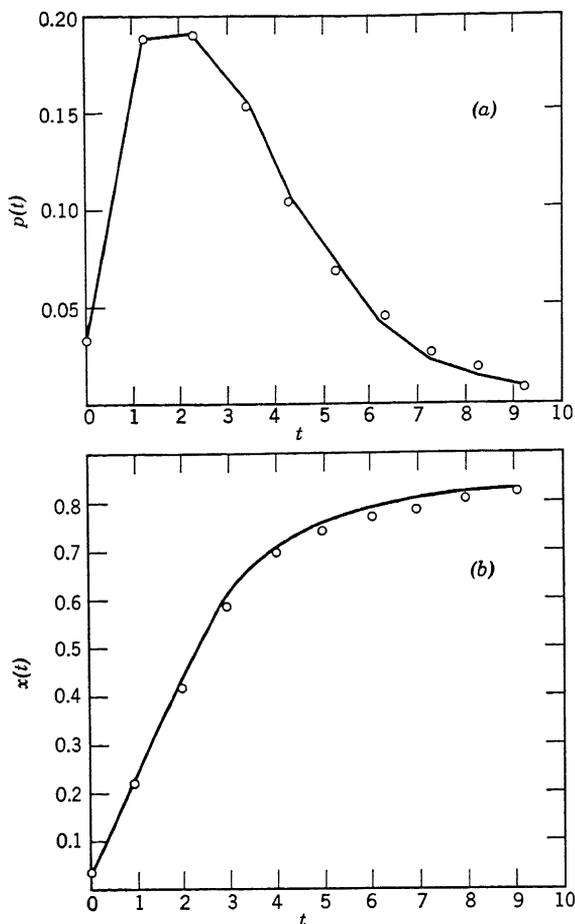


Fig. 3. Comparison of theoretical curves (Eq. 48) with data obtained from experimentally induced spread of information among school children (circles): (a) increments in the fraction of knowers against removes; (b) cumulative fractions of knowers against removes.

2.4 Application of the Biased Net Model to a Large Sociogram

We now abandon the study of the diffusion and contagion processes and concentrate on a statistical study of the population structure as such, which is, after all, the real subject matter underlying the recursion Formula

30. This formula traces not so much occurring contacts as existing channels for contacts. For example, if we were to follow the assumption of "guilt by association" to its logical conclusion, the Formula 30 would indicate the expected number of individuals who would be implicated in each tracing of acquaintanceship. The crucial variables involve the average number of acquaintance bonds and the nature of the sociometric bias.

To trace such a net in an actual population, we need only to ask each individual to list his potential contacts. The "tracing" would be done from these lists. Suppose we ask that the potential contacts of each individual consist of n individuals ordered according to the relative frequency of contact or according to contact intimacy. Now if we select any number from 1 to n for our a , we can, beginning with an arbitrary set of starters, trace our net. The "axone density" a is now no longer a free parameter but a parameter controlled by the experimenter. Since a represents only the average number of contacts, the value of a need not even be an integer. We can, for example, make $a = 3/2$ by tracing two contacts for one half of each set of individuals belonging to each remove and one contact for the remaining half.

By choosing the contacts to be traced high or low on the list, we would presumably be generating curves for high or for low values of θ , since it is likely that the closer contacts are more strongly "inbred." Finally, the parameter β , which appeared in our information spread model, should not enter at all in the sociogram tracing because the motivating factor of which β was a reflection is absent: our tracing is determined only by the structure of the acquaintance relationship and not by any acts of the individuals.

We could trace any number of curves in which the parameters p_0 and a would be chosen at will and only the single free parameter θ would be inferred. If all of these curves gave good agreements between predicted and observed values of p_t or x_t and if θ were found to be monotone decreasing as one chose the names to be traced farther down on the list, a rather strong confirmation of our theory would be obtained; or possibly indications would appear for promising modifications of the theory. The number of empirical investigations that could be conducted on the statistical properties of a large sociogram is large. Aside from the opportunities they offer for constructing a theory of sociometric space, they can serve as foundations of a "natural history" of sociometric nets.

Data were obtained from two junior high schools (population 800 to 1000) in Ann Arbor, Michigan, about two months after the beginning of the school year. Each pupil was asked to fill in the blanks in the statements, "My best friend in this junior high school is _____"; "My second best

friend in this junior high school is "———", etc., through "eighth best friend." Data on one of the schools are presented here.

The assumptions require that all choices be within the population. Therefore choices that could not be identified as pupils in the school (made through accidental or deliberate violations of the instructions) had to be disregarded. Thus some blanks appeared in the data cards. Absentees appeared as cards with all blanks. The same number of choices by everyone is not strictly required by the model, since the parameter a represents only an average axone density or a tracing. Therefore, if a tracing was made with an intended $a = 2$ (e.g., first and second friends only), the actual average a was reduced, in our case to about 1.78, and this is the value that was used in calculating the predicted curve.

The null hypothesis, namely, that the net is random, can be definitely rejected, as shown in Fig. 4. Next, we assume an overlapping clique bias with a free parameter θ whose value fixes $\alpha(t)$ for $t \geq 1$. The value of $\alpha = 1.1$ or $\theta = 0.8$ gives a reasonably good fit, as shown in Fig. 5. Obviously, the fit could be still better if we had another free parameter at our disposal. However, if the theory is to remain a rational one, such a parameter cannot be chosen *ad hoc*. It must be "rationalized." Rationalizations of additional parameters suggest themselves in biases that may be operating in the sociogram in addition to the bias of overlapping acquaintance circles, already taken into account.

Note that the overlapping acquaintance bias implicitly imposes a "metric"

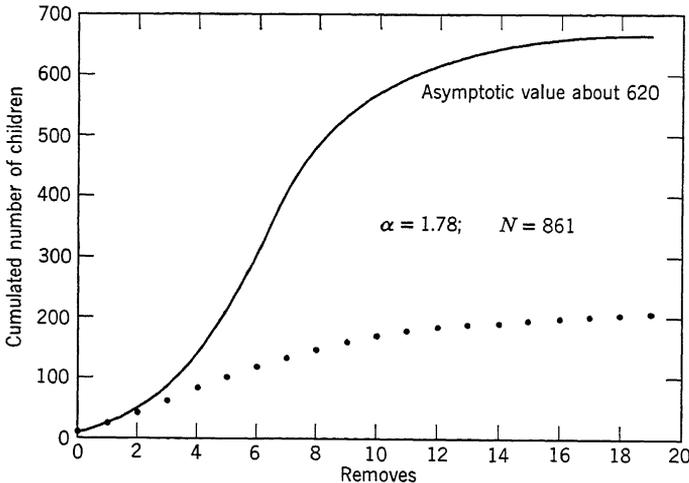


Fig. 4. Comparison of tracing through first and second friends (points) with the curve predicted by a tracing through a random net with the same actual axone density (1.78).

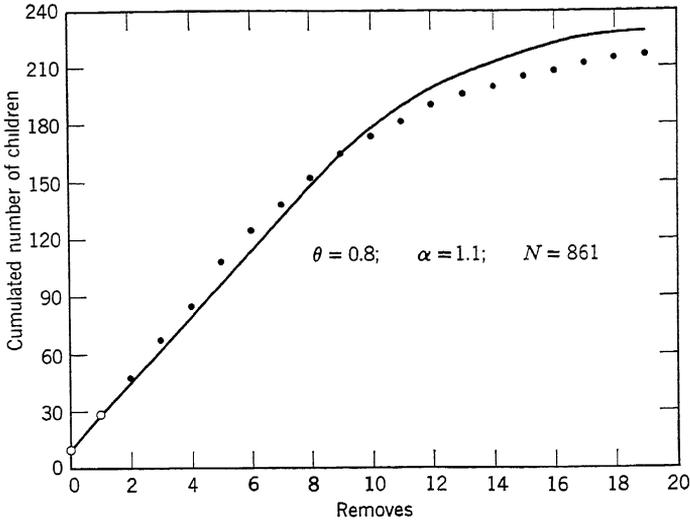


Fig. 5. Comparison of tracing through first and second friends (points) with the curve predicted by a tracing through a net with overlap bias $\theta = 0.8$.

on the social space, although the precise nature of this metric or of the underlying topology is not specified. Thus any other distance bias we would impose, for example, a “reciprocity bias” in which choices tend to be reciprocated, or “transitivity biases” of higher order, would either already have been implied by the overlap bias or might be implicitly inconsistent with it. If we wish to add an independent free parameter, we should seek a bias that is at least not obviously related to the overlap bias. Such may be the “popularity bias,” which results if some individuals “attract” axones and others “repel” them. This distinction defines an *inherent* property of the individuals in the population, not a relation among pairs of individuals, and therefore can be supposed to be unrelated to the overlap distance bias.

To introduce a popularity bias, we could assume or determine empirically a distribution of attractiveness in the population and modify the net model accordingly. A much simpler way is to take the grossest feature of this bias as a single free parameter. Note that a popularity bias tends to reduce the “effective” population through which the tracing is made. This can be seen in the extreme case in which only a fraction of the population can attract the axones at all, since this simply leaves the others out as members of the population.⁹ In any case the predominance of choices

⁹ Note the analogy with the notion of the susceptible population in the theory of contagion.

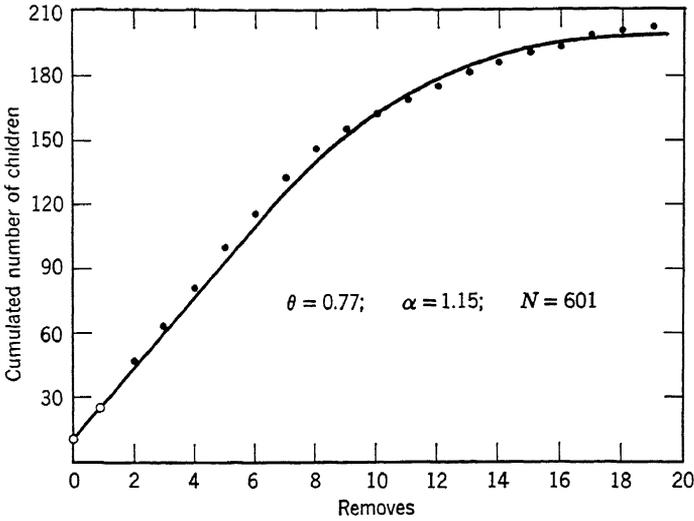


Fig. 6. Comparison of tracing through first and second friends (points) with the curve predicted through a net with $N = 601$ and overlap bias $\theta = 0.77$.

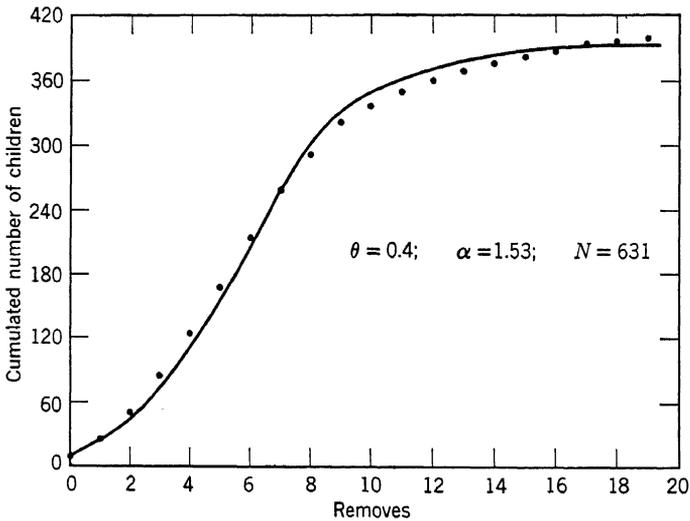


Fig. 7. Comparison of tracing through fourth and fifth friends (points) with the curve predicted by tracing through a net with $N = 631$ and overlap bias $\theta = 0.4$.

directed at some individuals at the expense of others reduces the possible number of targets of the axones and so reduces the size of the "effective" population in a tracing.

If we are free now to choose the size of this effective or apparent population as a free parameter, we obtain a fit shown in Fig. 6.

Figure 7 shows a further test of theory. Here the result of the average of 30 tracings is shown with intended $a = 2$ (actual $a = 1.73$), where the tracings were made through the fourth and fifth friends on each list. As can be seen from the values of the free parameters that give the best fits, the effective population is about the same size as the tracing obtained from first and second friends. The value of α , however, is 1.53. Calculating

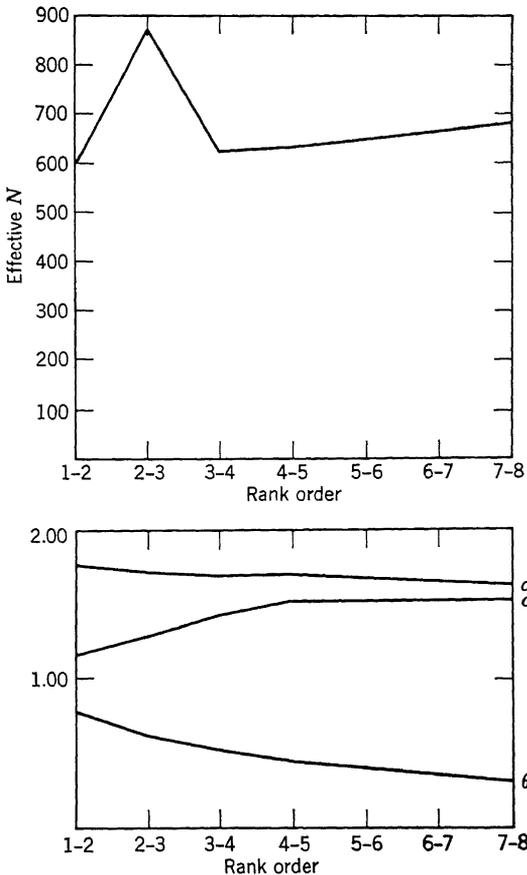


Fig. 8. The various parameters in tracings through pairs of friends of different consecutive ranks.

θ from Eq. 46, we have the value 0.4 for that parameter. We conclude that by the time the first three friends have been named the remaining ones will be named considerably more at random; that is, the resulting net approaches a random net.

Next, we test the hypothesis that θ is monotone decreasing with the numerical rank order of the friends through which the tracing is made. We make tracings through second and third friends, through third and fourth, and through seventh and eighth. Figure 8 summarizes the results.

We see from Fig. 8 that the actual α remains approximately constant, 11 to 14% below the intended value ($\alpha = 2$), except for the tracing through seventh and eighth friends, where it drops more. The deficiency in the actual α is accounted for by the absences and by some failures to name a friend or to name a pupil in the school, the last two failures becoming prominent in the tracings through seventh and eighth friends. The gradual rise in α reflects the decreasing tightness of the overlap bias with increasing numerical rank order of the friends. This effect is masked in the last value because of the drop in α , of which α is a monotone-increasing function. Looking at the behavior of θ , which expresses the actual tightness of the overlap bias, we see that it is indeed monotone decreasing.

If it were not for the abnormally high value of N^* , the "effective" population, associated with the tracing through second and third friends, this parameter would be monotone increasing with the numerical rank order. This would indicate that the popularity bias is *decreasing* with the rank order, a plausible result. The fact that the best fit for the tracing through second and third friends is obtained without popularity bias (i.e., with the actual $N = 861$) remains unexplained.

The existence of the popularity bias can be observed directly in the distribution of the numbers of choices received in the population. If every one had an equal chance to receive a choice, this distribution would be of the Poisson type. The observed distribution departs radically from the Poisson, in fact follows closely the so-called Greenwood-Yule distribution, as shown in Fig. 9.

Our model suggests a number of theoretical problems. One is to eliminate the *ad hoc* character of the "effective population," assumed to be a consequence of the popularity bias, that is, to deduce the value of this parameter theoretically. Another problem is to relate the overlap bias to some directly observed biases, such as the reciprocity and transitivity biases of several orders. Still another rather intriguing problem is to infer topological and metrical properties of social space, for example, to determine whether the population can be "immersed" into an Euclidean space of a few dimensions so that the distance between any two members would appear as the number of links in the path leading from one to the other.

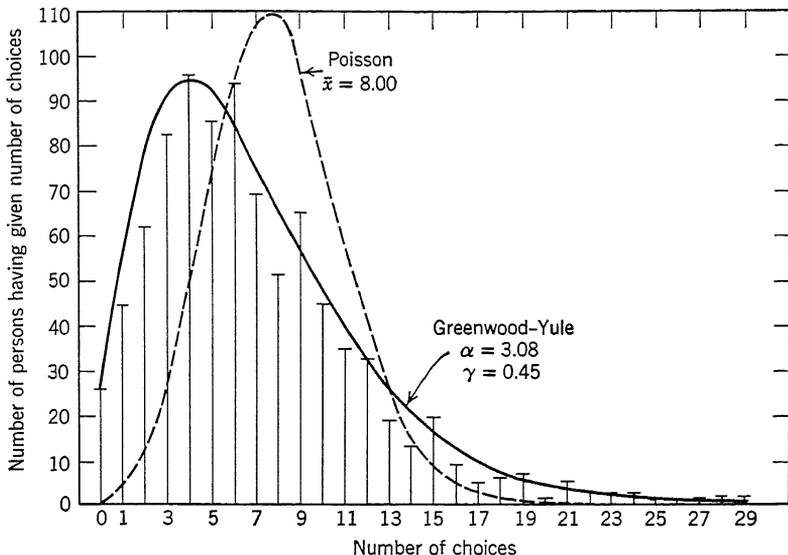


Fig. 9. Comparison of the distribution of the number of choices received with the Poisson distribution that would obtain if the choices were random. The actual distribution is fitted well by the Greenwood-Yule distribution. The parameter γ (which vanishes, as the Greenwood-Yule distribution degenerates into a Poisson distribution) can be taken as an index of the popularity bias.

Since distance must be a symmetric relation in any metric space, obviously an average of the two not necessarily equal distances separating a pair of members must be taken to define this variable.

We leave the large-scale models at this point and turn to theories of structure and interaction in small groups.

3. STRUCTURE OF SMALL GROUPS

In physics it is commonplace to describe the state of a system by a closed formula, from which we can “read off” the conditions in every one of the parts of the system. For example, if a formula gives the gravitational potential as a function of space coordinates, we know what the potential is at every point in the region considered. There is thus no limitation on the size of the region or on “how many points” are contained in it.

The necessity of dealing with *explicitly enumerated* sets of relations (instead of with global formulas from which an infinite number of relations can be read off) naturally places a severe restriction on the size of the

universe with which a social scientist can deal in these terms. At times this restriction can be circumvented if the relations are specified only statistically, as was done in Sec. 2. It is clear, however, that exact specification without the use of encompassing formulas is possible only if the number of entities and relations is not too large.

Fortunately, there are situations in which only a few individuals are involved, but in which, nevertheless, certain regularities can be observed, and so a portion of social science can be constructed to deal with the observed events. These are situations involving small groups.

Definitions of small groups based on content-oriented, sociopsychological concepts (motivation for existence, degree of involvement, etc.) are not our concern here. Only formal properties will be examined. Suffice it to say, then, that a small group is a collection of individuals, to each subset of which a definite value of each of a finite set of relations can be assigned. If all the subsets are pairs, the relations are binary. All subsequent discussions are confined to binary relations.

In the light of this definition, natural classification schemes for small groups follow at once. The simplest are those in which only one relation exists between each pair, and this relation has one of two values. (N.B. The presence or absence of a single-valued relation is equivalent to a two-valued relation existing between the members of every pair of the set of entities that constitutes the group.)

A finite set of points, among some of whose pairs a single symmetric relation, having a single value, exists is called a *linear graph*. Obviously, a linear graph can also be viewed as a set of points with a two-valued symmetric relation defined for all its pairs. The equivalence of the two definitions can be seen at once if the two values of the relation are "present" and "absent."

Related concepts are established by specifying relations between members of ordered pairs and/or by allowing more than two values for each relation. For example, a *directed graph* consists of a finite set of points, among some of whose *ordered* pairs a single relation, having a single value, exists. A *signed graph* allows the relation to have two values if present (plus or minus). Similarly, we define *signed directed graphs* and also graphs with more than one relation, which are, as we have seen, equivalent to graphs with a single relation of many values. When there is no danger of confusion, we shall refer to all species of graphs simply as linear graphs.

The mathematical theory of graphs is a branch of pure mathematics. In recent years some mathematicians have been developing the theory of graphs in the context of formalized social relations in small groups (e.g., Harary & Norman, 1953).

It is important to note that a system composed of a set of entities and a relation defined for each ordered pair can also be represented by a matrix (a_{ij}) , in which the entry a_{ij} in the i th row and j th column is the value of the relation associated with the ordered pair of individuals (i, j) . If the relation is symmetric (antisymmetric), the associated matrix is symmetric (antisymmetric.)

Besides the formalism of linear graph and matrix representations, certain other methods of describing small group structures are suggested by the context of the problems considered. Some of them are discussed in Sec. 3.4.

3.1 Descriptive Theory of Small Group Structures

The term "mathematical model" is sometimes used in two different senses. The usual meaning refers to a set of precisely stated postulates concerning some aspects of a system; for example, the interdependence of the variables which describe its state or the laws governing the time course of some process. Such postulates are *assertions*, and they lead via a deductive chain of reasoning to conclusions (theorems) that are also assertions. In another sense, a "mathematical model" is a collection of *definitions*. Definitions do not assert anything. They only indicate how words are going to be used.

When Harary and Norman (1953) speak of the theory of graphs as a "mathematical model in social science," they have in mind, at least in the beginning, this second kind of model, which we shall call descriptive, in distinction from the first kind, which we shall call predictive. In a descriptive model of a social group based on the theory of graphs the social group and a set of relations among its members is conceptualized as a linear graph (a directed graph, a signed graph, etc.). When this is done, theorems about the linear graph, which is assumed to be isomorphic to the social group, can be translated into corresponding statements about the social group. In this context the validation of such statements is a purely logical validation, a consequence of the assumed isomorphy between the graph and the social group. Only when additional statements are included as postulates about social groups *apart* from formal properties of linear graphs does the model become predictive (i.e., empirically *falsifiable*), and its adequacy becomes a matter of empirical validation.

In some instances graph theory has been used as a framework of predictive models in the theory of small-group structure and behavior. I shall first present the predominantly descriptive approach.

Table 2 contains in the left-hand column some definitions underlying

the theory of graphs. The right-hand column contains possible translations of these definitions into terms applicable to small social groups.

Like any branch of mathematics, graph theory is a collection of theorems deduced from the definitions of the sort listed. Like the definitions, the

Table 2 Correspondence Between Terms in Graph Theory and Terms Related to Description of Social Groups

| | |
|---|--|
| Connected graph: a linear graph in which a path (along the lines of the graph) exists between any two points. | Connected social group: a set of individuals among whom communication between any two (possibly via intermediaries) is always possible |
| Completely connected graph | Single clique group |
| Subgraph: a subset of points, and lines associated with them, of the set that constitutes a graph. | Subgroup: a subset of individuals and their associated communication channels. |
| Component: maximal completely connected subgraph. | Clique: maximal completely connected subgroup. |
| Articulation point: if it is possible to divide a graph G into two sets U and V having only point P in common, such that every path from a point of U to a point of V includes P , then P is an articulation point of G . | Liaison person: one whose removal would turn a single connected group into one not connected. |
| Bridge: a line of a connected graph whose removal separates the graph into two components, each of which has more than one point. | Critical communication channel: one which, if cut, results in the severing of communication between two subgroups. |

theorems too can be translated into the language of social relations. However, the significance of the theorems is not always easy to evaluate.

As an example, consider a theorem of König (1936) on a certain species of linear graph called "trees." A tree is a connected graph without cycles. Each point of a tree (or of any connected graph) has an "associated number," defined as the maximum of its "distances" to other points. (The distance between two points is the smallest number of lines necessary to traverse to get from one to the other.) König's theorem states that every tree has either one or two points whose associated number is minimal.

It is felt that the theorem may have relevance for the theory of small groups because of the following considerations. The analogue of "associated number" in social groups is the notion of "centrality." Whenever communication must go "through channels," the effectiveness and, perhaps, the morale of individual group members can be reasonably supposed to depend on their "distance" (in terms of the number of "relay stations") from the other members. One could even suppose that the individual whose associated numbers are smallest (those with largest "centrality" index) would be the more likely to assume positions of leadership in a group (cf. Bavelas, 1948). König's theorem can therefore be interpreted to mean that if the communication structure of a group is a tree (i.e., has no cycles) then there are *at most* two individuals in it with the largest possible centrality index.

It would be far-fetched to pursue the interpretation beyond this formal implication; for example, to conclude that a power struggle for leadership in a "treelike" group will either fail to materialize (if only one "central individual" exists in it) or will take place between no more than two candidates, etc. Such conjectures are not warranted in view of the immense complexity of real social situations compared with the schematized mathematical models. Nevertheless, it must be admitted that formal mathematical conclusions derived from structural characteristics of graphs have a certain suggestive potential.

On the other hand, certain types of social relations are sufficiently well defined to allow a translation of theorems strictly derived from structural postulates into assertions about such relations. Kinship relations and rules governing marriage are of this sort. The rules specify who may marry whom. In so-called "primitive" societies (i.e., among people who live in remote places), these rules typically involve the kinship relations (e.g., prohibitions of incest) and, in addition, specify certain "marriage types," so that marriage is permitted only among individuals of the same type.

If marriage is *always* to be permitted between any two individuals of the same type and if brother-sister and parent-children marriages are always to be prohibited (as they are in nearly all societies), then obviously sons and daughters must be assigned marriage types different from that of their parents (who are, we recall, of the same type) and different from each other. If, furthermore, the marriage type of a person's parents is to be uniquely inferred from the person's marriage type, there is a one-to-one correspondence between the marriage types of parents and that of sons on the one hand and that of daughters on the other; that is, two *permutations* operate in the assignment of marriage types, one for sons and one for daughters.

If there are n marriage types, they can be represented as a vector

$\mathbf{t} = (t_1, t_2, \dots, t_n)$. A permutation transforming this vector into a vector of marriage types assigned to the sons or to the daughters can be represented by an $n \times n$ permutation matrix. There are thus two permutation matrices, S for the sons, and D for the daughters.

Now a family tree can be represented as a directed linear graph, in which there are just two types of bonds, namely, a descendent bond (from parent to child) and a marriage bond. Moreover, the nodes of the graph will be of two kinds, male and female. Marriage rules imply that in a family tree marriage bonds are found only between individuals of opposite sex and of the same type.

The permutation matrices, S and D , now enable us to label each individual by marriage type, given the type of an ancestor. Thus the sons of a man of type t_i will have the type designated by the i th component of the once permuted vector $S\mathbf{t}$; the sons of his son will have the corresponding type in the twice-permuted vector $S^2\mathbf{t}$; the sons of his daughters will be read off from $SD\mathbf{t}$; the daughters of his sons from DSt , etc. We can infer the marriage type of an ancestor by applying the inverses of the permutation matrices: S^{-1} for a man's parent, D^{-1} for a woman's parent, $D^{-1}S^{-1}$ for a man's maternal grandparent, etc. Products of permutations matrices are also permutation matrices. The matrix that determines a permutation of marriage types associated with a given relationship is called the relation matrix M . Thus $M = DSD^{-1}S^{-1}$ is the relation matrix for a man's "cross cousin" (mother's brother's daughter.)

The mathematical method just described allows us to do more than determine whether marriage is permitted in any specific case. It allows us also to decide whether a given rule for the assignment of types is compatible with a given set of rules governing marriage and also to choose methods of type assignment to generate given marriage rules.

Examples of such general findings have been stated in the form of theorems, such as the following.¹⁰

A man is allowed to marry a female relative of a certain kind if and only if his marriage type does not belong to the effective set of M [that is to say, the component of the vector (t_1, t_2, \dots, t_n) corresponding to his marriage type remains invariant after transformation by M associated with the relationship].

If it is to be permissible for some of the descendants of any two individuals to marry (i.e., if marriage prohibition does not extend to all blood relations), then for every i and j there should be a permutation (in the group generated by S and D) which carries i into j .

¹⁰ For definitions of terms used in matrix theory and group theory, e.g., "effective set" and "generated group," see Kemeny, Snell & Thompson (1957).

If whether a man is allowed to marry a female relative of a given kind depends only on the kind of relationship (i.e., if the same rule is to apply to individuals of all types), then in the group generated by S and D every element except I (the identity permutation) is a complete permutation.

It is in a way remarkable that the intricate and explicit marriage rules in societies which have them are, as far as is known, consistent. One suspects of course, that the experience of many generations would have weeded out inconsistencies and that the rules have become second nature to the practitioners, who need no formal deductive system to arrive at conclusions. To the outsider, however, a formalized logical system is very useful not only in that it provides algorithms of reasoning but also because it suggests generalizations and applications in a variety of contexts.¹¹

Katz (1953) has used the method of matrix algebra to redefine the concept of "status" based on sociometric choice in a group. The conventional "popularity" definition of status is related to the number of times a group member is chosen by other members in situations in which sociometric choices are recorded. Katz redefines status by having it depend not only on how many times one has been chosen but also by whom. The idea is to make the choices by high status members count more.

If C represents the matrix of sociometric choices with $c_{ij} = 1$, if i chooses j and $c_{ij} = 0$ otherwise, let

$$T = aC + a^2C^2 + \dots + a^kC^k + \dots = (I - aC)^{-1} - I, \quad (49)$$

where I is the identity matrix and a ($0 < a < 1$) is an "attenuation factor" which weights indirect choices of various order of removes (choices of choices, etc.) by its corresponding successive powers. The columns of T contain components that represent the choices accorded to the corresponding member weighted by the statuses of the choosers and also indirect choices (of higher removes) attenuated by the powers of a . The column sums of T , then, divided by an appropriate integer (analogous to the total possible number of choices in the "popularity" definition of status) give the modified status index, in which the status of the choosers plays a part in determining the status of the chosen.

Katz shows how in an artificially constructed group the proposed status index reflects these properties (cf. also Forsyth & Katz, 1946).

¹¹ Hoffman (1959) has applied symbolic logic methods to similar problems and has derived a marriage rule in Pawnee society, which, he observes, is not found in the ethnographic record: *A man's marriage partner must be the granddaughter of the sister of the man's father's father.* (N.B. In Pawnee society cohabitation and marriage are not synonymous.)

3.2 The Detection of Cliques and Other Structural Characteristics of Small Groups

Festinger (1949), Luce and Perry (1949), Harary (1959), and others have used matrix algebra as a tool for detecting certain structural features of small groups.

To fix ideas, consider the matrix

$$A: \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Evidently, the corresponding directed graph is the one shown in Fig. 10.

Let us now examine A^2 :

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

The entries of A^2 , $a_{ij}^{(2)}$ are $\sum_k a_{ik}a_{kj}$, where the a_{ij} , the entries of A , are either 0 or 1. The nonzero entries of A^2 , therefore, come from all the paths of length 2 from i to j (the 2-chains in Luce and Perry's terminology) in the structure represented by A . Therefore the entry $a_{ij}^{(2)}$ in A^2 represents the number of distinct 2-chains from i to j .

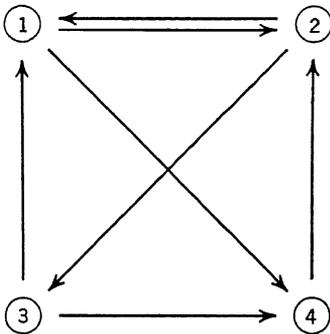


Fig. 10. Directed graph corresponding to matrix A .

This result is immediately generalizable: the entry $a_{ij}^{(k)}$ represents the number of distinct k -chains from i to j . We see that the structure of this group is represented in the entries of the successive powers of the associated matrix.

Examining now the diagonal elements of A^2 , we see that the same result implies that the i th diagonal entry represents the

numbers of elements in the group with which the i th member has symmetric relations (two-way bonds).

Pursuing the interpretation of structures in a social-psychological context, Luce and Perry (1949) define a *clique* as a maximal completely connected subset of the original structure containing at least three persons, that is, a completely connected subset which is not properly contained in a larger completely connected subset. In other words, all members of a clique have symmetric connections with one another, but no other group member stands in a symmetric relation to *all* the clique members (or he too would have to be counted as a clique member).

This definition seems natural enough, but on second thought it may seem somewhat restrictive, as Luce and Perry themselves point out. For example, let a subgroup of a given social group be "very tightly knit" according to any intuitive standard of judgment, except for a few bonds missing. This situation violates the clique definition. Therefore, the definition fails to distinguish cliques in a sense useful to the social psychologist or the sociologist. This objection is not a formal one, of course. It merely points out that many subgroups which the sociologist may consider well qualified to be called cliques may not satisfy the mathematical qualification. In a later paper Luce (1950) relaxes his definition of clique to include more general subgroups.

To a certain degree the clique structure of a group can be determined by examining the cube of a matrix S derived from A by eliminating all unreciprocated connections; that is, S is the element-wise product, $S = A \otimes A'$, where A' is the transpose of A .¹² The i th entry of the main diagonal of S^3 gives us information about whether i is a member of a clique: he is if the entry is different from zero.

Should there be only one clique of t members in the group (which does not mean, of course, that everyone is in it), the diagonal elements of S^3 will show it: the corresponding t elements will have entries $(t - 1)(t - 2)$, and the remaining entries will be zero, a result obtained earlier by Festinger (1949).

Harary (1959), working primarily with symmetrical structures, describes a procedure for detecting all the cliques by way of determining those with "unicliquical" members. A unicliquical member is a group member who belongs to only one clique. A noncliquical member is one who belongs to no clique. Harary's clique-detecting procedure begins with the deletion of all the noncliquical members. This is easily done by examining the element-wise product $S^2 \otimes S$, where S is the symmetric part of the original matrix. A

¹² In the element-wise product each entry of the product matrix is simply the product of the corresponding entries of the factor matrices.

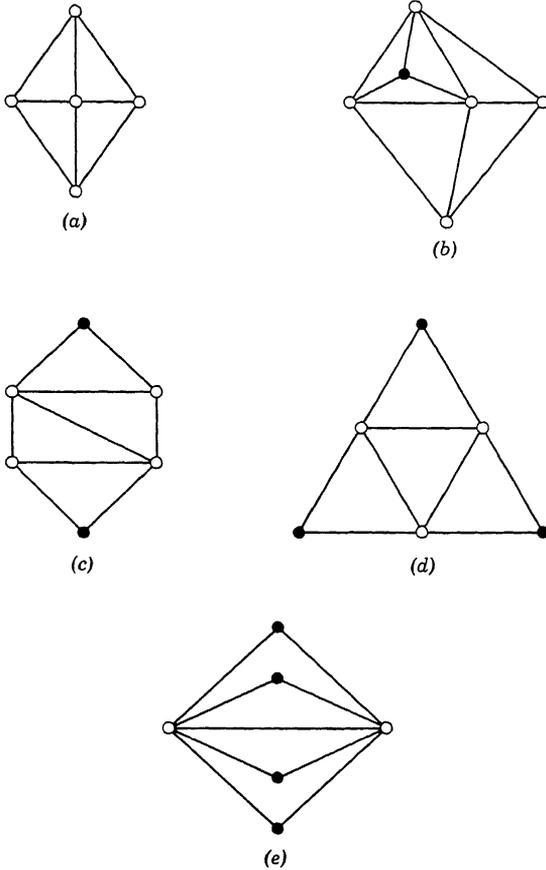


Fig. 11. Groups with four cliques. The black circles are uniclqual members.

member is noncliquial (as we can readily verify) if and only if the corresponding row of $S^2 \otimes S$ consists entirely of zeros.

Now, let G be the group from which all the noncliquial members have been removed. Then, obviously, if G has only one clique, all its members are uniclqual. It is also obvious that if G has exactly two cliques then both cliques must have uniclqual members (otherwise every member would belong to both cliques and the two cliques would be identical). Somewhat less obvious are the corresponding statements for the cases in which G has three cliques or more. In the first case at least two of the three cliques have uniclqual members; in the second case there is no restriction on the number of uniclqual members there may be in the group. All the possibilities for the four clique cases are illustrated in Fig. 11.

These results are useful in the method for detecting cliques proposed by Harary. After all the noncliquical members have been removed, we determine all the cliques having uncliquical members, delete these members, and iterate the process. If the resulting group has no uncliquical members, it is split into two subgroups in such a way that each has fewer cliques and each clique lies entirely within each subgroup. When the number of cliques in a subgroup becomes sufficiently small (<3), uncliquical members of that subgroup are sure to appear by the results cited. We can thus proceed to eliminate uncliquical members and be sure that we have counted all the cliques in the process.

The foregoing are samples of investigations of abstract structures. The investigations have been carried out in the spirit of pure mathematics; that is, the results obtained were sought because of the logical interconnections among the questions asked about the configurations studied and not necessarily because of direct "usefulness" of these results for understanding aspects of analogous structures in the real world, such as small groups of particular interest to the psychologist or the sociologist. Moreover, the investigations were descriptive in the sense that the results were a display of structural features (to be sure, mathematically deduced) and not behavioral predictions. To pass to the behavioral predictions, hypotheses are required that would relate structures to actual events or to some underlying tendencies for events to occur. In the next sections we shall be concerned with investigations of this sort.

3.3 The Theory of Structural Balance

Consider a sociogram determined by a set of individuals and a two-valued symmetric relation. Between the members of each pair there is either a "positive" or a "negative" bond or no bond. The relation can be psychologically interpreted as "liking," "disliking," or "indifference."

A hypothesis has been advanced in social psychology (Heider, 1946; Newcomb, 1953, 1956) to the effect that two persons' attitudes toward each other are influenced by their attitudes toward some third object. For example, two persons who both like the same things tend to like each other; two persons who dislike the same things also tend to like each other. But if two persons have opposite attitudes toward the same thing, they tend to dislike each other.¹³ The three situations are pictured in Fig. 12(a,b,c).

¹³ Dislike generated by conflict over coveted objects provides an obvious important exception.

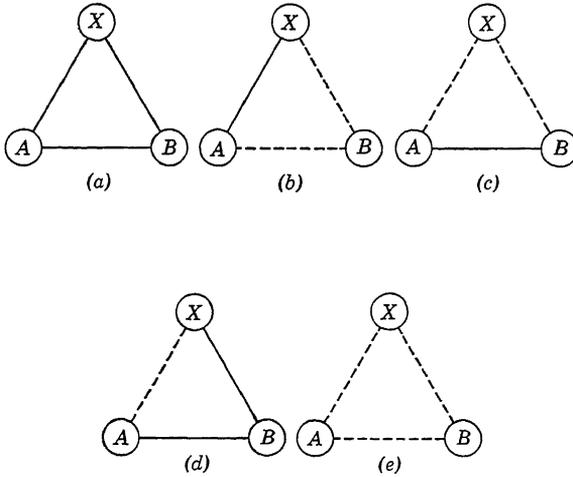


Fig. 12.

Situations (a), (b), and (c) in Fig. 12 are “balanced” in the sense that the hypothesis previously stated is satisfied. In (a), A and B both like X and each other; in (b), A and B have opposite attitudes toward X and dislike each other; in (c) they both dislike X and like each other. In (d), however, the hypothesis is violated: A and B like each other, in spite of having opposite attitudes toward X . In (e) the hypothesis is also violated because A and B dislike each other in spite of having the same (negative) attitude toward X .

Suppose, now, we assign a positive sign to solid lines and a negative sign to dotted lines. We define the “sign” of the cycle $A \rightarrow B \rightarrow X \rightarrow A$ as the product of the three signs of its three lines, following the algebraic convention that the product of like signs is positive and of unlike signs, negative. We see, then, that if and only if the hypothesis of balance is satisfied the sign of the associated cycle is positive.

The object X can, of course, be a person as well as a thing (or an institution or an idea). We may then examine the 3-cycles of any social group viewed as a signed symmetric graph to determine whether they satisfy the balance hypothesis. Moreover, we can generalize this procedure by examining the signs of cycles larger than 3-cycles, using the same rule of multiplication of signs.

A signed graph is called balanced if and only if all of its cycles are positive. It now becomes of interest to examine the evidence for the following hypothesis, which is an extension of the preceding one: a signed graph representing a sociogram of a social group tends to become balanced.

The hypothesis implies roughly that attitudes of the group members will tend to change in such a way that one's friends' friends will tend to become one's friends and one's enemies' enemies also one's friends, one's friends' enemies and one's enemies' friends will tend to become one's enemies, and moreover, that these changes tend to operate even across several removes (one's friends' friends' enemies' enemies tend to become one's friends by an iterative process). Another way of saying the same thing is that a social group tends to split into two subgroups (one of which may be empty) such that members within each subgroup like each other, whereas members from the two different subgroups (if there are two) dislike each other. The formal equivalence of the two statements has been proved by Harary (1954).

If a sociogram of a group is given, we can determine whether the associated graph is balanced by examining the sign of each cycle. Since, in general, there are many cycles in a moderately large and densely connected group (say a fraternity house), it is too much to expect the hypothesis of balance to be satisfied completely. However, the *trend* toward balance may still be verifiable, provided that we accept a quantitative instead of an all-or-none definition of balance, that is, a definition of the degree of balance of a graph and of its associated social structure. Such quantitative definitions were offered by Cartwright and Harary (1956) and by Harary (1959).

Evidence for the existence of secular trends toward structural balance, as defined by mathematicians, is meager. One longitudinal study comes close to establishing a result that may be related to this hypothesis. Newcomb (1956) conducted a set of consecutive observations on the 17 residents of a student house at the University of Michigan. In the course of time (the study lasted one semester and was replicated), the attitudes of those who were attracted to each other tended to come closer together, including views the subjects held of their own selves and their "ideal" selves.

Many social-psychological studies deal with related hypotheses in the realm of attitudes, interactions of attitudes, and resulting tensions or tension resolutions, but a rigorous testing of the mathematical theories of structural balance is still lacking. A review of the literature on this topic has been given by Zajonc (1960).

3.4 Dominance Structures

Let the relation of interest between any two individuals in a small group be one of dominance. For example, $A > B$ may mean that A tends to influence B 's decisions or that A tends to win in chess from B or that A

tends to be preferred to B in sociometric choices by others of the group when the choice involves A and B alone (a paired comparison). Such a relation is by its very nature antisymmetric, that is, if $A > B$ is true, then $B > A$ is not true. Further, we would ordinarily expect such a relation to be transitive, that is, $A > B$ and $B > C$ might be expected to imply $A > C$. If this is the case, a well-ordering of the members of the group is determined by all the relations between pairs. Interesting questions arise if the relation is not transitive, that is, when we may have $A > B$, $B > C$, and $C > A$.

Such cycles in dominance relations are actually observed, for example, in the behavior of hens which manifests the so-called "peck right." Although a complete hierarchy (well-ordering) according to peck right is often established in a flock of barnyard hens, cycles are also commonly observed. The violations of transitivity observed in social-dominance relations have given rise to various theoretical developments of social interaction, some of which we shall now consider.

The usual interpretation of the nontransitivity of the dominance relation rests on the assumption that this relation is established to a certain extent by chance events. The resulting models are analogous to certain stochastic models designed to explain intransitivities of preferences resulting from sequential paired comparisons. However, in the context of a stochastic theory of social structure, certain aspects of these models receive emphasis that they do not receive in the context of individual preference theories. We shall pursue the developments of some such models accordingly in the present context.

We take for our chance event the result of combat encounter between two individuals. We shall assume that the result is "victory" for one of them and that peck right is accorded to the victor. We assume that encounters have occurred among all pairs. Thus a peck-right structure representable by a directed completely connected graph is established.

Next, we seek to classify the structures. A natural classification would identify a structure with a class of linear graphs isomorphic to it. Naturally a renaming of individuals should not affect the type of structure. As the number of individuals increases, the number of possible nonisomorphic linear graphs becomes rapidly very large.¹⁴ It follows that the classification

¹⁴ The number of distinct antisymmetric matrices of order N is $2^{N(N-1)/2}$. Some of these represent isomorphic graphs obtained by renaming the individuals (interchanging some rows and corresponding columns). Each structure is therefore represented by at most $N!$ matrices, and so $2^{N(N-1)/2}/(N!)^{-1}$ is a lower bound on the number of nonisomorphic dominance structures of N -person groups. For $N = 8$ this lower bound is already more than 6000; for $N = 12$ it exceeds 10^{11} . Studies on the number of graphs of various types appear in mathematical literature (e.g., Katz & Powell, 1954; Davis, 1953, 1954).

proposed is too fine to be practical, since the number of distinct structures becomes so large even for moderate N that it is hopeless to observe the "frequency of occurrence" of each structure, which is the usual test of a postulated stochastic process supposed to underlie the establishment of the structure. We shall therefore introduce a rougher measure, namely, a "score structure," defined as follows.

In every group of N individuals, each will have $N - 1$ relations with the others, of which d will be dominant and $N - 1 - d$ will be submissive. The group can then be described by a set of N numbers (r_1, r_2, \dots, r_N) such that $\sum r_i = \frac{1}{2}N(N - 1)$. This set of numbers arranged conventionally so that $r_1 > r_2 > \dots > r_N$ will be called the score structure of the group. For $N > 4$ there are fewer nonequivalent score structures than social structures, and the difference increases rapidly as N increases. Therefore the score structure gives us a rougher classification. A still rougher index was introduced by Landau (1951). Note that the score structure ($N - 1, N - 2, \dots, 1, 0$) corresponds to the completely hierarchical structure of a group, in which the individual with the highest score dominates all the others; the one with the next highest score dominates all but one, etc. Landau's hierarchy index measures the departure of a given score structure from that of the complete hierarchy. He defines

$$h = \frac{12}{N^3 - N} \sum_j \left(r_j - \frac{N - 1}{2} \right)^2. \quad (50)$$

This definition ensures that in an "egalitarian" group, in which the score structure has all equal components and therefore $r_j = (N - 1)/2$, ($j = 1, 2, \dots, N$), $h = 0$; also h is maximal in a hierarchy. The factor outside the summation in Eq. 50 is a normalization factor, which makes $h = 1$ for a hierarchy.

Following Landau (1951), suppose each individual j is characterized by an "ability vector" $x_j = (x_{j1}, x_{j2}, \dots, x_{jm})$. The components of the vector are the various factors that presumably have a bearing on the probability that, in an encounter between two individuals, one will emerge the victor. Among these factors in hens are size, concentration of male hormones (making for the appearance of secondary male sex characteristics), etc.

We may assume that these characteristics are distributed according to a multivariate distribution in the population from which the individuals have been selected. By our definition of the factors x_j , it appears that the probability that individual j will dominate k is a function of the corresponding vectors:

$$\begin{aligned} \Pr(j > k) &= p(x_j, x_k) = p_{jk}, & (j, k = 1, 2, \dots, n); \\ p_{jk} + p_{kj} &= 1. \end{aligned} \quad (51)$$

The problem of determining the actual multivariate distribution of the components and the probabilities p_{jk} is not of the essence in the theoretical investigation to follow.

We seek instead some estimate of the expected value of the score structure h :

$$E(h) = \frac{12}{N^3 - N} \sum_j E(r_j - r)^2, \quad (52)$$

where $r = (N - 1)/2$.

Assuming the same multivariate distribution $F(x)$ for all of the individuals in the group, Landau proves the following general result:

$$E(h) = \frac{3}{N + 1} \left[1 + (N - 2) \int A^2(x) dF(x) \right], \quad (53)$$

where $A = \int g(x, y) dF(y)$, and $g(x_j, x_k) = p_{jk} - p_{kj}$.

Introducing various special assumptions concerning $F(x)$ and p_{jk} , Landau then obtains corresponding values of $E(h)$. In particular, in the completely unbiased case, in which the two possible outcomes of each encounter are equiprobable (independent of the ability vectors), he gets

$$E(h) = \frac{3}{N + 1}, \quad (54)$$

a result obtained previously by Rapoport (1949) directly in the special cases for $N = 3, 4$, and 5 .

To calculate $E(h)$ specifically in a biased case, assumptions must be made about the multivariate distribution $F(x)$ and about the way the probabilities p_{jk} depend on the ability vectors. Assuming the ability components to be normally distributed with variances σ_α ($\alpha = 1, 2, \dots, m$) and the probability of dominance to be a weighted sum of normal probability integrals with variances s_α ($\alpha = 1, 2, \dots, m$), Landau obtains the following expression for $E(h)$:

$$E(h) = \frac{3}{N + 1} \left[1 + \frac{2(N - 2)}{\pi} \sum_{\alpha=1}^m w_\alpha^2 \arcsin \frac{\sigma_\alpha^2}{s_\alpha^2 + 2\sigma_\alpha^2} \right], \quad (55)$$

where the w_α are the relative "weights" of the ability factors.

This reduces to the unbiased case if s_α becomes infinitely large (implying $p_{jk} = \frac{1}{2}$ for all j, k). If $s_\alpha = 0$, dominance is *determined* for every pair by the ability vectors alone, and $E(h)$ reduces to unity, as it should.

The main point of these calculations is to show that for any moderately large N the expected value of the hierarchy index h should be small, that is, considerably less than unity. The greater the number of (uncorrelated) ability components, the smaller this expected value. But even for *one*

component, Landau shows that $E(h)$ can be expected to be close to unity only if very small differences in ability are decisive in determining the direction of dominance established in an encounter.

Experimental evidence indicates that the dominance-determining power of known factors is actually quite small. Collias (1943) staged 200 combats between hens, taking each hen of a pair from a different flock, and correlated the outcomes with measured degrees of moult, comb size, weight, and rank in its own flock. The correlation coefficients were respectively .580, .593, .474, and .262. These correlations lead to 0.34 as the value of $E(h)$ for large N . On the other hand, the observed values of h in flocks of 10 hens (10 is a large number in this context) are in the nineties (Schjelderup-Ebbe, 1922).

Landau's conclusion is that the observed near-hierarchy established by "almost transitive" peck right cannot be accounted for by "inherent abilities" of the hens alone. It is natural to look for "social factors," that is, the role of experience within the flock, imitation, learning, etc., for reasonable explanations of the high values of the hierarchy index observed in nature. The conclusion is not surprising in view of the fact that workers in animal sociology have long suspected the operation of these factors. However, deriving this conclusion from mathematical and statistical considerations has for the mathematical sociologist a methodological interest because it puts the problem into theoretical perspective.

Leeman (1952) used a similar approach to construct a mathematical model of sociometric choice patterns in a small group. His model differs from that studied by Rapoport and Landau in that only one directed line issues from each group member; hence only one sociometric choice is made by each member at each specified time. Moreover, in the sociometric choice model the binary relation is not necessarily antisymmetric, as it is in the dominance structure model.

Specifically, Leeman assumes that sociometric choices are established by encounters between pairs. In each encounter either the person encountered is chosen or any of the other members of the group is chosen with equal probability. Thus at each moment of time a pattern of choices is established. The stochastic process leads to the probability distribution of all possible choice patterns.

Leeman computes these distributions for a three-person group (in which there are two possible nonisomorphic choice patterns) and for a four-person group, in which six nonisomorphic patterns are possible. The theory is extended to a biased probability case, and some experiments are cited, the results of which lead essentially to the rejection of the model based on equiprobable outcomes of encounter.

Luce, Macy, and Tagiuri (1955) treat a similar problem in which,

however, the relation between any pair of individuals can have 45 different values. This relation, called a "diad," is defined as follows. Assuming that an individual in a group can choose, reject, or ignore another individual and, in turn, can feel himself chosen, rejected, or ignored, it follows that individual A can relate himself to another individual B in the nine different ways in which his own attitude and his perception of the other's attitude can be combined. A 's nine relations can be combined with B 's nine in 45 different ways (disregarding order). Hence a diad can have 45 different values.

In a random model the choices and guesses of each individual are assumed to be governed by independent chance events. Thus, no psychological factors are supposed to operate except those governing the relative frequencies of choices and perceptions. In a biased model a dependency is introduced between an individual's attitude toward another individual and his guess about the other's attitude to bias the events toward greater congruence of attitude and perception.

Comparison with data obtained from a group-therapy session involving a 10-person group shows that the biased model accounts for a large part of the observed variation in diad frequency.

4. PSYCHOECONOMICS

Theories of population dynamics and those of interaction in small groups are linked by a common mathematical apparatus, first utilized extensively by Cournot (1927 translation of 1838 volume), an early mathematical economist. Recently his and related ideas have been cast into the framework of social-psychological experiments and have borne results of considerable interest and, perhaps, of sufficient importance to be considered as foundations of experimental psychoeconomics.

4.1 A Mathematical Model of Parasitism and Symbiosis

Before we examine these experiments let us first discuss a fictitious psychoeconomic model, closely related to the population dynamics models discussed in Sec. 1.6. In this way we shall introduce a conceptual link between population dynamics and psychoeconomics.

Consider two individuals, X and Y , each of whom produces a different commodity in the respective amounts x and y . Each, being in need of both commodities, agrees to exchange a fraction of his own output for

a fraction q of the other's output. Thus each keep the fraction $p = 1 - q$ of his own output.

Assume that there is a positive, logarithmic contribution to each individual's utility from what he receives in commodities and a negative contribution, proportional to the output (presumable because of the labor involved). Specifically, designating the utilities by S_x and S_y , we have

$$\begin{aligned} S_x &= \log(1 + px + qy) - \beta x, \\ S_y &= \log(1 + qx + py) - \beta y. \end{aligned} \tag{56}$$

The one's in the arguments of the logarithms were introduced to make the positive part of the utility vanish when $x = y = 0$.

The situation can be considered as a two-person, nonzero-sum game, in which each player has a continuum of strategies in the x - and y -space, respectively, so that the strategy space is the product space (x, y) .

In choosing his strategy each player naturally wishes to maximize his utility. But since he controls only one of the variables, all he can do is make the "best response" to each strategy chosen by his opponent, that is, choose that value of his variable which maximizes his own utility, *given* the choice of output by the other.

Setting $\partial S_x / \partial x$ and $\partial S_y / \partial y$ equal to zero, we obtain two "optimal lines" (so-called Cournot lines) in (x, y) space. Each individual will regulate his output to try to bring the points (x, y) on his own optimal line. The equations of these optimal lines are

$$\begin{aligned} L_x: px + qy &= \frac{p}{\beta} - 1, \\ L_y: qx + py &= \frac{p}{\beta} - 1. \end{aligned} \tag{57}$$

The intersection will be in the first quadrant if $p > \beta$, and the equilibrium will be stable if $p > q$. If the equilibrium is not stable, either X or Y will stop producing altogether and so become "parasitic" on the other, who must keep on producing to maximize his own utility in the absence of output by his partner. Which individual will become the parasite in this case depends on the initial conditions.

Thus the situation bears a formal resemblance to the two-population competition discussed in Sec. 1.6. The unstable case in the present model, which leads to parasitism, is analogous to the unstable case in the competition of populations, which leads to the extinction of one population.

However, the present example has another feature, which is not present in the population dynamics example, namely, the associated utilities. The specification of utilities enables us to determine how well each individual

does at the various possible outcomes: for example, at the stable equilibrium, if it exists, or as a parasite or as a "host," if he becomes one or the other. It is interesting to note that at the stable equilibrium neither of the two individuals does as well as he could at the "Pareto point," at which the joint payoff is maximized. But the Pareto point cannot be reached if each individual "tries" to maximize his own utility. It can be reached only if the two coordinate their outputs (for example, by a contract, in which each obligates himself to produce as much as the other) or if each attempts to maximize the joint payoff instead of his own. Even the parasite, who emerges in the unstable case and gets a considerably higher payoff than his host, can sometimes do better by *not* becoming a parasite but instead by maximizing the joint payoff (provided the other does the same). Whether this is so depends on the parameters p and β .

It turns out that X will be better off as a parasite than at the Pareto point if $\log(p\beta + qp - q\beta) - \log p + 1 - \beta > 0$. Taking into account that in the unstable case $q > p$, the inequality will hold if β is sufficiently small, the critical value depending on p .

Note that β measures the "reluctance to work." The qualitative conclusion, then, is the following: "It pays to be a parasite if the host is not too lazy." This sounds like a common-sense conclusion. But so does the statement, "It pays to be a parasite if you are sufficiently lazy." Since β was assumed equal for both individuals, the two common-sense conclusions are incompatible. The mathematical model decides between them.

Analogous situations appear in finite nonzero-sum games: for example, in the class of games called the Prisoner's Dilemma. Choices of strategy based on calculation of their own advantages leads the players to an outcome disadvantageous for both; choices of strategy based on calculation of joint interest lead to results advantageous to both.¹⁵

Experiments with two-person games of this sort indicate that if communication between players is not allowed the Nash point (analogous to a stable equilibrium in the continuous game described above) rather than the Pareto point is predominantly chosen (e.g., Scodel, Minas, Ratoosh, & Lipetz, 1959). Experiments simulating competing firms have yielded essentially similar results.¹⁶ A conclusion seems warranted that, at least

¹⁵ *Homo economicus* is assumed always to tend to maximize his own utility. However, utility is usually defined tautologically as the quantity that each individual attempts to maximize. In situations involving material payoffs it may be useful, in the context of a psychological theory, to separate the utility accruing from one's own payoffs and the vicarious utilities accruing from payoffs to others. Various weightings of these utilities, supposed to be summed, determine the "altruism vector" of an individual, and the set of such vectors determines the "altruism matrix" of a group. Use of this matrix in theoretical psychoeconomics was made by Rashevsky (1951) and by Rapoport (1956).

¹⁶ In many formalized situations of economic competition the intersection of Cournot lines (cf. p. 555) is analogous to the Nash point of a noncooperative nonzero-sum game.

in the cultures of the subjects in these experiments, choices guided by tacit mutual trust (which leads to the choice of the Pareto point) are, as a rule, not made. On the other hand, if communication and collusion are allowed, Pareto points are chosen with considerably greater frequency (Deutsch, 1958).

Now if there is a *unique* Pareto-optimal set of strategies and if the parties are inclined to effect an agreement by negotiation, we must expect the Pareto-optimal solution. (If they can agree at all, they can be expected to agree to do the best they can.) However, an interesting situation results if there are several Pareto-optimal solutions (or a continuum of such solutions) and, moreover, the interests of the players are directly opposed along this set: what one player wins, the other loses. Games of this sort lead to bargaining situations.

4.2 Bargaining

A typical bargaining situation is one in which two or more participants can each gain from an agreement entered into but in which there is a conflict of interest regarding the terms of agreement. Some would have it that bargaining is the prototype of all social interactions. The title of J. J. Rousseau's major work *Le Contract Social* attests to the influence on social philosophy of ideas stemming from economics (later explicitly formulated by Adam Smith and Ricardo).

Formal theories of bargaining are a modern development, an outgrowth of the theory of the so-called "cooperative" nonzero-sum game. It seems to me that the term "cooperative" is a misnomer in this context because it suggests that the interests of the players coincide. They do, as a matter of fact, *partially* coincide by virtue of the game being nonzero-sum, since this implies that in the set of outcomes there is a subset associated with a maximum joint payoff. If payoffs can be added and transferred conservatively from player to player (e.g., like money), it follows that it is in the players' joint interest to have an outcome with the maximum joint payoff. The term "cooperative," as it is used to designate such games, applies to the rules of the game which allow the players to communicate, that is, they presumably give them the opportunity to agree on such an optimum outcome. That this opportunity is not always utilized is well known, even if the outcome with the maximum joint payoff is unique. If there are several such outcomes, the difficulty of coming to an agreement is even more serious because in the choice of one of these outcomes the interests of the players often conflict. Indeed, this choice has the features of a constant-sum game, in which one player's gain is necessarily the

other's loss. I therefore prefer to call nonzero-sum games, in which communication (i.e., bargaining) is permitted, *negotiable* games.

A theory of such games constitutes a formal theory of bargaining. Like the theory of games, of which it is an extension, formal theories of bargaining are normative rather than descriptive. Typically, they are deduced from sets of axioms which reflect the features of "bargaining power" (e.g., the ability to make enforceable threats and promises) as well as certain equity considerations (usually conditions of symmetry or invariance of the outcomes with respect to the renaming of the players). The interested reader is referred to the researches of Nash (1950), Raiffa (1951), and Braithwaite (1955).

Experimental studies purporting to deal with bargaining situations are rapidly accumulating. However, many of them, although interesting in their own right, are not directly relevant to the formal theories of bargaining for two reasons. First, although the situations studied are clearly of the "mixed-motive" type, that is, involving partly coincident and partly conflicting interests of the participants, they do not aim at tests of explicit mathematical models of bargaining, such as are contained in the theoretical investigations we have discussed. They aim, rather, at tests of qualitatively stated hypotheses of interest to psychologists: for example, at comparison of outcomes under different imposed conditions. Second, many of these studies exclude direct communication between the participants and thus lack the principal feature of the bargaining situation. I suspect that this is done in the interest of simplicity of analysis: it is easier to record and to analyze formalized acts of the participants than a protocol of offers, counteroffers, threats, and promises.

To be sure it can be argued that *implicit* bargaining does occur if the same situation is repeated many times in succession because each participant can indicate to the other by his *acts* what he can be expected to do in response to the other's acts. Thus implicit offers, threats, and promises can be made and carried out.

An example of an implicit bargaining situation involving the distribution of priorities in the use of a one-way road by two "trucking companies" is given in Deutsch & Krauss (1960). The aim of this study was to test hypotheses concerning the effects of unilateral and bilateral "threats" on the outcomes, measured by the profits shown by the two firms who contend for the use of the one-way road, which is shorter than an alternate unimpeded road open to each company. Simultaneous attempts to use the short road compels one or the other to back up, thus losing time and profits. In some of the experimental conditions one or both of the firms can punish the other by blocking this road at one end, and the ability to do so constitutes the "threat."

The results show that when neither firm has the threat potential or when only one firm has it, thus being in control of the situation, both firms do better than if both can avail themselves of the threat. (The firms do not compete; each is instructed to maximize its own profits without regard for the profits of the other.)

In view of the existence of mathematical theories of economic behavior, simulations of classical economic situations (oligopolies, duopolies, etc.) offer greater opportunities for designing bargaining experiments along lines suggested by mathematical models. Experiments with oligopoly have been reported, among others, by Sauermann and Selten (1959). We shall examine in some detail the mathematical theory of the bilateral monopoly and a corresponding experiment reported by Siegel and Fouraker (1960).

4.3 Bilateral Monopoly

Consider a market in which there is only one seller and only one buyer. Assume that the buyer buys some manufactured product wholesale from the (only) seller and that he can sell this product in the retail market at a price that is determined by the demand for the product. The buyer's total profit, therefore, will be the difference between the market (demand) price and the price he pays to the seller, multiplied by the quantity of the product that the seller will sell him. The seller's profit, of course, will be the difference between the price he will receive from the buyer and the production costs, multiplied by this quantity.

The question before us is whether under these conditions, given the demand and the production schedules, the quantity sold by the seller to the buyer and the price paid by the buyer are determined by the economic situation alone, if it is assumed that each acts to maximize his total profit.

To fix ideas, assume that the retail demand price r decreases linearly with the quantity Q offered:

$$r = A - BQ, \quad (58)$$

whereas the production cost (per unit) c increases linearly with quantity produced, assumed the same as the quantity offered:

$$c = A' + B'Q. \quad (59)$$

The straight lines represented by Eqs. 58 and 59 converge as Q increases. Therefore the margin between production cost and demand price becomes smaller with larger Q . However, the *total* profit to seller and buyer combined is the distance between these two lines multiplied by the quantity

produced, sold, and resold. At some value of Q the combined profit accruing to both seller and buyer will be maximal. Let us compute this optimal quantity from the combined standpoint of the two individuals of this bilateral monopoly.

This total profit is obtained by maximizing the difference between total production cost and total retail sales, namely, letting $R = rQ$, $C = cQ$,

$$R - C = AQ - BQ^2 - A'Q - B'Q^2. \quad (60)$$

Setting the derivative of Eq. 60 equal to zero, we get

$$A - 2BQ - A' - 2B'Q = 0. \quad (61)$$

Solving for Q , we get

$$Q = \frac{A - A'}{2B + 2B'}. \quad (62)$$

It appears, therefore, that if the buyer and seller are to maximize their *joint* profit they should agree that the quantity to be produced and put on the market should be given by the value of Q in Eq. 62.

However, having agreed on the quantity, how are they to determine the price that the buyer will pay the seller? Obviously, the higher the price, the greater the share of the joint profit that will accrue to the seller but also the smaller the share of the buyer. There are no constraints on the system to fix the price at some point between the production cost and the demand price. Therefore the model does not lead to a determinate solution with regard to the price to be paid to the seller, although it does lead to a determinate solution with regard to the quantity to be sold.

Suppose, now, the buyer announces that he will pay the price p per unit. The seller, being also the producer, can then decide what quantity he will be willing to produce and sell at that price. The seller will wish to maximize $P - C$, where $P = pQ$. Since he controls Q , he will do so by differentiating $(P - C)$ with respect to Q and setting the derivative equal to zero. But from Eq. 59 we get

$$C = A'Q + B'Q^2, \quad (63)$$

$$\frac{dC}{dQ} = A' + 2B'Q.$$

Hence

$$\frac{d}{dQ}(pQ - C) = p - A' - 2B'Q = 0,$$

$$Q = \frac{p - A'}{2B'}. \quad (64)$$

Now the profit accruing to the buyer is

$$(r - p)Q = \left[A - \frac{B(p - A')}{2B'} - p \right] \frac{p - A'}{2B'}, \quad (65)$$

and the buyer will wish to maximize this quantity with respect to p , the quantity he presumably controls. Setting the derivative of Eq. 65 equal to zero, we obtain, after simplifying

$$p^* = \frac{AB' + A'B + A'B'}{B + 2B'}, \quad (66)$$

where p^* is apparently the price that should be quoted by the buyer to ensure the maximum profit for himself under the constraint of the seller's control of the quantity to be produced.

Substituting Eq. 66 into Eq. 64, we obtain the value of Q^* , the quantity determined by p^* , namely,

$$Q^* = \frac{A - A'}{2(B + 2B')}. \quad (67)$$

We see that Q^* does not correspond to the Q found by maximizing the joint profit.

Let us now see what profits accrue to the seller and to the buyer, respectively, if $p = p^*$ and $Q = Q^*$.

We have for the seller's profit

$$\begin{aligned} \pi_s &= (p^* - c)Q^* \\ &= \left[\frac{AB' + A'B + A'B'}{B + 2B'} - A' - \frac{B'(A - A')}{2(B + 2B')} \right] \frac{A - A'}{2(B + 2B')} \\ &= B' \left[\frac{A - A'}{2(B + 2B')} \right]^2. \end{aligned} \quad (68)$$

Similarly, we have for the buyer's profit

$$\pi_b = (r - p^*)Q^* = (B + 2B') \left[\frac{A - A'}{2(B + 2B')} \right]^2, \quad (69)$$

and accordingly, for the joint profit

$$\pi = \pi_s + \pi_b = (B + 3B') \left[\frac{A - A'}{2(B + 2B')} \right]^2. \quad (70)$$

Under these circumstances, we see that the buyer's share of the total profit (the buyer being the price leader) will amount to

$$\frac{\pi_b}{\pi_s + \pi_b} = \frac{B + 2B'}{B + 3B'}. \quad (71)$$

His share, therefore, will be at least $\frac{2}{3}$ (if B' is very large compared to B) and will approach unity if B is very large compared to B' .

The advantage is obviously with the price leader. We should not conclude, however, that the price leader will do best for himself under all circumstances by acting as price leader. Under certain circumstances, he would do better by negotiating a settlement with the seller, even to the extent of offering him the greater share of the joint profit. This is because negotiation makes possible the optimization of Q . Assume, in fact, that Q has been optimized to give the greatest joint profit. Then Q is given by Eq. 62 and the joint profit by

$$(r - c)Q = \frac{(A - A')^2}{4(B + B')^2}. \quad (72)$$

As price leader, the buyer can command the quantity given by Eq. 69 as his share of the total profit. Therefore he can afford to negotiate if he can get a fraction of the joint (maximized) profit, which amounts to at least

$$\frac{(B + B')^2}{(B + 2B')^2}. \quad (73)$$

This fraction is less than $\frac{1}{2}$ if $B/B' < \sqrt{2}$.

In other words, the buyer as price leader is in a position to offer the seller a greater share in the (maximized) joint profit if the slope of the demand curve is sufficiently smaller (by a factor of $\sqrt{2}$) than that of the supply curve. Recall the analogous situation with the two producers who share their product (Sec. 4.1). As in the present case, each can command a certain return at the intersection of the optimal lines. Therefore each is in a position to offer the other a somewhat greater share of the total utility (assuming the utility is transferable) in negotiating an agreement to maximize joint utility. If the equilibrium is unstable, the parasite can actually be better off (provided the parameters are within a certain range) receiving one half of the joint (maximum) payoff than getting his payoff as parasite. These findings concern an aspect of "bargaining power" not often emphasized. The usual emphasis is on bargaining power derived from being in a position to *threaten* the opponent with dire consequences if the terms are not accepted. But there is also bargaining power derived from being in a position to *promise* the opponent certain advantages if he goes along with a proposal.

Let us turn to some experimental data that have a bearing on the foregoing theoretical discussion of bilateral monopoly.

In the experiments described by Siegel and Fouraker (1960), pairs of subjects took the respective parts of a buyer and a seller in a simulated

bilateral monopoly situation. The object of the experiment was to determine which, if any, of the theoretical positions with regard to the outcomes of a bilateral monopoly would be corroborated in the simulated situation. The question is an interesting one, inasmuch as the theoretical positions of various economists have been quite distinct. The "solution" offered by Cournot on the basis of assuming a "price leader" was derived on p. 553 (Eqs. 66 and 67). In symmetric bargaining neither has the privilege of being a price leader. Thus the Cournot solution, which determines both price and quantity, does not apply. The opinions of economists regarding the expected outcome have been divided. They fall into three categories:

1. Neither the quantity nor the price (to the buyer) are determined by the economic factors. Other determinants must be known to predict the outcome of a specific situation (Bowley, 1928).
2. The quantity is determined à la Pareto, that is, by the maximization of joint profit. But the price is not determinate and will depend on factors not included in the model (e.g., psychological factors, reflecting the bargaining abilities of the participants) (Stigler, 1952).
3. Both quantity and price are determinate, quantity by maximization of joint profit and price by some bargaining principle, such as perceived symmetry of the situation (Schelling, 1960) and the intersection of the marginal revenue and marginal cost lines (Fouraker, 1957), or by some other bargaining principle derived from some plausible set of axioms (Nash, 1950; Raiffa, 1951).

The experiments were conducted under conditions of symmetric bargaining. The first bidder was chosen by lot from each pair, and the bidding was in terms of offers and counteroffers of prices paired with quantities. In different experiments certain supplementary conditions were varied in order to note their effects on outcomes. For example, the members of bargaining pairs could be informed either only of their own payoff schedules, that is, the supply-cost (or demand-price) curves, or of both. The maxima of joint profit could be relatively sharp or relatively flat. Finally, "levels of aspiration" could be induced in the bargainers by offers of incentives if they won for themselves an indicated minimum profit.

The results definitely corroborate the hypothesis that in symmetric bargaining the quantity agreed on is determinate and is chosen to maximize joint profit but that the division ratio of the profit depends on factors outside the economic model: for example, on the information possessed by the bargainers and on the induced levels of aspiration.

In every experiment, in which several pairs of subjects were involved, the quantities agreed on clustered closely around the profit-maximizing

quantity, whereas the prices agreed on were spread out along the "negotiation set." Increasing the information available to the bargainers and increasing the rate of decline of joint profit as one moves away from the maximum (i.e., "sharpening" the maximum) had the effect of decreasing (sometimes to zero) the variance of the quantity agreed on. Inducing different levels of aspiration in the bargainers (offering additional incentives if certain minimum profits were secured) produced unmistakable biases in the price agreed on—the bargainer with a higher aspiration came off with the greater share of the profit.

In this way the roles of both economic and psychological factors were separately and quantitatively demonstrated in a controlled bargaining experiment, in which bargaining was restricted to strictly formalized successive offers.

4.4 Formal Experimental Games

The relevance of psychological factors in bargaining situations suggests that psychoeconomics may well become another of the "border regions" (like social psychology, psychophysics, psycholinguistics) of behavioral science in which methods of more than one discipline are fused in forging the exploratory tools. Aside from the psychology of bargaining (of obvious importance in the study of social interactions), psychological considerations can be expected to be relevant wherever individuals are put into situations in which they perceive their interests to be divergent. Bargaining situations are special instances in which negotiations can take place. Of equal interest from the psychological point of view are situations in which explicit negotiations are impossible. In the literature of game theory such situations are called noncooperative games. Here they are called nonnegotiable games.

Although some psychological factors relevant to bargaining are obviously irrelevant in nonnegotiable games, other factors play a role perhaps equally essential. It is important to keep in mind that game theory, at least in its original formulation, completely bypassed psychological factors. All information available to the players by the rules of the game was assumed to be utilized; utility values of the outcomes were assumed to be given; the best available strategies were assumed always to be chosen, etc. It goes without saying that the application of game theory to *behavioral* science requires the introduction of psychological parameters, since human memory is not perfect, human decisions are not always "rational," etc. In principle, such parameters could be introduced to extend and to generalize game theory, and some work along these lines has been done.

Another, more purely empirical approach is taken by some experimenters who set up situations suggested by game theory with a view of recording any regularities that may be found relating the observed behavior to normatively prescribed "solutions" of game theory: for example, in experiments with zero-sum games or with n -person games in characteristic function form. Sometimes this cannot be done simply because game theory fails to prescribe even a normative solution, or a class of solutions, most notably in nonzero-sum, nonnegotiable games. Nevertheless, the behavior of people in situations isomorphic to such games is of great interest for what it may reveal about the underlying psychology.

The program of the empirical approach is an old-fashioned one. A good model of "rational behavior" for nonzero-sum, nonnegotiable games does not exist, so it is proposed simply to gather great quantities of data on actual behavior in such situations in the hope that regularities discovered in the data can suggest models to be tested by further experiments, for example, by varying experimentally controlled parameters.

The Prisoner's Dilemma is especially intriguing in investigations of this kind because it puts the players into a situation in which "collective rationality" (and its underlying assumption that the partner is also motivated by it) comes in conflict with "self-interest rationality" (and its underlying assumption that the partner is also motivated by it).

The essential feature of the Prisoner's Dilemma game is the choice open to each player, namely, to conform, that is, to play the "cooperative strategy" which, if played by both, rewards both; or to defect, that is, to play the "noncooperative strategy" which, if chosen by both, punishes both. If the two players choose different strategies, the conformist is punished more severely and the defector is rewarded more generously than if either had chosen the same strategy as his partner.

In all experiments with nonnegotiable Prisoner's Dilemma-type games, both the cooperative and the noncooperative strategies are chosen, to be sure with different frequencies in different games and under different conditions. The question naturally arises regarding the nature of the conditions that influence the propensity to conform or to defect.

Some recent studies indicate several such dependencies. Deutsch (1958) has investigated the propensity to make the cooperative ("trusting") choice that leads to the Pareto-optimal outcome in a Prisoner's Dilemma game, as it is influenced by the instructions given to the players, namely, a "cooperative orientation" (having joint payoffs in mind), "individualistic orientation" (having only one's own payoff in mind), and a "competitive orientation" (having the difference of the payoffs in mind). The propensity to choose cooperatively has been found to change in the expected direction with the instructions. The same author also investigated the effect of

the opportunity to negotiate and found that this also significantly increases cooperation.

Scodel, Minas, Ratoosh, and Lipetz (1959) have found that cooperation tends to decrease (or competition to increase) in the course of a run of 50 plays of the same nonzero-sum game. They also found evidence that the competitive motive plays an important part in the subjects' choices (even with neutral instructions), since even in the games in which no advantage accrued to the single defector, as many as 50 per cent noncooperative choices were made.

Lutzker (1960) found a higher propensity to cooperate among subjects rated high on an "internationalist" attitude scale, compared with subjects high on an "isolationist" attitude scale. A control group of unselected subjects was not significantly different from the "internationalists," but their cooperative choices tended to decrease in the course of the run, as did those of the "isolationists," whereas the frequency of cooperative choices of the "internationalists" showed no such trend. Deutsch (1960) found similar differences between subjects with extreme opposite ratings on the F ("authoritarian") scale.

Experiments with three-person, nonzero-sum, nonnegotiable games were undertaken at the University of Michigan. The main purpose was to observe the dependence of the over-all frequency of cooperative choices, f , on the payoff matrices. Accordingly, all games were played under presumably the same conditions. The instructions were approximately the same as the individualistic instructions in Deutsch's experiments. Communication was not allowed. Subjects played for one mill per point, the winnings or losses being added or subtracted from their subjects' fees.

One further feature was added to equalize the conditions in which each game was played, in particular, to avoid progressive learning from one game to another. A trio of players played eight games "at once" in each session; that is, the games were presented in randomized order. In this way no game was in any special position in the order of presentation, even though the same trio of subjects played eight different games.

There were two such sets of experimental runs. The first involved 16 three-person groups and the seven games shown in Table 3, the total number of plays ranging from 300 to 500. The second involved 12 three-person groups and the eight games shown in Table 4, with 800 plays in each session.

The frequency of cooperative choice f (averaged over all individuals in all plays of each game) was the dependent variable. It was recorded for each game so that the games could be arranged in a sequence in which f decreased monotonically.

The problem was to choose an independent variable, that is, an index,

Table 3 Seven Games Played in the First Set of Experiments. Game Four of a Different Type, Introduced as a Control, Is Omitted. The First Column Indicates the Triples of Choices of Right or Left Button by the Three Players, *A, B, C*. The Matrix Entries are the Corresponding Payoffs to the Three Players, Respectively

| <i>A B C</i> | I | II | III | V | VI | VII | VIII |
|--------------|----------|----------|----------|----------|----------|----------|----------|
| <i>R R R</i> | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 | -2 -2 -2 | 1 1 1 | 1 1 1 |
| <i>R R L</i> | 2 2 -2 | -2 -2 6 | 2 2 -2 | -1 -1 3 | -1 -1 1 | 1 1 -1 | -2 -2 3 |
| <i>R L R</i> | 2 -2 2 | -2 6 -2 | 2 -2 2 | -1 3 -1 | -1 1 -1 | 1 -1 1 | -2 3 -2 |
| <i>R L L</i> | 3 -2 -2 | 1 -2 -2 | -2 1 1 | -1 2 2 | 6 -1 -1 | -3 3 3 | 1 -2 -2 |
| <i>L L L</i> | 1 1 1 | -1 -1 -1 | 1 1 1 | -2 -2 -2 | 1 1 1 | -2 -2 -2 | -1 -1 -1 |
| <i>L L R</i> | -2 -2 3 | -2 -2 1 | 1 1 -2 | 2 2 -1 | -1 -1 6 | 3 3 -3 | -2 -2 1 |
| <i>L R L</i> | -2 3 -2 | -2 1 -2 | 1 -2 1 | 2 -1 2 | -1 6 -1 | 3 -3 3 | -2 1 -2 |
| <i>L R R</i> | -2 2 2 | 6 -2 -2 | -2 2 2 | 3 -1 -1 | 1 -1 -1 | -1 1 1 | 3 -2 -2 |

Table 4 Eight Games Played in the Second Set of Experiments

| <i>A B C</i> | IX | X | XI | XII | XIII | XIV | XV | XVI |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|
| <i>R R R</i> | 1 1 1 | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 | -3 -3 -3 |
| <i>R R L</i> | -2 -2 2 | 4 4 -2 | -2 -2 6 | 2 2 -4 | -6 -6 2 | 4 4 -4 | -6 -6 6 | 6 6 -6 |
| <i>R L R</i> | -2 2 -2 | 4 -2 4 | -2 6 -2 | 2 -4 2 | -6 2 -6 | 4 -4 4 | -6 6 -6 | 6 -6 6 |
| <i>R L L</i> | -2 2 2 | 4 -2 -2 | -2 6 6 | 2 -4 -4 | -6 2 2 | 4 -4 -4 | -6 6 6 | 6 -6 -6 |
| <i>L L L</i> | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 | -1 -1 -1 | 1 1 1 |
| <i>L L R</i> | 2 2 -2 | -2 -2 4 | 6 6 -2 | -4 -4 2 | 2 2 -6 | -4 -4 4 | 6 6 -6 | -6 -6 6 |
| <i>L R L</i> | 2 -2 2 | -2 4 -2 | 6 -2 6 | -4 2 -4 | 2 -6 2 | -4 4 -4 | 6 -6 6 | -6 6 -6 |
| <i>L R R</i> | 2 -2 -2 | -2 4 4 | 6 -2 -2 | -4 2 2 | 2 -6 -6 | -4 4 4 | 6 -6 -6 | -6 6 6 |

derived from the game matrix, against which f could be plotted, preferably an index of which f would be a linear function. Several such indices suggested themselves:

1. COMPARISON OF EXPECTED GAIN. Assume that each player views the four possible combinations of strategy choices by the other two players as four equiprobable "states of nature." Then, he compares his expected gains from his own cooperative and noncooperative choices. The algebraic difference constitutes the "cooperative index" of the game. (In all games the cooperative index was negative.)

2. COMPARISON OF PAYOFFS TO SELF AND OTHER. Except where the choice of strategy is unanimous, there are two payoffs associated with each outcome, the payoff to the defector and the payoff to the conformist. We assume that the player compares his payoff with that of the conformist, if he himself is a defector, and vice versa. The first of these differences is called the relative advantage of being a defector; the second the relative advantage of being a conformist. The second minus the first is the cooperative index of the game. Like the previous index, it is always negative.

3. COMPETITIVE ADVANTAGE OVER THE OTHER TWO PLAYERS. This criterion is like criterion 2, except that the comparison is made not between the "roles" but between payoff to self and *average* payoff to the other two.

4. MINIMAX. Each player assumes that he is playing a 2×4 game against a single opponent and chooses the strategy in which his possible loss is minimized. This calls for mixed strategies in Games I, V, and VII and in all the games of the second series. The minimax strategy so conceived should not be confused with the Nash equilibrium point, which is a *pure* strategy in Games I and V and all the games of the second series.

Of these criteria, criterion 3 yielded an index most closely correlated with the value of f . The regression line of the plot of f against i , the competitive advantage index, computed in arbitrary units is shown in Fig. 13.

We note that the regression line has approximately the same slope for 10 of the 15 games. Beyond that range, however, the regression line breaks. The cooperation frequency f still diminishes with the absolute value of the (negative) cooperative index but at a much slower rate. If the decrease had continued at the same rate, we would have expected all instances of cooperative choice to disappear at about $i = -30$. However, as much as 8% cooperative choices remain at $i = -48$.

The source of these cooperative choices can easily be seen if the records are examined. In every experimental run there is at least one individual who, apparently discouraged by continued losses resulting from unanimous

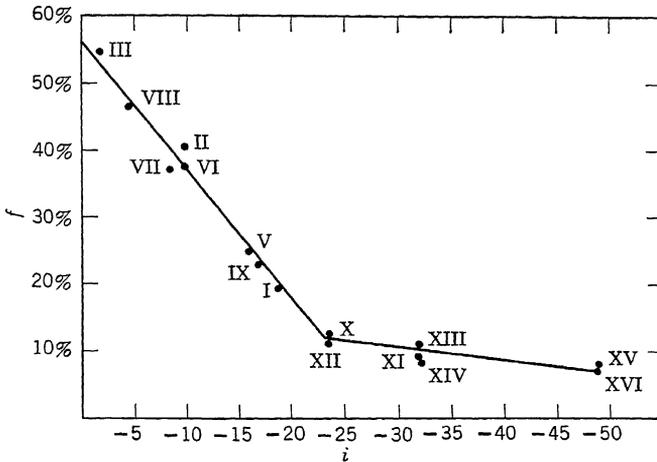


Fig. 13. Plot of f (over-all frequency of cooperative choice) against i (index for the 15 games listed in Tables 3 and 4, computed by Criterion 3, p. 560). Roman numerals indicate the positions of the games.

defection, now and then plays the cooperative choice many times in succession (sometimes 20 to 30 times), probably in an attempt to arouse in the others a sense of social responsibility to join in a collusion with him "against the house." Almost always he failed. Since these long stretches of cooperative choices are made indiscriminately, that is, with no regard for which game is being played, we have the "irreducible" residue of cooperative choices even in the games with high negative indices. Hope springs eternal in the human breast!

The static theory just described is, of course, a rather shallow one. Besides the *ad hoc* explanations of the way f varies from game to game, it has no additional predictive value. For example, nothing can be concluded from the question whether the f of a game depends only on the game itself or also on what other games are "mixed" with it in the same experimental run. No conclusion can be drawn concerning whether the cooperative choices of the three players in a given run are statistically independent, etc.

In order to form a theory with somewhat more depth, which would predict the present results and others too, we can seek some underlying process of which the observed data would be consequences. In other words, we seek a dynamic theory governing the time course of a process consisting of a sequence of states in which the subjects find themselves. A theory of this kind is examined in Sec. 5.

5. GROUP DYNAMICS

Group dynamics is a branch of social psychology in which the small (face-to-face) human group (e.g., a work group, a friendship circle, a family) is the object of study. Typically, workers in group dynamics view the group as a "system" or an "organism" (an expected development in a society in which a large portion of decisions is made by committees). Research is directed toward the discovery of regularities in the behavior of such systems. As in any theoretical approach, certain events and processes are singled out for study, and certain hypothetical constructs are offered with a view of providing economical descriptions of the events observed. If regularities are noted in the descriptions, hypotheses are suggested in which the theoretical constructs serve as variables. The hypotheses then become assertions about relations among these variables. In particular, if the assertions are related to the time courses of the variables, the resulting theory becomes "dynamic" in the strict sense.

At a certain stage of theory construction, in which quantitative observations are made but are not yet mathematicized, psychologists frequently state hypotheses in the form "the more . . . the more . . ." or "the more . . . the less . . ."; that is, they make assertions about direct or inverse relations among variables without specifying more definitely the nature of the functional relations. If the exact nature of the implied quantitative relations were specified, say by a mathematical function, we would have a mathematical model. A crucial difficulty in constructing a *verifiable* model of this sort relates to the lack of naturally suggested scales for the variables involved. Workers in group dynamics speak of "intensity of interaction" among the members of a group, of the "level of friendliness," of the "amount of activity," of "pressures" exerted on members from co-members and from outside, etc. Obviously, even hypotheses in the form "the more . . . the more" can be tested only when operational definitions of these "variables" are given. Mathematically stated hypotheses require such definitions to be appropriate to the type of model. For example, if explicit mathematical functions enter the equations, the variables must be expressible in a ratio scale (with zero point and unit precisely defined.)

It is not at all clear a priori how "the level of friendliness" in a group is to be measured. However, such a measure is not unthinkable. It is, in fact, quite easy to *suggest* measures. We could, for example, take for the "level of friendliness" the relative frequency of utterances of a certain type observed in the group: for example, those labelled "supportive" or "tension reducing" in the system worked out by Bales (1950). In a task group involved in a situation in which choices of cooperative or noncooperative acts must be made (cf. Sec. 4.4), the "level of friendliness" might be defined by the relative frequency of cooperative choices.

Thus it is quite possible to render operational the variables proposed by the group dynamicists. Of course, each of the variables can be defined in several different ways, none of them a priori more justifiable than others. We can only hope that justification of particular definitions can be made a posteriori, that is, in view of the fruitfulness of the models constructed with them.

It is also clear that the mathematical model builder need not concern himself with the justification of the definitions nor even with the definitions themselves to the extent that his job is to derive the consequences of the model he has constructed. Nevertheless, the model builder may be guided in the construction of the model by the content of the social psychologist's hypotheses. Presumably, Richardson was so guided (cf. Sec. 1.2) when he *translated* various verbally stated hypotheses on the causes of arms races into mathematical assertions.

A similar "translation" was undertaken by Simon (1957) when he formulated a number of mathematical models on the basis of hypotheses stated verbally in the literature of social psychology.

5.1 A "Classical" Model of Group Dynamics

One set of postulates, which Simon translated into differential equations, that derives from the work of Homans (1950) is as follows:

1. The intensity of interaction in a group increases with the degree of friendliness and with the degree of activity.

2. There is a level of friendliness "appropriate" to a corresponding activity level. The actual level of friendliness tends to this appropriate level at a rate proportional to the amount of departure from it.

3. Postulate 2 relates also to the rate at which "activity" tends to an appropriate level. This level depends on the level of friendliness and on the amount of activity imposed on the group externally (say, by a task).

The simplest "translation" of these postulates into mathematics is in terms of a system of linear algebraic and differential equations. If I , F , A , and E represent, respectively, intensity of interaction, friendliness, level of activity, and externally imposed activity (all functions of time), we have

$$I = a_1F + a_2A, \quad (74)$$

$$\frac{dF}{dt} = b(I - \beta F), \quad (75)$$

$$\frac{dA}{dt} = c_1(F - \gamma A) + c_2(E - A), \quad (76)$$

where all the coefficients are positive constants.

The system being linear, a general solution can be obtained, giving the variables as functions of time, of the parameters (the coefficients), and of the initial conditions. Clearly, the solution can be tested only if the variables can be appropriately measured. Even so, the large number of free parameters almost destroys the usefulness of an explicit solution. We therefore seek information of more general nature, for example, the static (equilibrium) properties of the system. Particularly, the conditions of stability, as we have seen (cf. Secs. 1.6 and 4.1), play an important part in all theories of this sort.

Setting the derivatives in Eqs. 75 and 76 equal to zero, we obtain expressions for each of the group variables I , A , and F in terms of the externally imposed activity. We also obtain in the usual way the conditions of stability for the system. These turn out to be

$$c_1\gamma + c_2 + b(\beta - a_1) > 0, \quad (77)$$

$$(\beta - a_1)(c_1\gamma + c_2) - a_2c_1 > 0. \quad (78)$$

We see that $\beta > a_1$ is necessary to satisfy Eq. 78 and sufficient to satisfy Eq. 77. Translated into words, if the system is to be stabilized, the coefficient β , which regulates the rate of change of F , as F departs from the level "appropriate" to a given intensity of interaction, should be greater than the coefficient a_1 , the proportionality factor that relates friendliness to intensity (in the absence of activity).

The mathematical model thus provides some leverage for a theory of interactions in a group. Questions of interpretation inevitably arise. If the system is unstable and F becomes negatively infinite, a reasonable interpretation is a dissolution of the group (or a brawl?). It is admittedly difficult to interpret an unlimited growth of F . However, we can always limit the meaningfulness of a model to a certain limited range of values of its variables.

Denote by I^* , F^* , A^* , and E^* the values of the corresponding variables at equilibrium. If equilibrium does obtain, the condition $dF^*/dE^* > 0$ can be deduced from the model. Furthermore, I^* , F^* , and A^* vanish with E^* . This is in accord with Homans' explanation of social disintegration on community and family levels resulting from the disappearance of externally imposed activities (e.g., with unemployment and atrophy of the economic functions of the family).

The relations $A^* > E^*$ and $A^* \leq E^*$ can be interpreted as positive or negative "morale." The conditions for positive or negative morale (while equilibrium is preserved) can also be obtained in terms of the relations between coefficients and appropriately interpreted in social-psychological terms.

If the restriction of linearity is dropped, general solutions of the differential equations are no longer available. In this case we proceed with

the investigation of the phase space, exactly as was done with population dynamics (cf. Sec. 1.6) and analogous problems.

5.2 A Semiquantitative Model

The variables examined in Sec. 5.1 related to the activity and the socio-emotional atmosphere in the group as a whole. If we inquire on what these variables, in turn, depend, we come upon concepts relating to the interactions among the group members. Examples of such concepts are the degree of unanimity or discord, receptiveness of members to each other's communications, interpersonal attractiveness, etc. These terms appear in group dynamics theory in contexts of more or less rigorous discussion. In particular, Festinger (1950) defines the following:

D: The perceived discrepancy of opinion among the members on an issue.

P: Pressure to communicate with each other.

C: Cohesiveness, that is, average (or total) strength of attractiveness among the members.

U: Pressure to achieve uniformity of opinion.

R: Relevance of the issue to the group.

We note that these terms, like those considered in Sec. 5.1, are aggregative; that is, they pertain to the whole group rather than to the individual members. However, they seem to be derived from a more detailed analysis. For example, we may well consider all of them as determinants of *F*, the over-all "friendliness," or of *I*, the intensity of interaction, discussed in Sec. 5.1.

Festinger's hypotheses are statements about the interdependence of the variables denoted by the terms. Simon (1957) translates these hypotheses in the same way as those of Homans (cf. Sec. 5.1) into mathematical statements, namely,

$$\frac{dD}{dt} = f(P, L, D), \quad (79)$$

$$P(t) = P(D, U), \quad (80)$$

$$L(t) = L(U), \quad (81)$$

$$\frac{dC}{dt} = g(D, U, C), \quad (82)$$

$$U(t) = U(C, R), \quad (83)$$

$$\frac{dR}{dt} = 0. \quad (84)$$

Equations 79 to 84 are weaker than Eqs. 74 to 76. The specific form of the functions on the right is not given. Naturally, the consequences to be derived will also be much weaker.

Next, we note that some of the equations involve derivatives and others do not. The latter imply an "instantaneous" adjustment of the dependent variable values to those of the independent variables, whereas the former imply *rates* of adjustment. A direct dependence of the variable on the left on those on the right obtains in the latter case only at equilibrium (if it is ever attained). Equation 84 says simply that R is a constant in a given situation, determined, say, by the topic under discussion by a group.

In Festinger's treatment the specific dependencies indicated in Eqs. 79 to 84 are stated in typical semiquantitative ("directional") form prevalent in investigations in which attempts at quantification but no attempts at mathematization have been begun.

1. The pressure on group members to communicate increases with increasing perceived discrepancy of opinion, with the degree of relevance of the issue in question, and with the pressure toward uniformity.

2. The amount of change in opinion resulting from a received communication increases with the pressure toward uniformity and with the cohesiveness related to the recipient.

3. The rate of change of cohesiveness suffers decrements (i.e., becomes smaller if positive and larger in absolute value if negative) as either perceived discrepancy or pressure toward uniformity increases. This rate depends also on the level of cohesiveness.

The last hypothesis relates to the changes in the mutual attractiveness among the members as they depend on discrepancies of opinion and on the importance to the group to preserve uniformity. Simon (1957) introduces this hypothesis in addition to those formally stated by Festinger in order "to make the dynamic system complete." He argues that in the interpretation of some empirical studies this hypothesis is actually implicit.

These "directional" hypotheses can be formalized by statements about the signs of partial derivatives of the functions on the right side of Eqs. 79 to 84. Thus P_D , L_U , U_C , P_U , and U_R are implied to be positive by the hypotheses, whereas f_P , g_D , g_U , and f_L are implied to be negative. (The subscripts indicate variables with respect to which the partial derivatives are to be taken.) Thus the signs of 9 of the 11 partial derivatives are implied by the verbally stated hypotheses. Two others remain, namely, f_D and g_C . To determine their signs, Simon examined the situation in equilibrium for a given value of R , that is, when dD/dt and dC/dt are equal to zero. Using the chain rule for differentiation, we obtain

$$f_P \delta P + f_L \delta L + f_D \delta D = 0, \quad (85)$$

$$g_D \delta D + g_U \delta U + g_C \delta C = 0. \quad (86)$$

If the system moves from one equilibrium to another (say, as the independent experimentally imposed parameter R changes in value) Eqs. 85 and 86 must hold throughout, quite analogously to the situation in thermodynamics in which a system goes through a sequence of reversible states. Also, if P and L are large, Eq. 79 implies that D will be pushed to a lower equilibrium level. Hence $\delta D/\delta P = -(f_P/f_D) < 0$ and $\delta D/\delta L = -(f_L/f_D) < 0$, from which we deduce $f_D < 0$. By a similar argument, Simon showed $g_C > 0$.

This completes the mathematical analysis of Festinger's model. Experimental studies can now be examined with a view to decide whether the results are relevant to the deduced predictions, and, if so, whether the predictions are corroborated or refuted. It goes without saying that experimental data are relevant to the model only to the extent that a method of measuring quantities involved is indicated. The "weakness" of the present mathematical model, however, necessitates only weak measurement scales. In fact, since only relative magnitudes are compared, only an ordinal scale of measurement is required.

Indices for such scales have been offered by many researchers. Simon analyzed the data obtained in an experiment by Back (1951) and in the field by Festinger, Schachter, and Back (1950) in the light of his mathematical interpretation of the proposed theory.

5.3 Markov Chain Models

The usefulness of the "classical" models, discussed in Sec. 5.1 and 5.2, is severely limited by difficulties of measuring the quantities represented in them. In recent years mathematical theories of group dynamics have been developing along quite different lines. Since most of these developments are being pursued by workers whose principal orientation is mathematical, the central variables of the models are characteristically not indices of "psychological states" of special interest to group dynamicists, for example, friendliness, cohesion, and rejection, but rather whatever happens to be easily and obviously quantifiable, such as easily identifiable acts, which can be quantified in terms of temporal or relative frequency.

Already in learning theory, this focusing of interest on countable acts has resulted in a rapid and fruitful development of mathematical models of the learning process, which, in some instances, have received strong corroboration. Mathematical group dynamics is basically an extension of similar methods to situations in which interactions among individuals are the central acts, and this provides justification for calling this area of research "group dynamics." It is dynamics because it involves the study of the time courses of processes; it is group dynamics because the events

The equilibrium is a dynamic one: the system still passes from one state to another, but in any long period of time the fraction of time that the system spends in any one state remains constant.

Elsewhere in this volume (cf. Chapter 10) it is shown how the Markov chain is applied to a stochastic theory of learning.

In the simplest stochastic learning model, the following assumptions are made:

1. The subject has a choice of two responses, A_1 and A_2 , to a single stimulus.
2. Regardless of the response is given, A_1 is reinforced (declared to be correct) with probability π and A_2 , with probability $1 - \pi$ (noncontingent reinforcement).
3. If A_1 is reinforced, the stimulus will become conditioned to A_1 with probability θ and will retain whatever conditioning it has had with probability $1 - \theta$. The situation is similar with respect to A_2 .

The "state" of the subject at time t is defined by the response to which the stimulus is conditioned at time t . That response is given on the next stimulus presented. The probabilities π and θ determine the matrix of transition probabilities among all the pairs of the two states, 1 and 2. Thus:

$$\begin{array}{c} 1 \\ 2 \end{array} \left[\begin{array}{cc} \theta\pi + (1 - \theta) & \theta(1 - \pi) \\ \theta\pi & 1 - \theta\pi \end{array} \right].$$

This matrix defines the Markov chain, which determines the entire time course of the process.

More involved models result if several stimulus elements are associated with each stimulus and a "sampled" element of this set is conditioned to a response at each presentation. Further generalizations involve larger number of stimuli, contingent reinforcement schedules, etc.

THE "TWO-HEADED SUBJECT." Burke (1959) extended the stochastic learning model to the case of what amounts to a "two-headed subject"; that is to say, the "subject" is a pair of persons in a stochastic reinforcement learning situation in which the reinforcements are in general contingent on what *both* subjects do. The learning processes of the two are thus inter-linked, and a kind of social interaction has been introduced into the learning situation.

A similar study is reported by Hays and Bush (1954).

DOMINANCE STRUCTURES. In our treatment of dominance structures (cf. Sec. 3.4), the states of a social group could be taken as particular

dominance configurations. If the dominance structure changes in time, we have a dynamic process. The corresponding Markov chain would involve the probabilities of the various dominance configurations and the transition probabilities from one configuration to another. The transition probabilities could be compounded of the probabilities of contact between pairs of individuals in the group and the probabilities of dominance reversals between them, if contact occurs, thus leading from one dominance structure to another. The equilibrium distribution would then represent the relative frequencies with which the different dominance structures would be expected to be observed. In Rapoport's treatment (1949) these transition probabilities are taken constant, and so the process becomes a Markov chain (see also Landau, 1953).

If the transition probabilities are such that reversals of dominance between two individuals with sufficiently disparate score structures become very rare, it is shown that the equilibrium dominance structure approaches a hierarchy. This is the mathematical statement of Landau's conclusion (Landau, 1951) that social factors (as distinguished from inherent biological ones) must be assumed to account for the near-hierarchies observed in moderately large flocks of hens. This *sociological* assumption (making transition probabilities depend on disparities of social rank) still allows the dynamics of dominance structure to be treated as a Markov chain. If the assumption were replaced by a *psychological* one (e.g., making the transition probabilities depend on the past histories of the particular individuals involved in contacts), the model would cease to be a Markov chain. We have, thus, an example of how the gross sociological assumptions lead to simpler mathematical models than do the finer psychological ones.

CONFORMITY PRESSURE. In another treatment (Cohen, 1958) a Markov process is used to describe the succession of probability distributions of states in which an individual is assumed to be when his own judgment conflicts with the judgments of others in his "reference group." The experiment is analogous to those initiated by Asch (1956). Except for one subject, the group consists of the experimenter's accomplices, who make judgments about relative lengths of lines contrary to the obviously perceived differences. The data consist of the subject's judgments made after the accomplices have expressed theirs.

Cohen constructs a Markov chain model in which the states are assumed to be the following:

State 1. If the subject is in this state on trial n , he responds correctly on that and every subsequent trial (i.e., he has decided to pay no attention to the judgments of the others).

State 2. If the subject is in this state, he responds correctly on that trial but may give wrong (conforming) responses on subsequent trials.

State 3. The subject conforms on that trial but may respond correctly on subsequent trials.

State 4. The subject then and thereafter conforms to the group's responses.

We see that State 1 and State 4 are "absorbing states," that is, once entered, they cannot be left. This implies that eventually, after a sufficiently large number of trials, a subject will either reject the group's judgment or reject his own judgment.

Observation of such end-states has previously led to hypotheses concerning the role of personality differences in determining the outcome. Although personality differences may indeed be decisive, it is important to note that the Markov chain model predicts the eventual separation of the subjects into conformists and nonconformists, even if all the subjects have the same "personality." According to this model, it is a matter of chance into which absorbing state each subject eventually will pass. More evidence is required for ascribing the final behavior of subjects to personality differences.

The model does, however, allow the estimation of the transition probabilities for groups of subjects, and these parameters can be taken as reflecting the prevalent or average personality characteristics of the population. Moreover, the model will give a good fit to the time course of the process only if the variance of these parameters is small. The good fits actually obtained therefore speak against large variations in the parameters. Whether this small variance reflects a homogeneity of the subject population studied or the small importance of personality characteristics in determining the transition probabilities, and so the propensity to conform, remains to be established in further studies.

GAME-LEARNING THEORY. Markov chains were used by Flood (1954a, 1954b), by Suppes and Atkinson (1960), and by others as models of social interaction in which aspects of stochastic learning theory and those of game theory were combined.

Game theory, as is now well known, is a static, not a dynamic, theory. Typically, the game matrix, whose entries are payoffs associated with each set of strategy choices by the players, is assumed known to all the players. Moreover, the players are assumed completely "rational" in the sense that each foresees all possible consequences of every choice open to him and is also aware that all the other players possess the same knowledge. Consequently, game theory examines the logical structure of social situations characterized by disparate interests rather than by a possible course of

behavior of the participants in such situations, as determined by psychological parameters. Indeed, there are no behavioral parameters in game-theoretical models, as originally formulated by von Neumann and Morgenstern (1944).

Theories of learning, on the other hand, do contain parameters. For example, in the Markov chain model of stochastic learning theory the transition probabilities are essentially learning parameters. They determine the magnitudes of changes in the probabilities of response as a result of conditioning operating in the learning process. Essentially, then, stochastic learning theory is a "mechanical" theory. Concepts such as "insight," "the logic of the situation," and "strategy," have no place in the currently proposed stochastic models of learning.

The assumptions of learning theory generally lead to predictions of behavior different from those based on game-theoretical considerations. This can be seen even in a one-person game (game against nature). Suppose a subject is presented with a single stimulus and has a choice of two responses, A_1 and A_2 . Suppose that response A_1 is reinforced (say, declared to be correct) with probability π , and A_2 , with probability $1 - \pi$, independently of the subject's responses. Suppose $\pi > 1 - \pi$. If this situation is viewed as a trivial game (against nature), game theory prescribes on the basis of maximizing expected gain (assuming only correct guesses rewarded) the choice of A_1 100% of the time. A stochastic learning model, based on noncontingent reinforcements, on the other hand, predicts an asymptotic frequency π for A_1 .

Some experiments, particularly with nonhuman subjects but some with human subjects, corroborate the result predicted by stochastic learning theory. Similar departures from game-theoretic results can be deduced for stochastic learning models applied to game situations.

For example, in the experiments of Suppes and Atkinson (1960) subjects play a 2×2 game in which the payoffs are reinforcement probabilities. Thus, if A_i and B_i ($i = 1, 2$) represent the responses of the two subjects, respectively, we have a game defined by the following matrix:

$$\begin{array}{cc} & B_1 & B_2 \\ A_1 & \left[\begin{array}{cc} a_1 & a_2 \end{array} \right] \\ A_2 & \left[\begin{array}{cc} a_3 & a_4 \end{array} \right]. \end{array}$$

Here the a 's are the reinforcement probabilities of A 's responses contingent on the joint responses of A and B . If B 's reinforcement probabilities are complementary to A 's, the situation is logically isomorphic to a two-person, constant-sum game. On the other hand, viewed as a learning situation in which the conditioning probabilities θ_A and θ_B characterize the two

subjects (cf. p. 569), we have the following Markov matrix:

$$\begin{array}{c}
 (1, 1) \\
 (1, 2) \\
 (2, 1) \\
 (2, 2)
 \end{array}
 \left[\begin{array}{cccc}
 (1, 1) & (1, 2) & (2, 1) & (2, 2) \\
 a_1(\theta_A - \theta_B) & a_1\theta_B & (1 - a_1)\theta_A & 0 \\
 + (1 - \theta_A) & & & \\
 a_2\theta_B & a_2(\theta_A - \theta_B) & 0 & (1 - a_2)\theta \\
 + (1 - \theta_A) & & & \\
 (1 - a_3)\theta_A & 0 & a_3(\theta_A - \theta_B) & a_3\theta_B \\
 + (1 - \theta_A) & & & \\
 0 & (1 - a_4)\theta_A & a_4\theta_B & a_4(\theta_A - \theta_B) \\
 + (1 - \theta_A) & & &
 \end{array} \right],$$

where the state (i, j) denotes the joint response of A and B .

The asymptotic probabilities of the states turn out to be functions of the a 's and of the ratio θ_A/θ_B . Since the latter is a parameter characterizing the pair of subjects, it is clear that the frequencies of the responses at equilibrium depend on the subjects, contrary to the game-theoretical conclusion which prescribes a solution to the two-person, zero-sum game as a function of the payoffs only.

Suppes and Atkinson's original aim was to design a set of experiments in which normative prescriptions of game theory could be compared with the predictions of stochastic learning theory in game situations where the subjects have the opportunity to modify their choices in successive plays on the basis of previous experience. As the authors themselves state, however, the emphasis soon shifted to testing stochastic learning models per se. Accordingly, the situations were not designed as they are assumed in game theory. Typically, the subjects did not have all the information that the players of a game are assumed to have. However, the experiments were increasingly "gamelike" in the sense that, in successive experiments, progressively more information was given to the subjects until, in the last experiments reported in the study, some groups of subjects had complete knowledge of the game matrix and were playing for money. Thus, for the most part, the corroboration of results predicted by learning theory (as against the strategies prescribed by game theory) does not in itself bespeak the greater accuracy of the learning theory except in those cases in which the requirements of the game situation were completely met. However, not all games were tested under all conditions. With increasing approximation to a real game situation, the games became also more complex, as will subsequently become apparent.

From the point of view of gaining "insight" into the strategic logic of a game, the simplest two-person, zero-sum game is one in which each

opponent has two strategies, of which one dominates the other. The choice of strategy is then determined by the "sure-thing principle" for both players. It is difficult to conceive that two adult players, knowing the game matrix, will do anything but choose the two dominating strategies.

Not quite so obvious, but nearly so, is the choice in a 2×2 game in which, although there is no dominating strategy for one of the players, there is nevertheless a saddle point, that is, a pure minimax strategy for both players. Note, in a 2×2 game the existence of a saddle point ensures a dominating strategy for at least one of the players.

Next in complexity, we might take the $m \times n$ games ($m, n > 2$) with saddle points. Finally, the most general two-person, zero-sum games are those without saddle points, which require mixed strategies for minimax solutions.

In the experiments reported by Suppes and Atkinson the entire range of complexity (except $m, n > 2$) was used. However, aside from the fact that in the simplest games the game matrix was not known to the players, the payoffs were probabilistic reinforcement schedules, as in the one-person game example.

The reason for introducing probabilistic payoffs in experiments with human subjects designed to test a learning theory is obvious. If responses were reinforced with certainty in 2×2 games with saddle points (especially where both players had sure-thing strategies), "insights" would have occurred very quickly, and, as a consequence, the situations would not lend themselves to treatment by stochastic models. We have already seen that even in games against nature with probabilistic reinforcement schedules human subjects sometimes fail to maximize expected gain: they do not choose the more frequently reinforced response exclusively. In two-person games this is even more likely to be true, as it is, in fact, in the experiments reported by Suppes and Atkinson.

It would seem that the introduction of determinate numerical payoffs in games other than 2×2 zero-sum games with saddle points would not impair the usefulness of the Markov model. This is particularly true in the Prisoner's Dilemma-type games, in which, in the absence of communication, a solution is not unequivocally prescribed. We have already seen (cf. Sec. 4.4) that in such games subjects typically oscillate between the cooperative and the noncooperative choice, even if the payoffs are determinate and known. One might therefore postulate the following four states, in which each subject playing such a game might find himself :

1. He has played cooperatively and was rewarded (i.e., the other has played cooperatively also).
2. He has played cooperatively and was punished.

3. He has played noncooperatively and was rewarded.
4. He has played noncooperatively and was punished.

The four states of the individual subject correspond in this case also to the four possible states of the system, namely, (CC) , (CD) , (DC) , and (DD) , where C stands for cooperation and D for defection of each player.

We see now that if we assign probability 1 to the event that a subject will stay with any rewarded state and probability 0 to the event that he will stay with a punished state, then (CC) will be an absorbing state into which the system will pass after at most two plays.

Since the data show nothing of the kind, we might try the next simplest model, namely, assign probability 1 to the event that a subject will stay in a rewarded *noncooperative* state (where the payoff is largest) and the probability θ ($0 < \theta < 1$) to the event that he will stay in the rewarded cooperative state, the probabilities of staying in either of the punished states being zero.

These transition probabilities of individual states then induce the following matrix of transition probabilities among the system states:

$$\begin{array}{c}
 \begin{array}{cccc}
 & (CC) & (CD) & (DC) & (DD) \\
 \begin{array}{l}
 (CC) \\
 (CD) \\
 (DC) \\
 (DD)
 \end{array}
 & \left[\begin{array}{cccc}
 \theta^2 & \theta(1-\theta) & (1-\theta)\theta & (1-\theta)^2 \\
 0 & 0 & 0 & 1 \\
 0 & 0 & 0 & 1 \\
 1 & 0 & 0 & 0
 \end{array} \right]
 \end{array}
 \end{array}$$

The asymptotic probabilities of the states are

$$p(CC) = \frac{1}{2 + 2\theta - 3\theta^2}, \quad (90)$$

$$p(CD) = p(DC) = \frac{\theta(1-\theta)}{2 + 2\theta - 3\theta^2}, \quad (91)$$

$$p(DD) = \frac{1-\theta^2}{2 + 2\theta - 3\theta^2}. \quad (92)$$

From these equations it follows that the over-all frequency of cooperative choice as determined by the parameter θ is

$$f = p(CC) + p(CD) = \frac{1 + \theta - \theta^2}{2 + 2\theta - 3\theta^2}. \quad (93)$$

If the choices were independent, we would have $p(CC) = f^2$. The Markov model, however, predicts $p(CC) > f^2$, which can be directly verified.

Thus, even if we confine ourselves to examining the static (equilibrium) aspects of the process, the Markov model suggests relations not implied in the purely static descriptive theory outlined in Sec. 4.4.

Elaboration of this approach through the assignment of finite "next choice" probabilities to the individual states and deducing the corresponding transition probabilities of the system states are straightforward.

Details of this method and of its experimental applications to three-person games are given in Rapoport et al. (1962).

References

- *Allanson, J. T. Some properties of a randomly connected neural network. In C. Cherry (Ed.), *Information Theory*. New York: Academic Press; London: Butterworth Scientific Publications, 1956. Pp. 303-313.
- Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monogr.*, 1956, **9**.
- Back, K. W. Influence through social communication. *J. abnorm. soc. Psychol.*, 1951, **46**, 9-23.
- Bailey, N. T. J. *The mathematical theory of epidemics*. New York: Hafner, 1957.
- Bales, R. F. *Interaction process analysis; a method for study of small groups*. Cambridge, Mass.: Addison-Wesley Press, 1950.
- Bavelas, A. A mathematical model for group structure. *Applied Anthropology*, 1948, **7**, 16-30.
- Bowley, A. C. On bilateral monopoly. *Econ. J.*, 1928, **38**, 651-659.
- Braithwaite, R. B. *Theory of games as a tool for the moral philosopher*. Cambridge, England: Cambridge Univer. Press, 1955.
- Burke, C. J. Applications of a linear model to two-person interactions. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 180-203.
- Cartwright, D., & Harary, F. Structural balance: a generalization of Heider's theory. *Psychol. Rev.*, 1956, **63**, 277-293.
- Cohen, B. P. A probability model for conformity. *Sociometry*, 1958, **21**, 69-81.
- Coleman, J. S. The mathematical study of small groups. In H. Solomon (Ed.), *Mathematical thinking in the measurement of behavior*. Glencoe, Ill.: The Free Press, 1960. Pp. 1-149.
- Collias, N. E. Statistical factors which make for success in initial encounters between hens. *Amer. Naturalist*, 1943, **77**, 519-538.
- Cournot, A. A. *Récherches sur les principes mathématiques de la théorie des richesses*. English translation: *Researches into the mathematical principles of the theory of wealth*. New York: Macmillan, 1927.
- Davis, R. L. The numbers of structures of finite relations. *Proc. Amer. Math. Soc.*, 1953, **4**, 486-495.
- Davis, R. L. Structure of dominance relations. *Bull. Math. Biophysics*, 1954, **16**, 131-140.
- Deutsch, M. Trust and suspicion. *J. Conflict Resolution*, 1958, **2**, 267-279.

* The starred items are relevant to, although not specifically mentioned in, the text.

- Deutsch, M. Trust, trustworthiness, and the *F* scale. *J. abnorm. soc. Psychol.*, 1960, **61**, 138-140.
- Deutsch, M., & Krauss, R. M. The effect of threat upon interpersonal bargaining. *J. abnorm. soc. Psychol.*, 1960, **61**, 181-189.
- Dodd, S. C., Rainboth, E. D., & Nehnevajsa, J., *Revere Studies on Interaction*. U.S. Air Force report, unpublished. Washington Public Opinion Lab., 1952.
- Festinger, L. The analysis of sociograms using matrix algebra. *Human Rel.*, 1949, **2**, 153-158.
- Festinger, L. Informal social communication. *Psychol. Rev.*, 1950, **57**, 271-282.
- Festinger, L., Schachter, S., & Back, K. W. *Social pressures in informal groups*. New York: Harper, 1950.
- Flood, M. M. A stochastic model for social interaction. *Trans. N.Y. Acad. Sci.*, 1954, **16**, 202-205. (a)
- Flood, M. M. Game learning theory and some decision-making experiments. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision Processes*. New York: Wiley, 1954. Pp. 139-158. (b)
- Forsyth, Elaine, & Katz, L. A matrix approach to the analysis of sociometric data. *Sociometry*, 1946, **9**, 340-347.
- *Foster, C. C., & Rapoport, A. The case of the forgetful burglar. *Math. Monthly*, 1958, **65**, 71-76.
- Fouraker, L. E. Professor Fellner's bilateral monopoly theory. *Southern Econ. Journal*, 1957, **24**, 182-189.
- Gause, G. F. *The struggle for existence*, Baltimore: William & Wilkins, 1934.
- Gause, G. F. Verifications experimentales de la théorie mathématique de la lutte pour la vie. *Actualités scientifiques et industrielles*. Paris: Hermann et Cie., 1935.
- Harary, F. On the notion of balance of a signed graph. *Mich. Math. Journ.*, 1954, **2**, 143-146.
- Harary, F. Graph theoretic methods in the management sciences. *Management Sci.*, 1959, **5**, 387-403.
- Harary, F., & Norman, R. Z. *Graph theory as a mathematical model in social science*. Ann Arbor: Institute for Social Research, 1953.
- Hays, D. G., & Bush, R. R. A study of group action. *Amer. Sociol. Rev.*, 1954, **19**, 694-701.
- Heider, F. Attitudes and cognitive organization. *J. Psychol.*, 1946, **21**, 107-112.
- Hoffman, H. Symbolic logic and the analysis of social organization. *Behav. Sci.*, 1959, **4**, 288-298.
- Homans, G. C. *The human group*. New York: Harper, 1950.
- Katz, L. A new status index derived from sociometric analysis. *Psychometrika*, 1953, **18**, 39-43.
- Katz, L., & Powell, J. H. The number of locally restricted directed graphs. *Proc. Amer. Math. Soc.*, 1954, **5**, 621-626.
- Kemeny, J. G., Snell, J. L., & Thompson, G. L. *Introduction to finite mathematics*. Englewood Cliffs, N.J.: Prentice Hall, 1957.
- König, D. *Theorie der endlichen und unendlichen Graphen*. Leipzig: Akademische Verlagsgesellschaft, 1936.
- Kostitzin, V. A. *Biologie mathématique*. Paris: Librairie Armand Colin, 1937.
- *Landahl, H. D. Outline of a matrix calculus for neural nets. *Bull. Math. Biophysics*, 1947, **9**, 99-108.
- Landahl, H. D. Population growth under the influence of random dispersal. *Bull. Math. Biophysics*, 1957, **19**, 171-186.

- Landahl, H. D., & Runge, R. Outline of a matrix algebra for neutral nets. *Bull. Math. Biophysics*, 1946, **8**, 75-81.
- Landau, H. G. On dominance relations and the structure of animal societies: I. Effect of inherent characteristics. *Bull. Math. Biophysics*, 1951, **13**, 1-19.
- Landau, H. G. On some problems of random nets. *Bull. Math. Biophysics*, 1952, **14**, 203-212.
- Landau, H. G. On dominance relations and the structure of animal societies: III. The condition for a score structure. *Bull. Math. Biophysics*, 1953, **15**, 143-148.
- Landau, H. G., & Rapoport, A. Contributions to the mathematical theory of contagion and spread of information. *Bull. Math. Biophysics*, 1953, **15**, 173-183.
- Leeman, C. P. Patterns of sociometric choice in small groups: a mathematical model and related experimentation. *Sociometry*, 1952, **15**, 220-243.
- Luce, R. D. Connectivity and generalized cliques in sociometric group structures. *Psychometrika*, 1950, **15**, 169-190.
- Luce, R. D., & Perry, A. D. A method of matrix analysis of group structure. *Psychometrika*, 1949, **14**, 95-116.
- Luce, R. D., Macy, J., Jr., & Tagiuri, R. A statistical model for relational analysis. *Psychometrika*, 1955, **20**, 319-327.
- Lutzker, D. R. Internationalism as a predictor of cooperative behavior. *J. Conflict Resolution*, 1960, **4**, 426-430.
- Nash, J. F. Equilibrium points in n -person games. *Proc. Nat. Acad. Sci., U.S.A.*, 1950, **36**, 48-49.
- Newcomb, T. M. An approach to the study of communicative acts. *Psychol. Rev.*, 1953, **60**, 393-404.
- Newcomb, T. M. The prediction of interpersonal attraction. *Amer. Psychologist*, 1956, **11**, 575-586.
- Neyman, J., Park, T., & Scott, Elizabeth L. Struggle for existence. The Tribolium model. Biological and statistical aspects. In J. Neyman (Ed.), *Proc. Third Berkeley Symposium on Math. Stat. and Probability*. Berkeley: Univer. of California Press, 1955. Pp. 41-79.
- *Polya, G. Sur la nombre des isomere de certains composes chimiques. *Comptes Rendus Acad. Sci.*, Paris, 1936, **202**, 1554-1556.
- Raiffa, H. *Arbitration schemes for generalized two-person games*. Rept. M 720-1, R 30, Engg. Res. Inst., University of Michigan, Ann Arbor, 1951.
- *Rapoport, A. Cycle distribution in random nets. *Bull. Math. Biophysics*, 1948, **10**, 145-157.
- Rapoport, A. A probabilistic approach to animal sociology: I, II. *Bull. Math. Biophysics*, 1949, **11**, 183-196; 273-282.
- *Rapoport, A. Nets with distance bias. *Bull. Math. Biophysics*, 1951, **13**, 85-91.
- *Rapoport, A. The probability distribution of distinct hits on closely packed targets. *Bull. Math. Biophysics*, 1951, **13**, 133-137.
- Rapoport, A. Spread of information through a population with sociostructural bias: I. Assumption of transitivity. II. Various models with partial transitivity. *Bull. Math. Biophysics*, 1953, **15**, 523-533, 535-543.
- Rapoport, A. Some game-theoretical aspects of parasitism and symbiosis. *Bull. Math. Biophysics*, 1956, **18**, 15-30.
- Rapoport, A., Gyr, J., Chammah, A., & Dwyer, J. Studies of three-person non-zero-sum, non-negotiable games. *Behav. Sci.*, 1962, **7**, 38-58.
- Rashevsky, N. Studies in mathematical theory of human relations. *Psychometrika*, 1939, **4**, 221-239.

- Rashevsky, N. *Mathematical biology of social relations*. Chicago: Univer. of Chicago Press, 1951.
- *Rashevsky, N. Topology and life: In search of several mathematical principles in biology and sociology. *Bull. Math. Biophysics*, 1954, **16**, 317-348.
- *Rashevsky, N. Some theorems in topology and a possible biological application. *Bull. Math. Biophysics*, 1955, **17**, 111-129.
- *Rashevsky, N. Some remarks on topological biology. *Bull. Math. Biophysics*, 1955, **17**, 207-218.
- *Rashevsky, N. The geometrization of biology. *Bull. Math. Biophysics*, 1956, **18**, 31-56.
- Rashevsky, N. Contributions to the theory of imitative behavior. *Bull. Math. Biophysics*, 1957, **19**, 91-119.
- Richardson, L. F. Generalized foreign policy. *Brit. J. Psychol. Monograph Supplements*, No. 23, 1939.
- Richardson, L. F. War moods: I, II. *Psychometrika*, 1948, **13**, 147-174; 197-232.
- Sauer mann, H., & Selten, R. Ein Oligopol experiment. *Z. Ges. Staatswissenschaft*, 1959, **115**, 427-471.
- Schelling, T. C. *The strategy of conflict*. Cambridge, Mass.: Harvard Univer. Press, 1960.
- Schjelderup-Ebbe, T. Beiträge zur Sozialpsychologie des Haushuhns. *Z. Psychologie*, 1922, **88**, 225-252.
- Scodel, A., Minas, J. S., Ratoosh, P., & Lipetz, M. Some descriptive aspects of two-person non-zero-sum games. *J. Conflict Resolution*, 1959, **3**, 114-119.
- *Shimbel, A. An analysis of theoretical systems of differentiating nervous tissue. *Bull. Math. Biophysics*, 1948, **10**, 131-143.
- *Shimbel, A. Applications of matrix algebra to communication nets. *Bull. Math. Biophysics*, 1951, **13**, 165-178.
- Siegel, S., & Fouraker, L. E. *Bargaining and group decision making*. New York: McGraw-Hill, 1960.
- Simon, H. A. *Models of man*. New York: Wiley, 1957.
- Slobodkin, L. B. Formal properties of animal communities. *General Systems*, 1958, **3**, 93-100.
- Solomonoff, R., & Rapoport, A. Connectivity of random nets. *Bull. Math. Biophysics*, 1951, **13**, 107-117.
- Stigler, G. J. *The theory of price*. New York: Macmillan, 1952.
- Suppes, P., & Atkinson, R. C. *Markov learning models for multiperson interactions*. Stanford: Stanford Univer. Press, 1960.
- *Trucco, E. On the information content of graphs: Compound symbols; different states for each point. *Bull. Math. Biophysics*, 1956, **18**, 237-253.
- Volterra, V. *Leçons sur la théorie mathématique de la lutte pour la vie*. Paris: Gauthier-Villars, 1931.
- Von Neumann, J., & Morgenstern, O. *Theory of games and economic behavior* (1st ed.). Princeton: Princeton Univer. Press, 1944.
- *Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. Sixth Int. Congress on Genetics*, 1932, **1**, 356-366.
- Zajonc, R. B. The concepts of balance, congruity, and dissonance. *Public Opinion Quart.*, 1960, **24**, 280-296.

Author Index

Page numbers in **boldface** indicate bibliography references.

- Ajdukiewicz, K., 411, 412, **415**
Allanson, J. T., 576
Alt, F. L., **417**
Anderson, N. H., 80, 99, 100, 108, **117**
Anscombe, F. J., 96, **117**
Apostel, L., **490**
Arrow, K. J., **118, 265, 266, 267**
Asch, S. E., 570, **576**
Atkinson, R. C., 119, 125, 133, 134, 140,
141, 154, 163, 164, 170, 173, 179,
181, 183, 187, 194, 195, 233, 234,
238, 250, 252, 256, 257, 258, 259,
262, 264, **265, 267, 268**, 571, 572,
573, 574, **579**
Attneave, F., 439, **488**
Audley, R. J., 17, 31, 67, 100, **117**

Back, K. W., 567, **576, 577**
Bailey, N. T. J., 39, **117, 507, 576**
Bales, R. F., 562, **576**
Bar-Hillel, Y., 333, 367, 378, 379, 380,
382, 383, 385, 386, 387, 391, 394,
411, 413, **415, 438, 488**
Barucha-Reid, A. T., 76, **117**
Baskin, W., **321, 418**
Bavelas, A., 533, **576**
Behrend, E. R., 62, **117**
Bellman, R., 83
Berge, C., 278, **319**
Berkson, J., 96, **117**
Billingsley, P., **266**
Birdsall, T. G., 250, 256, **268**
Bitterman, M. E., 62, **117**
Bloch, B., 311, 313, **319**
Bloomfield, L., 309
Booth, A. D., **418**
Bower, G. H., 130, 131, 133, 134, 136,
137, 139, 140, 164, 209, 243, 257,
266
Bowley, A. C., 555, **576**

Braithwaite, R. B., 550, **576**
Bruner, J. S., 319, **319**
Burke, C. J., 163, 181, 194, 195, 198,
234, 238, 250, **266, 267**, 569, **576**
Burton, N. G., 443, **488**
Bush, R. R., 4, 6, 9, 10, 11, 13, 14, 19,
20, 22, 31, 32, 34, 37, 42, 45, 48,
50, 53, 61, 62, 66, 74, 75, 76, 78,
79, 82, 83, 85, 86, 87, 89, 90, 92,
94, 95, 96, 97, 98, 99, 100, 101,
102, 103, 104, 107, 110, 112, 113,
117, 118, 119, 120, 125, 140, 159,
200, 213, 226, 234, 238, 250, 256,
257, **266, 267, 268**, 569, **576, 577**

Cane, V. R., 26, 116, **118**
Carnap, R., 438, **488**
Carterette, T. S., 201, 206, **266**
Cartwright, D., 541, **576**
Chammah, A., 576, **578**
Chapanis, A., 439, **489**
Cherry, C., 439, **489, 576**
Chomsky, N., 276, 285, 292, 293, 295,
297, 299, 302, 303, 304, 308, 309,
315, **319, 320**, 325, 334, 336, 347,
348, 360, 363, 365, 367, 369, 370,
376, 386, 393, 394, 395, 396, 398,
403, 408, 411, **415, 416, 418**, 444,
448, **489**
Chung, K. L., **489**
Cobb, S., **319**
Cohen, B. P., 570, **576**
Cole, M., 163, 181, **267**
Coleman, J. S., **576**
Collias, N. E., 545, **576**
Condon, E. V., 457, **489**
Coombs, C. H., **118, 577**
Cooper, F. S., 313, **320**
Cournot, A. A., 546, 547, 555, **576**

- Cox, D. R., 35, 96, 116, **118**
 Criswell, J., **265**
 Cronbach, L. J., 439, **489**
 Crothers, E. J., 128, 207, 211, 222, 223,
266, 267
 Culik, K., 293, **320, 333, 416**
 Curry, H., 411, **416**
- Davis, M., 291, **320, 354, 358, 416**
 Davis, R. L., **118, 542, 576, 577**
 Delattre, P., 313, **320**
 Detambel, M. H., 188, **266**
 Deutsch, M., 549, 550, 557, 558, **576,**
577
 Dodd, S. C., 520, **577**
 Dwyer, J., 576, **578**
- Edwards, W., 62, 115, **118**
 Eifermann, R. R., 482, **489**
 Elias, P., 444
 Elson, B., 410, **416**
 Estes, W. K., 4, 6, 7, 13, 32, 48, 61, 62,
 75, 85, 89, 98, **117, 118, 119, 120,**
 124, 125, 126, 128, 141, 153, 159,
 163, 164, 169, 170, 173, 179, 181,
 183, 193, 194, 195, 197, 198, 200,
 202, 207, 209, 211, 213, 216, 219,
 221, 222, 223, 226, 228, 229, 231,
 233, 234, 244, 250, **266, 267, 268,**
576
 Estoup, J. B., 457, **489**
- Fano, R. M., 452, **489**
 Fant, C. G. M., 310, **320**
 Feinstein, A., 451, 452, **489**
 Feldman, J., 62, **119**
 Feller, W., 31, 63, 68, **119, 123, 131,**
 146, 176, 199, **267, 424, 426, 489**
 Festinger, L., 536, 537, 565, 566, 567,
577
 Feys, R., 411, **416**
 Fletcher, H., 465, **489**
 Flood, M. M., 571, **577**
 Floyd, R. W., **416**
 Fodor, J., **319, 320, 321, 329, 416, 466,**
489
 Forsyth, E., 535, **577**
 Foster, C. C., **577**
 Fouraker, L. E., 551, 554, 555, 577,
579
- Frankmann, J. P., 194, 195, 250, **267**
 Frick, F. C., 443, 484, **489, 490**
 Friedman, E. A., 440, 464, **490**
 Friedman, M. D., **489**
 Friedman, M. P., 163, 181, **267**
 Fritz, E. L., 443, **489**
- Gardner, R. A., 188, **267**
 Gaifman, C., 411, 413, **415**
 Galanter, E., 9, 14, 50, 53, 74, 75, 90,
 92, 94, 95, 96, 97, 100, 102, 103,
 112, 113, **118, 119, 238, 430, 485,**
486, 490
 Garner, W. R., 439, **489**
 Gause, G. F., 511, **577**
 Ginsberg, R., 164, 215, 248, **268**
 Ginsburg, S., 370, 391, 393, 402, 409,
 410, **416**
 Gnedenko, B. V., 457, **489**
 Goldberg, S., 84, **119, 186, 267**
 Goodnow, J. J., 50, 70, 71, 72, 73, 74,
 75, 97, 98, 101, 108
 Grant, D. A., 108, **117**
 Greibach, S., 388, **416**
 Grier, G. W., Jr., 443, **489**
 Gross, M., 414, **416**
 Gulliksen, H., 31, 37, **119**
 Guttman, N., 201, **267**
 Gyr, J., 576, **578**
- Halle, M., 308, 309, 310, 313, 315, 319,
320, 465, 489
 Hanania, M. I., 13, 17, 106, 110, **119**
 Hannan, E. J., 116, **119**
 Harary, F., 530, 531, 536, 537, 539, 541,
576, 577
 Hardy, G. H., 443, **489**
 Harris, Z. S., 299, **320, 378, 410, 411,**
416
 Hartley, R. V., 431, 432, **489**
 Hayes, K. J., 109, **119**
 Hays, D. G., 569, **577**
 Heider, F., 539, **577**
 Heise, G. A., 465, **490**
 Hill, A. A., **319**
 Hillner, K., 223, **268**
 Hiz, H., 411, **416**
 Hodges, J. L., Jr., 97, **119**
 Hoffman, H., 535, **577**
 Homans, G. C., 563, 564, 565, **577**

- Hopkins, B. L., 128, 207, 211, 222, 223, 244, **267**
- Hovland, C. I., 481, **489**
- Huffman, D. A., 452, 454, **489**
- Hull, C. L., 11, 25, 30, 35, 39, 54, 55, **119**, 206, 213, **267**
- Humboldt, W. von, 319, **320**
- Irwin, F. W., 5, **119**
- Jackson, W., **490**
- Jakobson, R., 308, 310, 311, **319**, **320**, **416**, **417**, **490**, **491**
- Jarvik, M. E., 115, **119**, 179, **267**
- Jeffress, L. A., **417**
- Jonckheere, A. R., 17, 31, 67, 95, 100, **117**
- Jones, M. R., **266**
- Joos, M., **319**
- Jordan, C., 148, 186, **267**
- Kalish, H. I., 201, **267**
- Kanal, L., 75, 83, 85, 86, 89, **119**
- Karlin, S., 75, **118**, **119**, **265**, **266**, **267**
- Karp, R. M., 486, **489**
- Katz, J., 292, **319**, **320**, **321**, 329, **416**, 466, **489**
- Katz, L., 535, 542, **577**
- Kemeny, J. G., 123, 132, 146, 228, **267**, 534, **577**
- Kendall, D. G., 39, **119**, 507, 508
- Khinchin, A. I., 432, **489**
- Khristian, J., **321**
- Kinchla, R. A., 251, 253, 255, **267**
- Kleene, S. C., 333, 334, 336, **417**
- Koch, S., **118**, **266**
- Kohler, W., 328, **417**
- Kolmogorov, A. N., 457, **489**
- Konig, D., 532, 533, **577**
- Kostitzin, V. A., 510, **577**
- Kraft, L. G., 282, **320**
- Krauss, R. M., 550, **577**
- Kulagina, O. S., 411, **417**
- Kuroda, S.-Y., 379, 380
- Laberge, D. L., 202
- Lambek, J., 411, 413, **417**
- Lamperti, J., 62, 64, 75, 89, **119**, 236, **268**
- Land, V., 223, **268**
- Landahl, H. D., 508, 512, **577**
- Landau, H. G., 506, 543, 544, 545, 570, **578**
- Landweber, P. S., 379, 381, **417**
- Langendoen, T., 398, **417**
- Lashley, K. S., 240, 326, 376, **417**
- Leeman, C. P., 545, **578**
- Lees, R. B., 297, 304, **320**
- Lesniewski, S., 411
- Lieberman, A. M., 313, **320**
- Lichten, W., 465, **490**
- Licklider, J. C. R., 443, **488**
- Lipetz, M., 548, 558, **579**
- Littlewood, J. E., 443, **489**
- Locke, W. N., **418**
- Logan, F. A., 9, 20, **119**
- Lorge, I., 456, **491**
- Luce, R. D., 10, 18, 19, 25, 26, 27, 36, 37, 50, 53, 62, 63, 64, 74, 89, 94, 95, 96, 100, 101, 102, 113, **118**, **119**, 234, 238, 250, 256, **266**, **268**, 437, 439, **490**, 536, 537, 545, 546, **578**
- Lukoff, F., 315, **320**
- Lutzker, D. R., 558, **578**
- McCawley, J. D., **321**
- McGill, W. J., 107, **119**
- McMillan, B., 283, **321**, 425, **490**
- McNaughton, R., 331, 334, 352, **417**
- MacKay, D. M., 319, **320**
- Macy, J., Jr., 545, **578**
- Mandelbaum, D. G., **321**
- Mandelbrot, B., 283, **320**, 456, 457, 458, 463, **490**
- Markov, A. A., 409, 423, 424, **490**
- Marschak, J., 456, **490**
- Marx, M., **265**
- Matthews, G. H., 304, **320**, 370, 373, 374, 414, **417**, 469, 476, **490**
- Mehler, J., 482
- Mill, J., 275
- Miller, G. A., 107, **119**, 280, **321**, 334, 336, **416**, **417**, 429, 430, 439, 440, 459, 461, 462, 464, 465, 476, 482, 484, 485, 486, 489, **490**
- Miller, N. E., 213
- Millward, R. B., 163, 181, **267**
- Minas, J. S., 548, 558, **579**
- Morf, A., **490**

- Morgenstern, O., 572, **579**
 Mosteller, F., 4, 6, 9, 10, 11, 13, 14, 19,
 20, 22, 31, 32, 37, 38, 42, 45, 48, 50,
 51, 52, 53, 61, 62, 66, 75, 76, 82, 83,
 85, 86, 89, 90, 92, 94, 95, 96, 98, 99,
 101, 103, 104, 107, 110, 111, **118**,
119, **120**, 159, 200, 213, 226, 234,
 250, 256, 257, **266**
 Mourer, O. H., 213
 Murchison, C., **417**
 Myhill, J., 338, **417**
- Nagel, E., **319**
 Nash, J. F., 548, 550, 555, **578**
 Nehnevajsa, J., 520, **577**
 Newcomb, T. M., 539, 541, **578**
 Newell, A., 62, **119**, 340, **417**, 484, **491**
 Newman, E. B., 424, 461, 462, 464, **490**,
491
 Neyman, J., 511, **578**
 Nicks, D. C., 13, 115, 116, **119**, 179,
268
 Norman, R. Z., 530, 531, **577**
- Oettinger, A., 340, 343, **417**
 Osgood, C. E., 275, **321**
- Pareto, V., 457, **491**, 548, 549, 555
 Parikh, R., 366, 367, 389, 391, **417**
 Park, T., 511, **578**
 Patterson, G. W., 409, **418**
 Penfield, W., **319**
 Pereboom, A. C., 109, **119**
 Perles, M., 367, 379, 380, 382, 383, 385,
 386, 387, 391, 394, **415**
 Perry, A. D., 536, 537, **578**
 Peterson, L. R., 223, **268**
 Pickett, V. B., 410, **416**
 Pike, K. L., 410
 Polya, G., 443, **489**, **578**
 Popper, J., 250, **268**
 Post, E., 358, 382, 383, **417**
 Postal, P., 297, 304, **321**, 378, 414, **417**
 Powell, J. H., 542, **577**
 Pribram, K., 430, 485, 486, **490**
 Pushkin, A. S., 423, 424
- Rabin, M., 333, 334, 337, 338, 379, **417**
 Raiffa, H., 234, **268**, 550, 555, **578**
 Rainboth, E. D., 520, **577**
 Rappaport, A., 506, 519, 544, 545, 548,
 570, 576, **577**, **578**, **579**
 Rashevsky, N., 500, 501, 502, 503, 504,
 548, **578**, **579**
 Ratoosh, P., 548, 558, **579**
 Restle, F., 62, 104, **119**, 202, 256, 257,
268
 Rhodes, I., 340
 Ricardo, D., 549
 Rice, H. G., 370, 402, 409, **416**
 Richardson, L. F., 498, 500, 501, 502,
 509, 510, 563, **579**
 Ritchie, R. W., 352, **417**
 Rogers, H., 291, **321**, 354, **418**
 Rose, G. F., 391, 393, 402, **416**
 Rousseau, J. J., 549
 Runge, R., **577**
- Saltzman, D., 223, 268
 Sapir, E., 309, 310, **321**
 Sardinias, A. A., 409, **418**
 Sauer mann, H., 551, **579**
 Saussure, F. de, 309, **321**, 327, 328,
 329, 330, 414, **418**
 Schachter, S., 567, **577**
 Schatz, C. D., 313, **321**
 Scheinberg, S., 362, 367, 380, **418**
 Schelling, T. C., 555, **579**
 Schjelderup-Ebbe, T., 545, **579**
 Schmitt, S. A., **321**
 Schutzenberger, M. P., 279, 280, 281,
 282, **321**, 337, 345, 347, 348, 352,
 370, 376, 381, 383, 386, 388, 391,
 393, 403, 406, 407, 408, 409, **418**
 Scodel, A., 548, 558, **579**
 Scott, D., 333, 334, 337, 338, 379, **417**
 Scott, E. L., 511, **578**
 SeBreny, J., 193
 Selfridge, J. A., 336, **417**, 429, **490**
 Selten, R., 551, **579**
 Shamir, E., 333, 367, 374, 378, 379, 380,
 382, 383, 385, 386, 387, 391, 394,
 411, 413, **415**, **418**
 Shannon, C. E., 273, **321**, 336, **418**, 423,
 428, 431, 432, 439, 440, 441, 443,
 452, **491**
 Shapiro, H. N., 83

- Shaw, J. C., 340, **417**, 484, **491**
 Sheffield, F. D., 113, **120**
 Shepherdson, J. C., 338, **418**
 Shimbel, A., **579**
 Siegel, S., 554, **579**
 Silverman, R. A., **489**
 Simon, H. A., 340, **417**, 484, **491**, 563,
 565, 566, 567, **579**
 Skinner, B. F., 474, **491**
 Slobodkin, L. B., 511, **579**
 Smith, A., 549
 Smoke, K. L., 481, **491**
 Snell, J. L., 123, 132, 146, 228, **267**, 534,
577
 Solomon, H., **265**, **576**
 Solomon, H. C., **319**
 Solomon, R. L., 50, 53, 75, 92, 95, 96,
 97, 101, 103, 104, 110, 111, 215,
268
 Solomonoff, R., 378, 506, **579**
 Somers, H. H., 430, **491**
 Spence, K. W., 206, **268**
 Sternberg, S. H., 14, 33, 34, 50, 65, 66,
 70, 75, 85, 89, 90, 94, 95, 99, 100,
 102, 108, **118**, **120**, 140, 226, **266**
 Stevens, K. N., 319, **320**, **321**, 465, **489**
 Stevens, S. S., 202, 204, **268**
 Stigler, G. J., 555, **579**
 Straughan, J. H., 7, 13, 61, 98, **118**,
 141, **267**
 Suci, G. J., 275, **321**
 Sumbly, W. H., 443, **489**
 Suppes, P., 63, 64, 75, 85, 89, **118**, **119**,
 125, 133, 134, 141, 153, 154, 159,
 163, 164, 170, 173, 179, 181, 183,
 187, 200, 213, 215, 216, 226, 228,
 233, 234, 236, 238, 248, **265**, **266**,
267, **268**, **319**, 571, 572, 573, 574,
579
 Suszko, R., 411, **418**
 Sweet, H., 309, **321**
 Swets, J. A., 250, 256, **268**
- Tagiuri, R., 545, **578**
 Tannenbaum, P. H., 275, **321**
 Tanner, W. P., Jr., 250, 256, **268**
- Tarski, A., **319**
 Tatsuoka, M., 66, 82, 83, 86, **119**, **120**
 Theios, J., 213, 214, 215, 248, **268**
 Thompson, G. L., 20, **118**, 123, 132,
267, 534, **577**
 Thorndike, E. L., 456, **491**
 Thorpe, W. H., **118**
 Thrall, R. M., **118**, **577**
 Thurstone, L. L., 11, 31, 39, 40, **120**
 Toda, M., 442, **491**
 Tolman, E. C., 326, **418**
 Trakhtenbrot, B. A., 291, **321**
 Trucco, E., **579**
- Underwood, B. J., 109, **120**
- Volterra, V., 510, **579**
 Von Neumann, J., 572, **579**
- Wald, A., 96, **120**
 Wallace, A. F. C., 275, **321**
 Wason, P. C., 481, **491**
 Weaver, W., 336, **418**
 Weiner, N., 431, 432, **491**
 Weinstock, S., 114, **120**
 Weiss, W., 481, **489**
 Wells, R., 410, **418**
 Wilks, S. S., 106, **120**
 Willis, J. C., 457, **491**
 Wilson, T. R., 62, 101, 104, **118**
 Witte, R., 223
 Wright, S., **579**
 Wunderheiler, A., 411, **418**
 Wunderheiler, L., 411, **418**
 Wyckoff, L. B., Jr., 257, **268**
 Wynne, L. C., 50, 53, 75, 92, 95, 96,
 97, 101, 103, 104, 110, 111, 215, **268**
- Yamada, H., 334, 352, **417**, **418**
 Yngve, V. H., 471, 474, 475, 484, **491**
 Yule, G. U., 457, 464, **491**
- Zajonc, R. B., 541, **579**
 Zangwell, O. L., **118**
 Ziff, P., 292, **321**, 466, **491**
 Zipf, G. K., 457, 461, 463, **491**

Subject Index

- Abbreviation, law of, 463
Absorbing state, 82, 498, 571, 575
Absorption, probability of, 82–83, 88
Acquaintance circle, 516–520
Activity, amount of, 562–563
Additive increment model, 38, 51–52
Algebraic model (for language user),
422, 464–483
 assumptions about, 472–475
ALGOL, 403, 409
Algorithms, 354–357
Allophone, 311
All-or-none learning assumption, 126,
153, 170
Alphabet(s), 273
 coding, 450, 452–453
 information capacity of, 273, 439
 input, 338
 output, 338
 universal, 339
Alternation tendency, 34
Altruism, 548
Ambiguity, of grammar, 405, 470
 of language, 274, 280, 466
 of segmentation, 280
 structural, 387–390
Amount of information, 431–432, 437,
439, 481
 see also Information
Analyzable string, 301, 303
Animal sociology, 545
Antisymmetric relation, 542
Approximation, continuous, 42, 45
 deterministic, 39–41, 47–49
 differential equation, 85, 87
 to English, 428–429
 expected operator, 42–43, 45–47
 k-order, 336
 model as an, 102
Arms race, 500–501, 563
Artificial language, *see* Language
Association, law of, 153
Associative chain theory, 376
Asymmetry, 292
 in grammar, 373, 414
 left-right, 473
 of responses, 10
 of sentences, 399, 472
Asymptotic, *see* specific topics
Attitudes within social groups, 539–541
Authoritarianism, 558
Autoclitic responses, 474
Autocorrelation, of errors, 71, 73, 79,
136, 214
 of responses, 33, 69
Automata, abstract, 326–357
 behavior of, 332
 codes as, 283
 with counters, 345, 352
 definite, 376
 deterministic, 334, 343, 379–380, 389,
406
 equivalent, 334
 finite, 331–338, 343, 345, 369, 376,
378–379, 382, 389, 390–401, 406,
421, 424, 426–427, 467, 469–470,
486
 k-limited, 336–337, 426–427, 430,
441–443
 linear-bounded, 338–339, 342, 353,
371, 379–380
 nondeterministic, 379
 one-way, 338
 PDS (pushdown-storage), 339–345,
351–352, 371–374, 376, 378–379,
391, 413, 469, 484
 real time, 352
 restricted-infinite, 352, 360, 371–380,
407, 484
 two-tape, 379
 two-way, 337–338

- Avoidance learning experiment, 111, 213, 215
 component model for, 213–215
- Axiom(s), bargaining, 555
 of component model, 192, 199
 conditioning, for component model, 192
 for linear model, 226–227
 for mixed model, 244
 for pattern model, 155
 of linear model, 226–227
 Luce's, 26–27, 36
 of pattern model, 154–155
 response, 155, 191–192, 226–227, 244
 sampling, 155, 192, 199
- Axone, 513
 Axone density, 514–515, 520, 523–524
- Background, 195–197, 251–252
 Balance (of graph), 541
 Bargaining, 549–551, 554–555
 Barrier, absorbing, 81–82
 Baseline studies with models, 49–50, 106–109
 Behaviorism, 328
 Bernoulli sequence, 214–215
 Beta model, 19, 25–30, 36–37, 50–54, 58, 60, 66, 83, 89, 96, 106, 111
 asymptote for, 62–64
 commutivity of, 64
 and damping, 67
 and experimenter–controlled events, 58
 explicit formula for response probability in, 29
 and linear model, 35, 50–51, 57, 58, 96, 113
 and logistic function, 36
 nomogram for, 51, 54–55
 parameter estimates for, 53–54, 97
 rate of learning in, 52
 recursive formula for response probability in, 29–30
 response-strength for, 25
 responsiveness of, 60
 and shuttlebox data, 28–29, 53–54
 sufficient statistics for parameters of, 96
 and urn scheme, 50
 validation of, 111
- Bias, circularity, 515
 distance, 515, 525
 interaction of sources of, 231, 528
 overlay, 518–519, 524, 528
 popularity, 525, 528
 reciprocity, 525, 528
 response, 142
 sociometric, 523
 sociostructural, 517–519, 521
 symmetry of, 515
 transitivity of, 515, 525, 528
see also Net
- Bilateral monopoly, 551–556
 Biomass (of prey and predator), 509
 Birth rate, 509
 Bit, 435, 462
 Block-moment method of estimation, 101
 Boundary condition, 128–129
 Boundary marker (symbol), 280, 287, 292–293, 334, 338, 452, 459–460
see also P-marker
- Branch (of tree), 290
 Branching process, for N -element model, 156
 left, 474
 multiple, 474–475
 for one-element model in two-choice contingent case, 152
 for one-element model in two-choice noncontingent case, 142, 145
 for paired-comparison learning, 184–185
 right, 473–474
 for two-process discrimination-learning model, 261
- Bridge (of graph), 532
- Calculus, first-order predicate, 355
 sequential, 370, 406
 Categorical grammar, 410–414
 Categories, primitive, 411
 system of, 444, 447–449
 Categorization of order i , 444
 Centrality (in social groups), 533
 Chain of infinite order, 226
 Channel, critical, 532
 noiseless, 450
 Channel capacity, 431, 448

- Choice(s), 4
 cooperative, 558–563, 575
 distribution of number of, 528
 Greenwood-Yule distribution of, 528–529
 independence of, 517
 matrix of sociometric, 535
 pattern of, 141
 Poisson distribution of, 528–529
 of strategy, 557–561, 574
 sociometric, 516, 535, 542, 545
see also Response
- Choice point, 278
- Classificatory matrix, 310–311
- Clique, 516, 532, 537–539
- Closure, 380–381
- Code(s), 277–278, 409
 anagrammatic, 281
 artificial, 277
 as automata, 283
 binary, 453–454
 classification of, 280–281
 error-correcting, 455
 general, 280–281
 left tree, 280–281, 452
 memory of, 279
 minimum redundancy, 450–456, 462
 natural, 277, 281–282, 452
 nontree, 283
 right tree, 280–281
 self-synchronizing, 281
 tree, 280–281, 452
 uniform, 281
 word boundaries in self-synchronizing, 281
- Code symbol, 452
- Coding, 277–283
 efficiency of, 455
 optimal, 450
- Coding alphabet, 450–453
- Coding theorem, 452
- Coding tree, graph of, 278, 289, 484
 for minimum redundancy, 455
 for *P*-markers, 289
- Cohesiveness, 565–566
- Combining-classes condition, 19–24, 26
- Communality, degree of, 205
- Commutativity, *see* Event
- Competence (of language user), 326, 330, 352, 390, 464, 467, 472
- Compiler (for computer), 410
- Complementarity, assumption of, 7, 9–12
- Complete family of operators, 10–11
- Completely connected graph, 532
- Complicated behavior, 483–488
- Component model, 123–125, 153, 191–238
 asymptotic response probability in, 249–250
 asymptotic response variance in, 215–216
 autocorrelation of errors in, 214
 for avoidance learning experiment, 213–215
 axioms of, 192, 199
 for discrimination learning, 249–250
 with fixed-sample-size, 198, 207–219
 with fixed sampling probabilities, 198
 and linear model, 206–238
 mean learning curve for, 214, 228, 250
 for multiperson interactions, 234–238
 probability of reversal in, 214
 for simple learning, 206–219
 with stimulus fluctuation, 219–226
 and total number of errors, 214
- Comprehension (of language), 275
- Computer, 356–357
 handling of natural language by, 343
 memory in, 468–469
 program of instructions for serial, 486
- Computer program(s), simulation of behavior, 485
 theory of, 283
- Concatenation, 273–274, 277–278, 283, 292–295
- Concept-attainment experiment, 481–482
- Conditional expectations, 16, 78–83
- Conditional probability, 16, 131
- Conditioning, operant, 47–49
- Conditioning assumptions, 131, 141
- Conditioning axioms, *see* Axioms
- Conditioning experiment, 239
- Conditioning parameter, 127, 131, 133
- Conditioning state, 125, 130–131, 155, 192
 changes in, 143

- Configuration, 341–342, 350–351
 initial tape-machine, 339–340
- Conflict, logical structure of, 495
- Conformity, 560
- Confusion errors, 154
- Connected graph, 532, 542
- Connotation, 275
- Constituent-structure grammar, *see* Grammar
- Contact, channels for, 523
 frequency of, 504
 randomization of, 520
- Contagion, 496–499, 504–508
 theory of, 525
- Context-free grammar, *see* Grammar
- Context-free language, *see* Language
- Context-sensitive grammar, *see* Grammar
- Context-sensitive language, *see* Language
- Contingent-noncontingent distinction, 15
- Contingent reinforcement, 157–158
 and one-element model, 151–153
- Continuum (of states), 497
- Control unit, 331
- Convergence (of response probability), 19
- Co-occurrence relation, 296–297
- Cooperation (related to attitudes), 558
- Copying machine, 365–366
- Correction (role in language learning), 276
- Correlation, *see* specific topic
- Correspondence problem, 382–384, 387–388
- Counter, automata with, 345, 352
- Countersystem, 345, 378
- Cournot lines, 547
- Criterion-reference learning curve, 109
- Critical channel, 532
- C-terminal string, 299, 306
- Cues, background, 251–252
 discarded, 174
 irrelevant, 240–242, 257
 relevant, 240–242
 verbal, 174, 431–432
see also Stimulus and specific topics
- Curve, *see* specific topics
- Cycle (of sociogram), sign of, 540–541
- Damping of response effects, 67, 72
- Death rate, 509
- Decidability of sets, 354–356
- Decipherability, unique, 389
- Decision problem, recursively solvable, 354, 356
- Decision theory, 256, 512
- Delayed effect, 17
- Denotation, 275
- Dependency system, 288
- Depth of postponed symbol, 474, 484
- Derivation(s), 286, 292
 completely determined set of, 292
 left-to-right, 373–374, 414
n-embedded, 374
 right-to-left, 373–374, 414
- Derived categories, 411
- Derived utterances, 474
- Detection experiments, 251–255
- Deviant utterances, 444–446
- Diad, 546
- Difference equation, 42–43, 82–85
 partial, 84–85
 power-series expansion, 85
 solution of, 85, 139, 148, 152, 159
- Differential equation, 82, 85, 87, 199, 495–496, 501, 563
- Diffusion, 497, 508, 512
- Discourse, initial, 448
- Discriminating statistics, 49
 of sequential properties, 70–73
 of variance of total errors, 73–75
- Discrimination learning, 238–265
 component model for, 249–250
 defined, 238–239
 mixed model for, 243–249
 multiple-process model for, 257–264
 observing responses in, 258
 orienting response in, 257
 pattern model for, 239–243
 probabilistic experiment on, 194–198
 relevant cues (patterns) in, 242, 257
 stimulus sampling model for, 250–256
- Dissipation rate, 499
- Distance bias, 515, 525
- Distinctive features (of phoneme sequence), 310
- Distribution of asymptotic response probability, 45, 144–146, 150, 167–168, 237, 242, 249–250, 253–254

- Distribution, cumulative normal, 25–30
 equilibrium, 569–570
 of errors, 71, 94, 135–139
 Greenwood-Yule, 528–529
 length-frequency (of words), 461
 nonnormal, 458
 Poisson, 528–529
 rank-frequency (of words), 457–459,
 460, 462
 of response probabilities, 13, 186
see also specific topics
- Dominance relation, 541–546, 570
- Domination (between string deriva-
 tions), 293
- Drive stimuli, 197–198
- Element(s), junctural, 308
 left-recursive, 290, 293, 394, 399,
 471–472
 neutral, 210
 nonrecursive, 293
 nonterminal, 294
 recursive, 290, 293, 295, 394
 right-recursive, 290, 293, 394, 399,
 471–472
 self-embedding, 290, 293, 394, 399,
 472
 stimulus, 123
 terminal, 294
- Embedding, degree of, 474
 in natural language, 286
see also Self-embedding
- English, coding efficiency of, 439–440
 double-negative in, 481–482
 letter approximation to, 428
 passive construction in, 482
 probabilities of strings in, 440–441
 redundancy in, 440, 443
 rewriting rules for, 447
 self-embedding in, 471
 speed of transformation in, 482
 transformational grammar for, 477–
 478
 word approximation to, 428–429
 word frequency in, 456
- English grammar, 288
 context-sensitive, 365
- Entropy, 436
- Environment, competition for, 510
- Epidemic, 39, 497, 507–508
- Equations, definability of language by
 system of, 401–409, 501
see also specific topics
- Equilibrium, dynamic, 511, 569
 in interspecies competition, 510–511
 Nash, 548, 560
 in noncooperative nonzero-sum game,
 548
 of probability distribution, 570
 stability of, 500–503, 510–511, 547–
 548, 554
 static, 554, 556–567
- Equivalence class, 8, 12, 14
- Equivalent events, 7
- Equivalent grammars, *see* Grammar
- Error(s), autocorrelation of, 71, 73,
 130, 136
 autocovariance of, 75
 confusion, 154
 distribution of number of, 137–138
j-tuples of, 78
 last, 76, 135, 214
 number of (expected), 70, 77, 81, 86,
 90, 94, 130, 133–134, 226
 and component model, 214
 frequency distribution of, 135–139
 predicted and observed values of,
 136
 variance of, 73–75, 135–136
 types of, 141
 variance of, 130
- Error runs, 70, 77, 90–91
 as discriminator among models, 103
 distribution of lengths of, 71, 94
 lengths of, 70–72
 mean number of, 70, 214
 observed and predicted number of, 71
- Error statistics from models, 130, 134–
 138
- Estimation of parameters, block-mo-
 ment, 101
 errors in, 94
 maximum-likelihood, 89, 93–98
 method of, 52, 94
 minimum-chi-square, 93, 96–97
 for single-operator model, 94
 use of statistics in, 76
- Eugene Onegin*, 423–424

- Event(s), commutative, 7, 17–19, 22, 24, 28, 32, 38, 41, 56, 58, 61, 64, 67
 complementary, 7, 9–12
 contingent, 14–15
 equivalent, 7, 9
 experimental, 6, 8, 12
 experimenter-controlled, 13–14, 23–24, 29–30, 44, 52, 58, 65, 115
 experimenter-subject controlled, 13–14, 45–47, 57, 63
 model, 6, 12–15
 neutral, 227
 path-independent, 7, 16–18
 regular, 333
 reinforcing, 15, 123, 142, 154, 177, 182–183, 227, 235
 repeated occurrence of, 18–19
 subject-controlled, 13–15, 23, 28, 31, 65, 69
 trial, 5
 trial-independent, 17
see also Operator and specific topics
- Event effects, invariance of, 36–38, 61
- Excitatory tendency, 206
- Expectation, conditional, 16, 78–83
- Expected gain, *see* Payoff
- Expected operator, 44–45
- Expected-operator approximation, 42–43, 45–47
- Experiment, *see* specific topics
- Experimental event, *see* Event
- Explicit formula for response probability, 5, 15–18, 110
 analysis of, 65
 for beta model, 29–30, 50, 57
 for commutative events, 18, 23
 for experimenter-controlled event model, 62
 for linear model, 50, 57
 for logistic model, 35
 for perseveration model, 34
 for prediction experiment, 24, 29
 for shuttlebox experiment, 23, 29, 32
 for subject-controlled events, 23, 28
 transformation of, 50–55
 trial-to-trial change in, 5
 for two-event experiment, 51
 and urn scheme, 30–32, 50–51
- Explicit formula for state probabilities, 162
- Exposure time, 211
- Feedback process, learning model as, 69–70
- Fidelity criterion, 273
- Finite automata, *see* Automata
- Finite transducer, *see* Transducer
- First success, trials before, 90–91
- Fixed point of operator, 21
- Fixed-sample-size component model, *see* Component model
- Forgetting (in learning), 18, 24, 56, 65–66, 127, 221, 230
- Formal power series, *see* Power series
- Free-recall verbal learning experiment, data from, 107
- Functional equations, 81–89
 differential equation approximation to, 85, 87–89
 power-series solution, 86–87
- Game-learning theory, 571–576
- Games, against nature, 572
 cooperative nonzero-sum, 549
 experimental, 556–561
 mixed-motive, 550
 noncooperative, 548, 556
 nonnegotiable, 556–560
 nonzero-sum, 547–549, 558
 Prisoner's dilemma, 548, 557, 574
 theory of, 234, 556–557, 571
 zero-sum, 572–574
- Generalization, stimulus, 200–206
- Generative capacity (of grammar), strong, 325–326, 357, 371, 377–378, 406
 weak, 325–326, 357, 377–378, 379
- Generative grammar, 290, 292, 296, 326, 411–412, 465–467
- Goodness-of-fit, 76, 96, 103, 133–134, 179, 215
- Grammar(s), 271, 276, 284
 adequacy of, 283–284, 291–292, 297–300
 ambiguity of, 405, 470
 asymmetry in, 373, 414
 categorical, 410–414

- Grammar(s), constituent-structure, 295
 component of, 296–298
 context-free, 343
 deficiencies of, 297–298, 378–379
 generalization of, 414
 grammatical transformations in,
 300–306
 context-free, 294, 352, 366–410, 413, 469–470, 472, 474
 ambiguity of, 387–388
 constituent-structure, 343
 linear, 383–390
 nonself-embedding, 396, 467
 power-series satisfying, 406
 and restricted-infinite automata,
 371
 special classes of, 368–371
 sufficient condition of, 394
 theory of, 294–295, 340
 undecidable properties of, 382–388
 context-sensitive, 294, 360–368, 373–374, 378, 468–469
 asymmetrical, 469
 general property of, 364
 strictly, 373–374
 undecidable properties of, 363
 definition of, 284–285
 discontinuous, 414
 English, 288, 365
 equivalent, 293, 297, 356, 362, 395–396, 400, 413
 strongly, 297, 395–396, 400
 weakly, 293, 395, 397, 413
 generative, 290, 292–296, 326, 356, 411–412, 465–467
 strong, 325–326, 357, 371, 377–378, 406
 weak, 325–326, 357, 377–379
 linear, 369–370, 379–390, 393, 399
 meta-, 369–370, 380, 385
 minimal, 386, 388
 one-sided, 369–370, 379, 389–390, 409, 421, 467, 470
 of natural language, 366
 normal, 369–371, 393, 396
 modified, 374–375, 377
 nonself-embedding, 400
 phonological component of, 288, 306–313
 phonological rule of, 288, 313–319
- Grammar(s), of programming language, 409
 properties of, 363–364, 382–387
 recursive rules of, 284, 328–329
 self-embedding, 394
 sequential, 369–371, 389, 409
 syntactic component of, 306
 theory of, 285, 295
 transformational, 296–306, 357, 364–365, 476–483
 type *i*, 360–367
 undecidable properties of, 363
 universal, 295–296
 well-formed, 291, 364, 367–368
- Grammatical, *see* specific topic
- Grammaticalness, categorization by, 445, 447
 degree of, 291, 295, 443–449, 466
 deviation from, 291, 444
 hierarchy of, 292
 and well-formedness, 449
- Graph, articulation point of, 532
 balanced, 541
 bridge of, 532
 component of, 532
 connected, 532, 542
 directed, 530, 536, 542
 linear, 495, 512, 530–531, 542
 signed, 530, 540–541
 symmetric, 540
 tree, 278, 289, 455, 484, 532
- Greenwood-Yule distribution, 528–529
- Group, egalitarian, 543
 single-clique, 532
 small, 529–535
 social, 531–532
 symmetric relations among members of, 537
- Group dynamics, 562–576
 classical model of, 563–565
 game-learning theory of, 571–576
 Markov chain model for, 567–576
 qualitative hypothesis for, 564–567
 semiquantitative, 565–567
- Group members, attitudes of, 539–541
- Guessing, in one-element model, 129, 137
 in RTT experiment, 209
- Guessing procedure, 441–443

- Guessing-state model, 134–140, 170–172
- Hebrew, double-negative in, 482
- Heuristics, 277, 318
- Hierarchy, 292, 543
 - of grammaticality, 292
 - in groups, 543
 - index of, 543
 - of tote units, 486–487
- Homo economicus*, 548
- Homogeneity assumption, 99–101, 104
- Homonyms, 447–448
- Hullian model, 54–55
- Hypothesis model, 154
- Identification learning experiment, *see* Discrimination learning
- Imitation, experiment on, 10
 - linear model for, 224
 - in mass behavior, 500–504
 - in peck hierarchy, 545
- Immediate constituent analysis, 370, 411
- Immediate constituents, 289
- Independence from irrelevant alternatives, assumption of, 26–27, 437
- Independence of responses, 13
- Independence of unit, assumption of, 26
- Independent sampling restriction (in RTT experiment), 221
- Index, of cliquishness, 516
 - hierarchy, 543, 545
 - of similarity, 201
 - structure, 301
- Indicator random variables, 77–78
- Individual differences, 13, 99–101, 111, 230
 - in amount of retention loss, 230
 - in learning rate, 133, 174, 229–230
 - in parameter values, 111
 - in rate of forgetting, 230
- Infection rate, 498, 505, 507
- Information, 484
 - amount of, 431–432, 437, 439, 481
 - bits of, 435, 462
 - chunks of, 462
 - levels of processing of, 280
 - measure of, 431–439
 - model of, 105
 - spread of, 505–506, 519–522
- Information capacity, 431–455
 - of alphabets, 273, 439
- Initial symbol, 292
- Innate *faculté de langage*, 327, 329
- Input alphabet, 338
- Input tape, 339
- Insight model, 103, 248
- Intention, 486
- Interaction, intensity of, 562–563
 - model for, 501–512
- Internal state, 331
- International behavior, 501
- Intertrial interval, 141, 219, 224–226
- Invariance condition, 310–313, 318
- Irrelevant alternatives, independence of, 26–27, 437
- Irreversibility, 498
- Item difficulty, differences in, 229–230
- Joint profit, maximization of, 552–555
- Kernel string, 299
- Kinship relation, 533
- k*-limited automaton, 336–337, 426–427, 430, 441–443
- Liaison person, 532
- Language(s), 271, 283
 - accepts a, 332, 342–343
 - ambiguity of, 278, 280, 387–390, 408–409, 466
 - artificial, 272, 283, 285–286, 343, 364
 - complement of, 380–381, 386
 - comprehension of, 275
 - computer, 273, 343, 402–403, 409–410
 - context-free, 294, 351–352, 366–367, 373–374, 376–377, 380, 386, 392–393, 402–403, 408–409
 - context-sensitive, 379–381
 - definable, 402
 - definition of, 283
 - formal, 272, 411
 - generates a, 322, 342–343
 - intersection of, 380–381
 - k*-limited, 336
 - knowledge of, 326, 352, 441–443, 464
 - learning of, 272, 275–279, 307, 314, 330, 430
 - meta-linear sequential, 380

- Language(s), mirror-image, 342, 383
 natural, 271–272, 274, 280–281, 283, 286, 288, 295, 343, 366, 378, 389–390, 421, 450, 475, 483
 programming, 273, 343, 409–410
 regular, 333–335, 338, 347–348, 376–378, 380, 383, 386–387, 393–394, 407–409, 470
 terminal, 293
 theory of, 329
 type *i*, 360–367
 union of, 380–381
 user of, 325–326, 330, 352, 390, 421–422, 441–443, 464, 467, 472–475, 483, 487
- Langue, 327–329
- Last error, 76, 135–136, 214
- Late Thurstone model, 11
- Learning, all-or-none, 126
 avoidance, 111, 213–215
 criterion of, 130
 discrimination, 194–198, 238–265
see also Discrimination learning
 language, 472, 275–277, 307, 314, 330, 430
 paired-associate, 123, 126–141, 239
 paired-comparison, 181–189, 243
 in peck hierarchy, 545
 probability, 141–163, 167, 169, 173–174, 179, 193–194
 rate of, 52
 rote serial, 141
see also specific topics
- Learning assumptions, *see* specific models
- Learning curve, 6, 46–47, 134, 140, 172, 225
 asymptote of, 173
 for avoidance learning, 213–215
 for component model, 153, 213–214, 228, 250
 criterion-reference, 109
 as discriminator among models, 72, 102
 form of, 37, 153
 individual, 72
 effect of individual differences on, 174
 for linear model, 140, 153, 228, 233
 mean, 40, 70, 72, 74–75, 173–174, 213–215, 228, 233, 250
- Learning curve, for one-element model, 140, 153
 for paired-associates learning, 134, 140
 for pattern model, 215, 233
 in probability learning, 173
 of stat-organisms, 46–47
- Learning model, *see* specific models
- Learning-rate parameter, 22, 61, 90, 101, 572
 individual differences in, 133, 174, 229–230
- Learning-to-criterion experiment, one-element model for, 130
- Left-branching, 474
- Left-recursive element, *see* Recursive element
- Left tree code, 280–281, 452
- Length-frequency distribution, 461
- Lexical morphemes, 308
- Lexicon, 370
- Liaison person, 532
- Likelihood-ratio test, 96
 and comparison of models, 106
- Limit point of operator(s), 21, 28
- Linear, *see* specific topic
- Linear (interaction) model, 499–504
- Linear (operator) model, 19–24, 37, 50–51, 58–59, 67, 82–83, 226–227
 asymptotic properties, 62, 82, 226–228, 233
 axioms of, 226–227
 and beta model, 37, 50–51, 57–58, 96, 113
 commutativity of, 67–68, 79
 and component model, 206–238
 conditioning axiom of, 226–227
 and damping, 67
 explicit formula for response probability, 56
 and fixed-sample-size component model, 216, 227–228
 for imitative behavior, 234
 learning curve for, 140, 153, 228, 233
 as limiting case of stimulus sampling model, 226–234
 and multiperson interaction, 234–238
 nomogram for, 51, 54–55
 and one-element model, 140

- Linear (operator) model, and pattern model, 228–233
 and perseveration, 34, 108
 probability matching in, 62
 recursive formula for response probability in, 23–24, 56
 for RRT experiment, 228–232
 for simple learning, 206–234
 variance for, 216
 and urn scheme, 50–51
see also Single-operator linear model
- Linguistics, 271, 274, 283–293, 325–331
- List structure, 484
- Logic, 271
 artificial language of, 283
 methods of symbolic, 535
- Logistic, 30, 35–36, 96–97, 504–508
 sufficient statistics for parameters of, 96
- Luce's axiom, 26–27, 36
- Many-trial perseveration model, 66
- Markov chain, 103, 123, 131, 145–147, 186, 227, 245, 260–261, 424–425, 463, 561–576
 aperiodic, 145
 conditioning states as a, 157, 212
 discrete-time, 17
 ergodic, 150
 higher order, 426–427
 irreducible, 145
k-limited, 426–427
 limit vector of, 145, 186
- Markov source, 424–430, 437
- Marriage types, 533–534
- Mass behavior, model for, 501–504
- Matching theorem, *see* Probability matching
- Maximization of joint profit, 552–555
- Maximum-likelihood method of estimation, 89, 93–98
- Maze experiment, 7, 9, 10–12, 14, 20, 65, 73–75, 92, 103, 113–114
 with correction procedure, 7
 effect of reward in, 73–75, 114
 experimenter-subject controlled events in, 13
 overlearning in, 96
 reversal of reward in, 75, 92, 112–113
 and single-event model, 14
- Meaning, 275, 329, 456
- Meaningfulness, 429
- Mean learning curve, *see* Learning curve and specific topics
- Memory, 279, 471
 computer, 468–469
 human, 16, 471–472, 475–476, 556
 long-term, 476
 short-term, 471, 476, 480
- Mentalism, 327–328
- Messages, 432–435
- Metathetic stimulus dimension, 202
- Minimax strategy, 560, 569, 574
- Minimum chi-square method of estimation, 93, 96–97
- Minimum redundancy code, 450–456, 462
- Mirror-image language, 342, 383
- Mixed model, 243–249
- Mob effect, 504
- Mobility, 508
- Model, *see* specific topic
- Model-free test, 107
- Model type, 6
 testing of, 90, 102–104
- Mohawk (language), 378
- Monogenic system of rules, 359
- Monogenic type 1 grammar, 361
- Monoid, 274, 277
- Monomolecular autocatalytic reaction, analogy to, 37
- Monopoly, bilateral, 551–556
- Monte Carlo method, 76–77, 89, 94
- Morale, 564
- Morpheme, 282, 289, 295–296, 299, 302, 308, 414
- Morpheme structure rules, 314
- Morphophonemics, 309
- Multi-element pattern model, *see* *N*-element pattern model
- Multiperson games, 234
- Multiperson interaction, 234–238
- Multiple-alternatives, 10, 19–21
- Multiple-branching, 474–475
- Multiplicative learning model, 27
- Multiprocess model, 125, 257–264
 asymptotic predictions from, 263–264
 branching process in, 261
- Mutual attractiveness, 566

- Nash equilibrium point, 548, 560
 Natural code, *see* Code
 Natural language, *see* Language
 Negative recency effect, 115, 179
 Negative response effect, 74, 113
 Negotiation set, 556
 Neighborhoods, 515
 Neighbors, 512
 N-element pattern model, 153–191
 asymptotic variance of, 233
 branching process in, 156, 184–185
 mean learning curve for, 173, 233
 and one-element guessing state model, 170
 for paired comparisons, 187–188
 for probability learning, 173–174
 sequential statistics for, 173–174
 for two-choice noncontingent reinforcement experiment, 162–181
 Nesting, degree of, 480
 of dependencies, 470–471, 475
 of phrases, 343
 Net(s), acquaintance relation, 515
 biases in, 515–519
 information-spreading, 520, 522
 neural, 513
 of social relations, 515
 statistical aspects of, 512–519
 tightness of, 519
 tracing of contracts in, 514–515, 523–528
 Neurophysiological correlates, 35
 Neutral element, 210
 Neutral event, 227
 Node, 278, 289, 513
 -to-terminal-node ratio, 480, 485
 Nomogram, for beta model, 51, 54–55
 of clique structure, 536, 538
 of Hullian model, 55
 for linear-operator model, 51, 54–55
 for urn model, 51
 Noncontingent choice experiment, 163–166
 Noncontingent-contingent distinction, 15
 Noncontingent reinforcement schedules, 142–151, 157
 Nondeterministic automaton, 379
 Nondeterministic transducer, 406–407
 Nonlinear interaction models, definition of, 503
 Nonlinear operator(s), 38
 Nonrecursive element, 293
 Nonreward, effect of, 19
 see also Reward, effect of
 Nonsense syllables, 314
 Nonstationary time series, 116
 Nonterminal element, 294
 Nonterminal vocabulary (universal), 357
 Nontree codes, 283
 Normal grammar, *see* Grammar
 Number theory (formalized), 356
 Observing responses, 258–260, 263
 Oligopoly, 234, 551
 One-element guessing-state model, 170–172
 One-element model, 125–153
 all-or-none property of, 126
 autocorrelation of errors, 136
 branching process for, 142, 145, 152
 conditioning assumptions of, 131, 141
 conditioning parameter for, 127
 errors, distribution of, 135–137
 following *k*th success, 140
 last, 135
 and fixed-sample-size component model, 208, 211
 learning curve for, 131, 134, 140, 153
 for learning-to-criterion experiment, 130
 for paired-associates experiment, 126, 128–141
 reference experiment for, 141
 reinforcement schedules in, 142, 145–153
 sequence of responses in, 127–130
 special cases of, 125, 147–151
 for stimulus-response association learning, 126
 for two-choice learning problems, 141–153
 One-element pattern model, 126–128
 One-person game, 572
 One-sided linear grammar, *see* Grammar
 One-trial perseveration model, linear, 33, 66, 108
 logistic, 36, 98, 108

- One-trial perseveration model, recursive formula for response probability in, 34, 56
- Operant conditioning, 47–48
model for, 47–59
- Operationalism, 328
- Operator(s), 5, 17, 56
average, 42
classification of, 56
commutative, 7, 17–19, 22, 24, 28, 32, 38, 56, 58, 61, 64, 67
complete family of, 10–11
fixed point of, 21
identity, 22
ineffectiveness of, 22
limit point of, 21
linear, 9, 21
nonlinear, 17, 27
trial-dependent, 110
trial-independent, 17
unidirectional, 28
see also Event
- Opinion, amount of change in, 566
- Ordinal scale, 567
- Orienting response, 257
- Outcome(s), 4, 7–14, 154
contingency of, 14
definition of, 14
differential effect of, 33
equivalent classes of, 12
Pareto-optimal, 549, 557
response-controlled, 12
response-correlated, 33
symmetry of, 10–12, 33, 45
- Outcome probability, 20
- Outcome sequence (in two-choice prediction experiment), 24
- Output alphabet, 338
- Output tape, 346
- Overlap, degree of, 219
- Overlap bias, 518–519, 524, 528
- Overlearning, 75, 92, 96, 103, 112–113
- Paired-associates experiment, 123, 126–127, 239
data from, 128, 134, 140
interpreted in terms of one-element model, 128–141
- Paired-associates learning-to-criterion experiment, 130
- Paired-associates model, 128–141
- Paired-comparison experiment, 181–191, 243
and pattern model, 243
reference experiment for, 181–182
response probability in, 187
- Pandemic, 508
- Panic, 497
- Paradise fish, 104
- Parameter(s), 5
in avoidance learning, 53–54
choice of statistic to estimate, 95
comparison of methods (of estimation), 97
conditioning, 127, 131, 133
as descriptive statistics of the data, 103–104
estimates of, 53–54, 76, 89–99, 127, 129, 133, 167, 169, 171, 214, 222, 229–233, 253, 256
see also Estimation of parameters free, 106
learning-rate, 22, 61, 90, 101, 572
- Parameter-free properties, 76, 93, 190
- Parameter invariance, 36, 104
- Parameter space, 89
- Pareto-optimal strategy, 548–549, 557
- Parasitism, 547–548
- Parole, 327–328
- Passive transformation, 300, 482
- Path dependence, 34, 56
see also Path independence
- Path independence, 7, 16–21, 26, 38, 52, 56
and combining-classes condition, 19–21
and commutativity, 19
and event invariance, 19
quasi-, 32
- Path length, 17
- Pattern model, 123–124, 125–153, 153–191, 222–223
asymptotic properties, 161–162, 213, 233
axioms for, 154–155
comparison with linear model, 233
for discrimination learning, 239–243
and fixed-sample-size component model, 212, 215

- Pattern model, mean learning curve for, 215, 233
 N-element, 153–191
 one-element, 126–128
 transition probabilities for, 156
 and verbal discrimination experiment, 243
 see also N-element pattern model
 Pattern of stimulation, 123
 Pawnee marriage rules, 535
 Payoff(s), expected, 560
 joint, 557
 matrix of, 234, 237, 571–573
 maximization of in bilateral monopoly, 552–553
 maximization of by *homo economicus*, 548
 maximization of joint, 548–549, 552
 probabilistic, 574
 PDS (pushdown storage), 339–352, 371, 400–401, 469
 generation of a string with, 344
 PDS automaton (pushdown storage automaton), 339–345, 351–352, 371–380, 391, 413, 469, 484
 Peck right (of hens), 542–545
 Percept, 329
 Perceptual capacity, 471
 Perceptual model, 318, 377, 401
 decision theory, 256
 incorporating generative processes, 483
 left-to-right, 472
 optimal, 469–470
 single-pass, 472
 see also Speech perception
 Perceptual process, 329–330
 Performance, 6, 123
 of language user, 326–330, 390, 464, 467
 Permutation, 304–305, 534
 Perseveration model, linear, 34, 108
 logistic, 36, 98, 108
 many-trial, 66
 one-trial, 33, 66, 108
 Phase space, 496, 511–512
 Phoneme, 308–310
 Phones, 308
 Phonetic alphabet, 295, 307–308
 Phonetic representation and related topics, 288, 308–314
 Phonological component and related topics, 288, 306–319
 Phrase-marker, *see P*-marker
 Phrases, nesting of, 343
 Phrase structure, 288
 Phrase types, categorization into, 410–411
 Plan, 486–487
 Player, rational, 571
P-marker(s), 293–294, 296, 298–299, 304, 306, 359, 363, 365, 405, 468, 473–474, 477–481
 and ambiguity of grammar, 405
 and attachment transformation, 305
 construction of, 301
 derived, 301, 303–304, 307, 478–481
 generated by rewriting rules, 477
 graph of, 289
 and singularity transformation, 305
 strong derivation of, 368
 and structural complexity, 481
 Poisson distribution, 528–529
 Polish notation, 370, 406
 Polynomial expression, 401–402
 Popularity, 535
 Popularity bias, 525, 528
 Population, assumption of well-mixedness, 505
 dissemination of genes in, 497
 homogeneity of, 503
 nonhomogeneity of, 503
 predator, 509
 size of, 504
 statistical study of, 522–529
 Positive response effect, 72, 74, 108
 Postponed symbol, 474–475
 Postponement, depth of, 484–485
 Power series, 406
 algebraic elements of, 407
 characteristic, 406
 closed, 403
 formal, 403–407
 ordinary, 403
 solution, to function equations, 86–87
 to difference equations, 85
 Practice effect in signal detection experiment, 250–251

- Predator population, 509
- Prediction experiment, 7–9, 11–14, 44, 65, 84
 asymptotic behavior in, 61–65
 and beta model, 29–30, 56, 63–64
 experimental event in, 8
 experimenter-controlled event in, 13
 explicit formula for response probability for, 24, 29
 and linear model, 56
 and single-event model, 14
 and stimulus-fluctuation model, 223–226
- Preference, intransitivity of, 542
- Pretraining, 36
- Price, 456
- Price leader, 553–555
- Primitive categories, 411
- Prisoner's dilemma game, 548, 557, 574
- Probabilistic reinforcement schedule, 141–153
- Probability, of absorption, 82–83, 88
 of choice, 517
 conditional, 16, 131
 of contact, 499
 of reversal in component model, 214
 transition, 156–158, 212–213, 498, 575
see also Response probability
- Probability learning, 141–162
 asymptote in, 144–146, 150, 153, 167, 169
 and contingent reinforcement, 151–153
N-element model for, 173–174
 and noncontingent reinforcement, 142–144, 162–163
 one-element model for, 147–151
 pattern model for, 153–162
 and probability matching, 151, 179
 reference experiment for, 141–142
 and stimulus compounding, 193–194
- Probability matching, 61–64, 151, 179
 by paradise fish, 105
 and urn scheme, 65
- Probability vector, 5, 146
- Profit, *see* Payoff
- Programming language, *see* Language
- Pronunciation, 247–275
- Proper analysis, 301
- Property-space, 89–90
- Prothetic stimulus dimension, 204
- Psychoeconomics, 546–561
- Psycholinguistic model, 327, 329
- Psychophysical experiments, 33, 256
- Pure reinforcement model, 226
- Pushdown storage, *see* PDS
- Quantification theory, 355
- Random net, 506, 513–517, 519, 528
 connectivity of, 506
 rejection of (hypothesis of), 514, 524
- Random walk, 463
- Rank-frequency relation (of words), 457–464
- Rational function, 407
- Rationality, collective, 556–557, 571
- Ratio scale, 26, 562
- Reaction potential, 25, 35, 54
- Reaction probability, 54
- Reading head, 331
- Real-time automaton, 352
- Receiver operating characteristic (ROC) curve, 256
- Receiver's uncertainty, measure of, 432, 435
- Reciprocal bias, 525, 528
- Recognition routine, 377, 469
- Recognition of words, 465
- Recovery, 497
- Recursive element, 290, 293, 295, 394
 left, 290, 293, 394, 399, 471, 472
 right, 290, 293, 394, 399, 471, 472
 self-embedding, 290, 293, 394, 399, 472
 types of, 290
- Recursive formula for response probability, 16, 18, 23–24, 28, 30, 34, 44, 56
 approximate, 44
 for beta model, 29–30
 classification of, 56
 for commutative events, 18
 general, 30
 for linear operator model, 23, 56
 for one-trial perseveration model, 34, 56
 for prediction experiment, 24, 29, 44
 for shuttlebox experiment, 23, 28

- Recursive formula for response probability, for single-operator model, 56
 for subject-controlled events, 23, 28
 for two-event experiment, 51
 for urn scheme, 30–32, 56
- Recursive generative process, 290
- Recursively enumerable set, 355–356, 361–362
- Recursive rules, 284, 328–329
- Recursive set, 355
- Redundancy, 431, 439–443, 449, 455, 484
 in English, 440, 443
 estimation of, 440–442
 maximization of, 449
 minimum, 450–456, 462
 sequential, 442
- Reflexivity, 279, 293
- Regression analysis of binary sequence, 35
- Regular event, 333
- Regular language, 334–335, 347–348, 376–378, 380, 383, 386–387, 393–394, 470
 defined, 333
 and formal power series, 407–409
 structural characterization theorem for, 334
 and two-way automata, 338
- Reinforcement, 123
 conditions of, 20
 contingent, 151–153, 157–158
 noncontingent, 142–151
 probability of, 572
 schedule of, 142, 158
- Reinforcing event(s), *see* Event
- Relation, acquaintance, 515, 518
 antisymmetric, 542
 asymmetric, 292
 binary, 495, 530
 co-occurrence, 296–297
 dependency, 286
 dominance, 542–546, 570
 equivalence, 7, 9, 14, 279
 grammatical, 477–478, 480
 kinship, 533
 rank-frequency, 462–464
 reflexive, 279, 293
 submissive, 543
 symmetrical, 25, 33, 45, 279, 530
- Relation, transitive, 279, 293, 518, 542
- Removal rate, 507
- Repetition tendency, 33
- Representing expression, 334–336
- Reproduction, rate of, 499
- Resolution, rules of, 411, 413
- Response(s), 4
 asymmetric, 10–11
 autoclitic, 474
 autocorrelation of, 33, 69
 dependence of, 33, 56
 discriminative, 258–260
 independence of, 13
 observing, 258–260, 263
 orienting, 257
 problem of definition of, 20
 repetition tendency of, 108
 set of, 5, 123
 symmetry of, 9–12, 45
 variance of, 174–177, 233
see also specific topics
- Response axioms, 155, 192, 226–227, 244
- Response bias, 142
- Response effect, accumulation of, 67
 damping of, 67, 72
 direct, 66–68
 erasing of, 67, 72
 indirect, 68–69
 magnitude of, 67
 negative, 74, 113
 positive, 72, 74, 108
 undamped, 67
- Response-outcome event, 7, 12, 33–34, 78
- Response probability, 5
 asymptotic, 45, 61–65, 102–103, 105, 144–146, 150, 159–162, 167–169, 176, 187–188, 224–227, 233, 237, 242, 249–250, 253–254
 convergence of, 19
 distribution of, 13, 186
 explicit formula for, 5, 15–16, 18, 20, 22–24, 28–32, 34–35, 50–57, 62, 65, 110
 in independent sampling model, 224
 mean of distribution of, 172–173
 and moments in pattern model, 158–162

- Response probability, nonlinear transformation on, 27
 recursive formula for, 16, 18, 23–24, 28–32, 34, 44, 51, 56
 for stimulus fluctuation model, 224
 variance of distribution of, 159, 172–173
- Response sequences, 75, 127–133
- Response strength, 25, 32
see also Beta model
- Responsiveness of model, 56–61
- Restricted-infinite automaton, 352, 360, 371–380, 407, 484
- Retention curve, 219–220
- Retention loss, 209, 211
- Reversal, 112–113
 of dominance, 570
 in learning, 214
- Reversibility, 497–498
- Reward, effect of, 19, 33, 45, 48, 57, 64, 73, 107, 110
 in prediction experiment, 44
 on response probability, 113
 in shuttlebox experiment, 74
 in T-maze, 73
 in urn scheme, 48
 on variance of total errors, 73–74
- Reward and nonreward parameters, estimates of, 107
- Rewriting rules, 468–475, 477, 481
 unrestricted, 357–360, 379
- Right-branching, 473–474
- Right recursive element, *see* Recursive element
- Right tree code, 280–281
- ROC (receiver operating characteristic) curves, 256
- Rote serial learning, model for, 141
- RTT experiment, 207
 fixed-sample-size component model for, 207–211, 221–232
 linear models for, 228–230
 neutral element model for, 210
 retention loss in, 209, 229
 stimulus fluctuation model for, 207, 221–223, 232
 stimulus-sampling model for, 229–232
- Rules (grammatical), left-linear, 369
 linear, 369–370
 meta-linear, 369
- Rules (grammatical), monogenic system of, 359
 recursive, 284, 328–329
 of resolution, 411, 413
 rewriting, 357–360, 379, 468–475, 477, 481
 right-linear, 369
 selection, 301
 for synthesizing sentences, 466
 terminating, 369
- Rumor, 497
- Runs of errors, *see* Error; Error runs
- Runway experiment, 8–10
- Saddle point, 574
- Sampling axiom, 155, 199, 252
- Sampling model, *see* Stimulus fluctuation, Stimulus sampling, Component, *N*-element, One element, and Pattern models
- Satisfies, 404
- Saussurian view of linguistics, 327–330
- Scale, ordinal, 567
 ratio, 26, 562
- Score structure, 543
- Secondary reinforcement, 19
 model for, 105
- Segmentation, 280
- Selective information, measure of, 431, 438–439
- Selective sampling, effects of, 108
- Self-embedding, 286, 343, 470, 473–475, 480
 degree of, 396, 400, 468, 470, 474, 480, 484
 in English, 471
- Self-embedding elements, 290, 293, 394, 399, 472
- Self-embedding grammar, 394
- Self-synchronizing code, 281
- Semantic information, measure of, 438
- Semantics, 328, 466
- Semigroup, 274, 280
- Sentence(s), 283, 292
 asymmetry of, 399, 472
 definition of, 332–333
 recognizing device for, 318, 465
 rules for synthesizing, 466
 structure of, 228, 297–298, 326–327, 399

- Sentence(s), structural complexity, measure of, 480–481
 structural description of, 289, 297–298, 399
- Sentence-matching test, 482
- Sequence, of conditioning states, 132
 index, 382–383
 of responses, 75, 127–133
 of trials, fixed-sample-size model, predictions of, 211
- Sequential calculus, 37, 406
- Sequential grammar, 369–371, 389, 409
- Sequential statistics, 5, 70–73, 188
 for error runs, 70–71
 estimation procedures for, 166–169
 for fixed-sample-size component model, 211–213, 216–219
 for linear model and pattern model, 190
 and mean learning curve, 173–174
 for N -element model, 173, 188–191
 observed and estimated values for, 167, 169–170, 254–256
 for one-element model, 148
 for paired-comparison learning experiment, 188–191
 for pattern model, 164, 169–170
 technique for deriving, 177–178
 for visual detection experiment, 254–256
- Serial autocorrelation of errors, 71
- Serial computer, program of instructions for, 486
- Set(s), computable, 354
 decidable, 354–355
 recursively enumerable, 355–356, 361–362
 stimulus, 123–124, 182, 192
 of strings, 356–357, 362–363
 theory of, 495
- Shuttlebox experiment, 8–10, 13, 40, 65, 74–75, 93, 95–96, 110
 beta model for, 28–29, 53–54
 data from, 50, 103, 111
 explicit formula for response probability for, 23, 29, 32
 linear-operator model for, 22, 53–54, 105, 109
 model-free analysis of, 53
 recursive formula for response probability of, 23, 28
- Shuttlebox experiment, Restle model for, 104
 urn model for, 32, 53, 54
- Signal detection experiment, 250–256
- Signed graph, 530, 540–541
- Similarity, index of, 201
- Simplicity, 38
- Simulation of behavior, computer program for, 485
- Single-event model, 14, 66
- Single-operator linear model, 34, 66, 89, 90, 92, 140
 estimation of parameters for, 94
 expected number of errors in, 90
 mean learning curve for, 140
 in prediction experiment, 84
 recursive formula for, 56, 84
- Single-pass device, 469, 472
- Singularity transformation, 303–305
- Sink, 497, 499, 509
- Small group, *see* Group
- Social group, *see* Group
- Social disintegration, 564
- Social dominance, 541–546, 570
- Social interactions, measurability of, 495
- Social rank, 570
- Social space, distance in, 528–529
 index of cliquishness in, 516
 metrical properties of, 528
 topology of, 515, 528
- Social structure, 543
- Sociogram, 522–529, 539–541
- Sociometric choice, 496, 516, 523–529, 542, 545
- Sociostructural bias, 517–519, 521
- Sound structure, 306–319
- Source, 497–498, 509
- Speech perception, 273, 311, 314, 318
- Spontaneous recovery, 221
- Spontaneous regression, 220–221
- Stability of equilibrium, 500–503, 510–511, 547–548, 554
- State(s), absorbing, 498, 571, 575
 asymptotic probabilities of, 573, 575
 change in, 498–499, 504, 509, 520
 conditioning, 125, 130–131, 143, 155, 192
 continuum of, 497
 diagram, 332
 final, 334

- State(s), internal, 331
 irreversible, 497–498
 steady, 463, 496
 of subjects, 575
- State probability, 162
- Stat-organism, data from, 46–47
- Status, 535
- Stereotypy, 484
- Stimulus, 123
 background, 195–197
 communality of, 124
 intensity of, 124
 overlapping of samples of, 219
 in paired-comparison experiment, 182
 scaling of, 203–205
 transfer, 206
 variability of, 124
see also Cues
- Stimulus compounding, 193–198
- Stimulus elements, 123
 background, 195–196
- Stimulus fluctuation model, 219
 applied to RTT experiment, 221–223,
 232
 linear model as limiting case of, 226–
 228
 for noncontingent case, 223–226, 228
 for prediction experiment, 223–226
- Stimulus generalization, 200–206
- Stimulus pattern model, 153–191
 asymptotic distribution, 159–162, 169
 axioms for, 155
 matching theorem, 179–181
 for paired-comparison experiment,
 181–191
- Stimulus-sampling model, 31, 61, 123–
 125
 for discrimination learning, 250–256
 limiting case, 226–234
 exposure time in, 211
 urn scheme, 31
- Stimulus similarity, 124
- Stochastic learning model, 4, 569
see also specific models
- Stochastic source, 427–430, 432
- Stochastic theory of communication,
 422–423
- Storage, pushdown, *see* PDS
- Storage tape, 339, 346
- Strategy, choice of, 557–558, 574
 cooperative, 557–563, 575
 in cooperative nonzero-sum games,
 549
 dominating, 574
 in finite nonzero-sum game, 548
 minimax, 560–569, 574
 mixed, 574
 in negotiable games, 550
 noncooperative, 557, 575
 Pareto-optimal, 549, 557
 sure-thing principle in choice of, 574
 in two-person nonzero-sum game, 547
 in two-person zero-sum game, 574
- Strictly finite automata, *see* Automata
- String(s), accept a, 332, 337–338, 340,
 342, 353, 359
 analyzable, 301, 303
 binary operation on, 292
 blocked, 348
 constituent, 304
 C-terminal, 299, 306
 derivation of, 286, 292–293, 373–374,
 414
 domination of, 293
 finite, 362–363
 generate a, 332, 337, 342, 353, 356–
 363
 generated with PDS, 344
 infinite, 362–363
 kernel, 299
 length of, 340
 null, 362–363
 probability of (in English), 440–441
 recursively enumerable, 356
 reduce a, 348
 redundancy in, 440, 443
 terminal, 288, 293
 termination of a, 293
 transformed, 303–306
 T-terminal, 299
 unique, 294
- Structural ambiguity, 387–390, 406
- Structural balance, theory of, 539–541
- Structural complexity, 480–481, 485
- Structural description, 285, 289, 297–
 298, 307, 479
 left-right asymmetry in, 399
 possible, 295
- Structural regularities, 330

- Structure, of clique, 526, 537–539
 dominance, 541–546, 569–570
 grammatical, 488
 group, 543
 index of, 301
 peck-right, 542–543
- Stylistic transformations, 471–472
- Subgraph, 532
- Subgroup, 532
- Subject-controlled event, *see* Event
- Subject-controlled event model, 34, 65
- Submissive relations, 543
- Subspace, model type, 89–93
- Substitution transformation, 304
- Substitutive stimulus dimension, 202–203
- Success, trials before first, 90–91
- Successive samples, independence of, 219
- Successive symbols, correlated, 423
- Sufficient statistics, 38, 96
- Summation effect, 196
- Sure-thing principle, 574
- Survival, conditions of, 510–511
- Syllabaries, 273
- Symbiosis, model of, 546–549
- Symbol(s), amount of information per, 432
 boundary, 287, 292–293, 334, 338
 code, 452
 correlation of successive, 423
 initial, 292
 postponed, 474, 478
 string of, *see* String
 terminal, 359, 474
- Symmetry, 25, 33, 45, 279, 530
- Symmetry bias, 515
- Synchronization, 278
- Syntactic category, 430
- Syntactic component (of sound structure), 306
- Syntactic structure, 285, 328
- Tagmemics, 410
- Tape, blocked, 331, 348
 contents of, 341
 as counter, 345
 input, 339
 output, 346
 storage, 339, 346
- Terminal element, 294
- Terminal language, 293
- Terminal situation of scanning device, 340
- Terminal string, 288, 293
- Terminal symbols, 359, 474
- Terminal vocabulary, universal, 357
- Testing models, insensitivity of methods of, 100
see also specific models; Goodness-of-fit
- Theory, *see* specific topics
- Threats, 550–551, 554
- Threshold, 11, 54
- Threshold model, 11, 25
- Thurstone model, late, 11
- Tightness of net, measure of, 519
- Time series, nonstationary, 116
- T-maze experiment, *see* Maze experiment
- Total errors, *see* Errors
- Tote unit, 485–486
- Tracing (of information flow), 513–514, 524–528
- Transducer, 351, 374–375, 377, 387, 391–392, 397, 399–400
 bounded, 347, 392–393
 finite, 346–348, 395, 410, 466, 468–469
 generation of a structural description by a, 396
 information lossless, 347
 nondeterministic, 406–407
 strong equivalence of, 396, 400
 understanding of all sentences by, 469
 with PDS, 348
- Transfer, effect, 206, 247
- Transformation(s), additive, 11
 adjunction, 306
 attachment, 304–305
 behavioral, 487
 classes of, 304
 complexity of, 485
 deletion, 306
 elementary, 301–302
 generalization, 303–304
 grammatical, 296–306, 357, 365, 377, 379, 387, 477, 481–482
 interrogative, 482
 negative, 482

- Transformation(s), obligatory, 299
 optional, 299
 passive, 300, 482
 permutation, 304, 534
 phonological, 314
 psychological correlates of, 482
 singularity, 303–305
 speed of, 482
 stylistic, 471–472
 substitution, 304
- Transformational cycle, 315
- Transformational grammar, 296–306, 357, 365, 476–483
- Transformed string, constituent structure of, 303–306
- Transition probabilities, 156–158, 498
 for fixed-sample-size component model, 212–213
 of individual state, 575
 of system state, 575
- Transition rule, *see* Operator
- Transitivity, 279, 293, 518, 542
- Transitivity bias, 515, 525, 528
- Tree, 278, 289–290, 455, 528, 532
see also Graph; Branching process; Tree code
- Tree code(s), 280–282, 452
- Trial-dependent operators, 110
- Trial event, 5
- Trial independent event, 17
- Trials, before first success, 90–91
 to last error, 76, 135, 214
- T*-terminal string, 299
- Turing machine, 352–362, 371, 379
- Two-armed bandit experiment, 9, 15, 65–66, 73, 75, 101
 data from, 34, 50, 70–73, 105
 and linear one-trial perseveration model, 108
- Two-choice prediction experiment, *see* Prediction experiment
- Two-event experiment, 51
- Type *i* (grammars and languages), 360–367
- Uncertainty measure of, 440
 reduction of, 432
- Undecidability (of grammar), 363
- Understandability, measure of, 480
- Uniform codes, 281
- Uniformity of opinion, pressure to achieve, 565–566
- Union of languages, 380–381
- Uniqueness, proof of, 86, 294
- Unit, independence of, 26
- Universal, *see* specific topic
- Unrestricted rewriting system, 357, 359–360, 379
- Urn scheme, 31–36, 48, 50–51, 66
 approximations for, 40–43
 error probability in, 52
 explicit formula for response probability in, 30–32, 50–51
 nomogram for, 51
 quasi-independence of path, 32
 recursive formula for response probability in, 56
- Utility, 547–548
- Utterance, derived, 474
 deviant, 444, 446
 grammatical, 429
- Variance, asymptotic, 219, 228, 233
 of errors, 72–75, 130, 135–136
 for linear model, 216–228
- Verbal behavior, 271
- Verbal cues, 174, 431–432
- Verbal experiments, 107, 243
- Verbal operant response, 474
- Visual detection experiment, 253–254
- Vocabulary, 273, 292, 357
- Vowel reduction, 314–315
- War moods, 498
- Well-formed grammar, 291, 364, 367–368
- Well-mixedness, 499, 512
- Word(s), 315
 distribution of length-frequency of, 457–464
 distribution of rank-frequency of, 457–464
 distribution of sequences of, 464
 frequencies of, 456–464

UNIVERSAL
LIBRARY



116 194

UNIVERSAL
LIBRARY