ESD-TR-66-404
ESTI FILE COPY

ESD RECORD COPY
RETURN TO
SCIENTIFIC & TECHNICAL INFORMATION DIVISION
(ESTI), BUILDING 1211

ESD ACCESSION
ESTI Call No. AL 54826
Copy No. 1

ESD-TR-66-404

(FINAL REPORT)

NONCONSERVATIVE PROBABILISTIC INFORMATION PROCESSING SYSTEMS

TECHNICAL DOCUMENTARY REPORT NO. ESD-TR-66-404

December 1966

Ward Edwards

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Distribution of this document is unlimited.

(Prepared under Contract No. AF 19(628)-2823 by the Institute of Science and Technology, The University of Michigan, Ann Arbor, Michigan)

AD0647092

When U. S. Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Do not return this copy. Retain or destroy.

ESD-TR-66-404

(Final Report)

NONCONSERVATIVE PROBABILISTIC INFORMATION PROCESSING SYSTEMS

Prepared by

Ward Edwards

Engineering Psychology Laboratory

INSTITUTE OF SCIENCE AND TECHNOLOGY

The University of Michigan

Ann Arbor

for

Decision Sciences Laboratory

Electronics Systems Division

Air Force Systems Command

United States Air Force

L. G. Hanscom Field, Bedford, Massachusetts

Distribution of this document is unlimited.

Prepared under Contract No. AF 19(628)-2823

FOREWORD

The research reported here was conducted in support of Project 2806, Task 280609 under Contract AF 19(628)-2823. The work was performed between September 1963 and April 1966 at the Institute of Science and Technology, with Dr. Ward Edwards as the principal investigator. Contracts and grants to The University of Michigan for the support of sponsored research by the Institute of Science and Technology are administered through the Office of the Vice-President for Research.

The report summarizes experimental studies of human processing of probabilistic information, particularly with respect to the basic concepts of a PIP (Probabilistic Information Processing) system that might be useful in "diagnostic" information-processing systems.

The author wishes to acknowledge the contributory work of Stanley D. Bielby, Wesley M. DuCharme, Barbara Goodman, William L. Hays, Richard W. Pew, Lawrence D. Phillips, Andries F. Sanders, Paul Slovic, Mary Ann Swain, Richard G. Swensson, Gloria E. Wheeler, Donald L. Zink, and Marilyn T. Zivian.

This technical report has been reviewed and is approved.

Walter E. Organist

WALTER E. ORGANIST
Research Psychologist
Decision Sciences Laboratory

for Donald W. Carr

ROY MORGAN
Colonel, USAF
Director, Decision Sciences Laboratory

ABSTRACT

This report is concerned with two large-scale simulation experiments on probabilistic information processing (PIP) systems. One, a very large and prolonged study of four systems, yielded the conclusion that PIP is indeed an efficient philosophy for information processing systems—at least twice as efficient as its next-best competitor, and four times as efficient as a representative of current processing techniques. The second PIP experiment was concerned with whether likelihood estimators in PIPs should be allowed to know the state of system opinion; the data confirm the suggestion that it might be undesirable. These experiments required the use of an on-line computer system.

This comparison of PIP and its competitors clearly indicates that PIP is superior, but does not indicate how PIP compares with theoretically optimal performance since no objective model of the data-generating process was available. A smaller-scale laboratory experiment is reported that compares PIP with a posterior-odds estimation system (POP) in a task sufficiently complex to be difficult for subjects and yet allowing an objective standard of correct performance. PIP was far superior to POP. PIP and calculations of optimal performance were roughly comparable, with PIP sometimes more extreme than optimal performance and sometimes less extreme. Another small laboratory study, concerned with the development of a response mode in which subjects report on probabilities by making choices among bets, is reported. Its original purpose was to develop a response mode for one group in the first PIP experiment, but it proved to be considerably more important than that. A study is also reported in which the fact of human conservatism in information processing, the fact with which PIP is designed to cope, is again demonstrated under conditions of realistic complexity that have a military flavor.

People are shown to be conservative information processors. To cope with this it is appropriate to design information processing systems in which human estimates of likelihood ratios are followed by computer aggregation of these into posterior distributions by means of Bayes's theorem. Such procedures extract information from data more efficiently than any other way of exploiting human judgment yet tried, and produce data roughly comparable with theoretically optimal calculations when such calculations are possible.

CONTENTS

Foreword ii

Abstract iii

List of Figures vii

List of Tables viii

1. Introduction 1

2. The Design and Evaluation of Probabilistic Information Processing Systems 2
   - 2.1. Information Processing and Bayes's Theorem 2
     - 2.1.1. The Measurement of Uncertainty 4
     - 2.1.2. Bayes's Theorem and Bayesian Statistics 5
     - 2.1.3. Bayes's Theorem in Diagnostic Information Processing 7
     - 2.1.4. A Probabilistic Information Processing System 9
     - 2.1.5. Articulation of Diagnosis with Action Selection 13
     - 2.1.6. Research Requirements for PIPs 14
   - 2.2. Evaluation of an Experimental PIP 16
     - 2.2.1. Method 18
     - 2.2.2. Design of the Four Information Processing Systems 22
     - 2.2.3. Selection and Training of Subjects 26
     - 2.2.4. Results 28
     - 2.2.5. Discussion 31

3. Probabilistic Information Processing Systems with Cumulative and Noncumulative Displays 32
   - 3.1. Method 32
     - 3.1.1. Task and Display 32
     - 3.1.2. Subjects 33
     - 3.1.3. Scenarios 33
   - 3.2. Results 33
   - 3.3. Discussion 35

4. Conservatism in Estimating Probabilities 35
   - 4.1. The Experimental Problem 38
     - 4.1.1. Method 40
     - 4.1.2. Results from the Training Task 45
     - 4.1.3. Results from Deductive Inference Tasks 47
     - 4.1.4. Results from Inductive Inference Task: Inferred Likelihood Ratios 50
     - 4.1.5. Results from Inductive Inference Task: Posterior Odds Comparisons 52
   - 4.2. Discussion 55
   - 4.3. Implications 56

5. Choice Among Bets and Revision of Opinions 58
   - 5.1. The Inference Task 59
     - 5.1.1. Response Modes 60
     - 5.1.2. Subjects and Experimental Conditions 60
   - 5.2. Results 62
   - 5.3. Discussion 63

6. Estimation of Probabilities in Military and Abstract Settings 64
   - 6.1. Method 64
     - 6.1.1. Material 64
     - 6.1.2. Design 64

6.1.3. Apparatus 66
6.1.4. Subjects 68
6.1.5. Procedure 68
6.2. Results and Analysis 71
6.2.1. The Measures Employed 71
6.2.2. The General Pattern of Response 72
6.2.3. The Effect of the Experimental Variables 75
6.2.4. Recent vs. Old Information 75
6.2.5. The Rate of Working 76
6.2.6. General Features of the Subjects' Reaction to the Task 76
6.2.7. Fallacies Exhibited 77
6.2.8. Ways in Which Instructions Were Disregarded 77
6.2.9. Affective Reaction 78
6.3. Discussion 78

References 80

Appendix I: Publications Under Contract AF 19(628)-2823 83

Distribution List 85

FIGURES

1. A PIP for Threat Evaluation 11
2. PDP-1 Computer with a CRT Display 24
3. Final Odds Favoring War 29
4. Final Odds, Cumulative Display Group vs. Noncumulative Display Group 34
5. Scatterplots of Estimated Likelihood Ratios from the Second Deduction Task 48
6. Scatterplots of Inferred Likelihood Ratios 51
7. Posterior Odds as a Function of Number of Draws (Medians Across Subjects) 53
8. Posterior Odds as a Function of Number of Draws (Medians Across Subjects), with ELR_1 and ELR_2 Plots Corrected for Bias 54
9. Map of Enemy Territory Within Battle Area 67
10. Experimental Setup 67
11. Response to the Critical Hypothesis: Mean for all 64 Subjects 72
12. Probability of the Critical Hypothesis Indicated by the Most Responsive and the Least Responsive Subjects 74

TABLES

I. A Functional Analysis of Decision Making 3

II. Characteristics of PIP, POP, PEP, and PUP 27

III. Odds of Each War Hypothesis to the Peace Hypothesis 30

IV. Comparative Efficiencies of the Four Systems 30

V. Distribution of Likelihood Ratios Averaged over Subjects Within Groups 31

VI. Geometric Means of Distribution of Likelihood Ratios 34

VII. Individual Likelihood Ratio Distributions 35

VIII. Reflection in Differing Posterior Odds of Optimal Combination of Data and/or Veridical Deductions Regarding the Data-Generating Process 40

IX. Bigrams Used in the Inference Tasks and the Geometric Means (Across Subjects) of Their Veridical Likelihood Ratios 44

X. Correlation of Each Subject's LVLRs with All Other Subjects' LVLRs 45

XI. Values of c for the Training Task and for the Data Obtained by Phillips and Edwards (1966) 46

XII. Coefficients and Regression Parameters of the Linear Correlations Between Log Estimated Likelihood Ratios (Dependent Variable) and Log Veridical Likelihood Ratios (Independent Variable) for the First and Second Deductive Inference Tasks 47

XIII. Parameters and Statistics of the Bayesian Analysis Described in the Text 49

XIV. Coefficients and Regression Parameters of the Correlation Between Log Inferred Likelihood Ratios and Log Veridical Likelihood Ratios 52

XV. Quadratic Gain Bets 60

XVI. Samples in First Block of Subexperiment One 61

XVII. Samples in First Block of Subexperiment Two 61

XVIII. Accuracy Ratios as a Function of s - f 63

XIX. The Conditional Probabilities and Coding of Events 65

XX. Splitting the Conditional Probability Distributions for S1 and S3 66

XXI. A Priori Probabilities for the Nine Subjects Who Did Not Rate the Hypotheses as Equally Likely 71

XXII. Frequency with Which the Evidence Led to No Adjustment of the Posterior Probabilities During Trials 1-20 73

XXIII. Frequency Distribution of Settings of Probability of Critical Hypothesis on Trial 60 73

XXIV. The Effect of Experimental Variables on Response to the Critical Hypothesis 75

NONCONSERVATIVE PROBABILISTIC INFORMATION PROCESSING SYSTEMS

1
INTRODUCTION

This is the final report of Contract AF 19(628)-2823, concerned with research on probabilistic information processing systems. The research program reported here overlapped in time and content with the program of a previous contract, AF 19(604)-7393. Since originally this was planned to be a more extended program than it turned out to be, it has not been possible to complete all studies that were begun under the present contract. Of those completed, not all are reported here. Some are published or in press as journal articles; a list of these is given in appendix I. Others require additional data before they will be ready for publication; this is particularly true of two experiments on purchasing information. Work on these experiments will continue under other Air Force sponsorship, and will eventually be reported in journal articles.

The main effort under this contract was concerned with large-scale simulation experiments on probabilistic information processing (PIP) systems. Two experiments were completed during the life of the contract. One, a very large and prolonged study of four processing systems, was the heart of the work and is the heart of this report; it is reported in section 2. The main conclusion to be reached from it is that PIP is indeed an efficient philosophy for designing information processing systems — at least twice as efficient as its next-best competitor, and five times as efficient as a system representative of those currently used. The second PIP experiment, reported in section 3, was concerned with a technical problem of PIP design: whether likelihood estimators in PIPs should be allowed to know the state of system opinion. Theory suggests that it might be undesirable, and the data confirm the suggestion.

The comparison between PIP and its competitors, though it clearly indicates that PIP is superior, does not indicate how PIP compares with theoretically optimal performance, since no objective model of the data-generating process was available. Section 4 reports a smaller-scale laboratory experiment that compares PIP with a procedure for estimating posterior odds (POP) in a task sufficiently complex that it was difficult for subjects and yet allowed an objective standard of correct performance; PIP proved to be far superior to POP. PIP and calculations of optimal performance were roughly comparable, with PIP sometimes more and sometimes less extreme than optimal performance.

Section 5 reports a small laboratory study concerned with the development of a response mode in which subjects report on probabilities by making choices among bets. The original purpose of the study was to develop a response mode for the PEP group (personal processing) of the first PIP experiment, but it turned out to be of considerably more general import than that.

Section 6 reports a study initiated under an earlier contract (AF 19(604)-7393), in which the fact of human conservatism in information processing, with which PIP is designed to cope, is again demonstrated, this time under conditions that are realistically complicated and have a military flavor.

The conclusions from this program are simple. People are conservative processors of information. Practical information processing systems that cope appropriately with human conservatism can be designed so that computers aggregate human estimates of likelihood ratios into posterior distributions by means of Bayes's theorem. Such procedures extract information from data more efficiently than does any other way of exploiting human judgment yet tried, and produce data roughly comparable with theoretically optimal calculations when such calculations are possible.

2
THE DESIGN AND EVALUATION OF PROBABILISTIC INFORMATION PROCESSING SYSTEMS*

This section has two parts. The first is a discussion of the notion of a probabilistic information processing system. The information in it extensively overlaps the information in Edwards (1962, 1963), and is included here both to make this report self-contained and to explain the second part of the section. The second part is the first report of a large experiment comparing a Bayesian information processing system (designed according to the principles presented in the first part) with three competitive systems.

2.1. INFORMATION PROCESSING AND BAYES'S THEOREM

The bare outlines of a formal system for processing information and making decisions have been apparent for some time. They start, of course, with a payoff matrix. In order to obtain such a matrix, one must specify acts, one of which will ultimately be executed; states of the world that influence the payoffs obtained by means of these acts, exactly one of which will eventually obtain; and a quantitatively specified payoff for each combination of a state and an act. If any information is available concerning the probabilities of the various states, it is processed by means of Bayes's theorem into a posterior distribution over these states. This posterior

*Research reported in this chapter was conducted under Contracts AF 19(628)-2823 and AF 19(604)-7393.

distribution is used to calculate an expected value for each act and the act with the highest expected value is chosen.

The sketch contained in the preceding paragraph raises many problems of implementation, most of them unsolved. If this approach to the process of making optimal decisions is to be applied to real-world decision-making situations of substantial importance and complexity, at least 13 steps must be taken. Table I summarizes these tasks; identifies whether they are best performed by men, by machines, or by both; and specifies whether, when a system intended to make decisions is being designed, these tasks can be performed ahead of time or must be done at the time the decision is to be made.

The experiment reported here deals with steps 10 and 11. It assumes, therefore, that the prior steps have been executed, and it is not primarily concerned with how to go about executing subsequent steps. It is concerned, then, chiefly with the diagnostic part of the problem of

TABLE I. A FUNCTIONAL ANALYSIS OF DECISION MAKING

	Function	Performed by	When Performed
1.	Recognize the existence of a decision problem	Men	Ahead of Time
2.	Identify available acts	Men	Ahead of Time
3.	Identify relevant states that determine payoff for acts	Men	Ahead of Time
4.	Identify the value dimensions to be aggregated into the payoff matrix	Men	Ahead of Time
5.	Judge the value of each outcome on each dimension	Men	Ahead of Time
6.	Aggregate value judgments into a composite payoff matrix	Machines	Ahead of Time
7.	Identify information sources relevant to discrimination among states	Men	Ahead of Time
8.	Collect data from information sources	Men and Machines	Now
9.	Filter data, put into standard format, and display to likelihood estimators	Men and Machines	Now
10.	Estimate likelihood ratio (or some other quantity indicating the impact of the datum on the hypothesis)	Men	Now
11.	Aggregate impact estimates into posterior distributions	Machines	Now
12.	Decide among acts by using principle of maximizing expected value	Machines	Now
13.	Implement the decision	Men and Machines	Now

diagnosis and action selection. The emphasis is on the design of systems intended for diagnosis and action selection, in part because such systems are interesting and important and in part because they provide a convenient focus for orderly thought about how to make real-world decisions. However, these ideas are not only or even primarily appropriate to military command systems. They are relevant to any setting in which formal diagnosis is important — medical, legal, governmental, and business settings included. In all such settings, the decision-maker must come to grips with uncertainty. He typically feels that he has too little information. Much of the effort spent on uncertainty, therefore, has been spent in providing decision-makers with more and more information. Unfortunately, it has become increasingly clear that more information, while nice to have, is not the answer all by itself. Some way of providing better information would be ideal, of course — in military settings, a copy of the enemy's battle plans would often be just right. But such information is not often available. Abundant and often accurate information about questions only peripherally related to what the decision-maker really wants to know must somehow substitute. The problem of diagnosis is in large part that of making quantity of information substitute for quality.

Why is there no well known technology of diagnosis? A possible answer is that during most of the period in which managerial techniques became responsive to scientific and technological advance, science was under the misapprehension that uncertainty cannot be measured. Recent recognition that uncertainty can be measured in a way that depends on expert human judgment has led to the beginnings of a technology of diagnosis. This paper first presents that technology abstractly and then indicates a possible application to the problem of diagnostic information processing in command-and-control systems. The point of view underlying that technology, based on the personalistic view of probability, has come to be called the Bayesian viewpoint. Although the Bayesian viewpoint has developed rather recently within mathematical statistics, none of the applications considered in this paper are statistical.

2.1.1. THE MEASUREMENT OF UNCERTAINTY. Probabilities quantify uncertainty. A probability, in the definition appropriate to this paper, is simply a number between zero and one that represents the extent to which a somewhat idealized person believes a statement to be true. The reason the person is somewhat idealized is that the sum of his probabilities for two disjoint (i.e., mutually exclusive) events must equal his probability that either of the events will occur. This additivity is the essence of probability theory; its ramifications extend so far that it is quite difficult always to behave consistently with respect to it. It is convenient, both for exposition and as an operational definition of probability, to point out that the probability of event A for you is the amount of money that you will pay me now in return for my trustworthy promise to pay you one dollar if event A happens. (A more precise statement would replace dollars with a measure of utility.)

This simple definition makes the idea of probability as applicable to the task of describing uncertainty about the next heavyweight champion's identity (a unique event) as it is to the result of the flip of a coin (also a unique event, but one for which a relevant relative frequency may be more easily identified). Most of the hypotheses encountered in command-and-control environments are about unique states; probabilities can be assigned to these states only by means of a definition like that in the preceding paragraph. Since that definition makes probability a matter of opinion, probabilities so defined have come to be called personal probabilities (see Savage, 1954).

The hypothesis that the object just seen by BMEWS is a satellite being launched rather than a missile is either true or false; it cannot have a meaningful relative frequency (other than zero or one). But a commander's opinion about which it is may well be describable by a personal probability other than zero or one. If so, the mathematically appropriate rule for revising that opinion on the basis of new information is Bayes's theorem.

2.1.2. BAYES'S THEOREM AND BAYESIAN STATISTICS. The mathematical definition of conditional probability of an event (D) given another hypothesis (H) is:

$$P(D|H) = \frac{P(D \cap H)}{P(H)} \tag{1}$$

unless $P(H) = 0$. Although the events D and H are arbitrary, the initial letters of data and hypothesis are suggestive names for them. The probability $P(D \cap H)$ is the probability of the simultaneous occurrence of two events regarded as one event.

A little algebra now leads to a basic form of Bayes's theorem:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{2}$$

provided $P(D)$ and $P(H)$ are not zero.

In equation 2, $P(H)$ is the prior probability of some hypothesis H. Though not so written, it is a conditional probability; all probabilities are really conditional. It is the probability of H conditional on all information about H available before D is learned. Similarly, $P(H|D)$ is the posterior probability of H conditional on that same background knowledge together with D. $P(D|H)$ is formally the probability that the datum D would be observed if the hypothesis H were true. For a set of mutually exclusive and exhaustive hypotheses H_i, the $P(D|H_i)$ represent the impact of the datum D on each of the hypotheses. Obtaining the values of $P(D|H)$ for each D and H is the key step in applying Bayes's theorem to scientific or military information processing. In statistical applications, $P(D|H)$ typically is obtained by computation from a so-called statistical

model (like the assumption that a set of observations is normally distributed); in the applications of Bayes's theorem that are of interest to this report, $P(D|H)$ typically will be obtained from direct human judgment.

The probability $P(D)$ is usually of little direct interest. Ordinarily it is calculated, or eliminated, as follows. The hypothesis H is one of a list, or partition, of mutually exclusive and exhaustive hypotheses H_i. Since the definition of probability requires that $\sum_i P(H_i|D) = \sum_i P(H_i) = 1$, equation 2 implies that $P(D) = \sum_i P(D|H)P(H_i)$. The choice of the partition H_i is of practical importance but largely arbitrary. For example, tomorrow will be "fair" or "foul," but these two hypotheses can themselves be subdivided and resubdivided. Equation 2 is true for all partitions, but is more useful for some than for others. In principle, a partition should always leave room for some other explanation. Since it would be difficult to obtain $P(D|H)$ for the hypothesis that "some other explanation" is the true one, the catch-all hypothesis usually is handled in part by studying the situation conditionally on denial of the catch-all and in part by informal appraisal of whether any of the explicit hypotheses fit the facts well enough to maintain this denial.

A particularly convenient version of Bayes's theorem for some of the applications to be discussed in this paper is the odds-likelihood ratio form. For two hypotheses, H_A and H_B, and the datum D, Bayes's theorem may be written twice as follows:

$$P(H_A|D) = \frac{P(D|H_A)P(H_A)}{P(D)} \tag{3}$$

$$P(H_B|D) = \frac{P(D|H_B)P(H_B)}{P(D)} \tag{4}$$

It is convenient to divide equation 3 by equation 4, which yields

$$\frac{P(H_A|D)}{P(H_B|D)} = \frac{P(D|H_A)P(H_A)}{P(D|H_B)P(H_B)} \tag{5}$$

or

$$\Omega_1 = L\Omega_0 \tag{6}$$

In equation 6, Ω_0, the prior odds, is simply the ratio of the prior probability of H_A to that of H_B. The ratio $L = P(D|H_A)/P(D|H_B)$ is called the likelihood ratio. The word odds here means exactly what it does at the racetrack, and the notion of likelihood ratio is just what it was in classical statistics. Equation 6 is as valid and appropriate a way of writing Bayes's theorem as is equation 2—and in some applications is considerably more convenient.

It is worthwhile to notice that in equation 6 multiplication of $P(D|H_A)$ and $P(D|H_B)$ by any constant would make no difference to the posterior odds. To put it another way, $P(D|H)$ need

be defined only up to a multiplicative constant. This fact, which is of far-reaching importance for all applications of Bayes's theorem, is known as the likelihood principle (for a fuller discussion, see Edwards, Lindman, and Savage, 1963). Only likelihood ratios, not values of $P(D|H)$ itself, are required in applications of Bayes's theorem. This fact is the basis for the procedure used by the PIP group in the experiments reported in sections 2 and 3.

Resistance to application of Bayes's theorem in scientific contexts has focused on the difficulty of estimating $P(H)$, the prior probability of the hypothesis before the data have been collected. Without examining the reasons why Bayesian statisticians consider that resistance to be misguided, for the purpose of this paper it is enough to point out that in almost any military information processing system, the process of interpreting incoming information is continuous and progressive. Whatever the initial probabilities that the system may have entertained at the beginning of the day, week, or month over which observations are being taken, those probabilities will very quickly (e.g., in five minutes) be swamped by the mass of incoming information, and so can be chosen arbitrarily (so long as the arbitrary choices are not too close to 0 or 1). In any application of equation 2 or equation 3 to the design of information processing systems, $P(H)$ will simply be the output of the previous calculation, $P(H|D)$. Since $P(D)$ can be treated simply as a normalizing constant, it follows that the only information that must be supplied in order to use equation 2 or equation 3 in such a cumulative information processing system is $P(D|H)$.

2.1.3. BAYES'S THEOREM IN DIAGNOSTIC INFORMATION PROCESSING. There are two ways of substituting quantity for quality of data in the diagnostic processing of information. The cumulative character of Bayes's theorem makes it easily applicable to both. If the data are inherently poor so that repeated observations are used to refine estimates, each new observation can simply be processed as it arrives, with the output of the previous calculation being used as the necessary prior probability. The result of each calculation will be the current posterior distribution over the hypotheses with which the inference is concerned, based on all the data so far available. Exactly the same statement applies to inference from several lines of evidence. It makes no difference in the use of Bayes's theorem whether one item of information is qualitatively like or unlike the preceding item; in either case it can be used to modify the current distribution into a new one that represents the appropriate impact of the new information, when it is combined with the old, on the hypothesis under consideration. The usefulness of Bayes's theorem for information processing, then, arises because it is formally an optimal way of aggregating information, whether from one source or from many.

That the items of information reaching a system may be correlated rather than independent has not been explicitly taken into account in the preceding discussion. The topic is important for many applications and difficult; this discussion must be sketchy and incomplete. A correlation between two items of information ordinarily implies that the two are related to some under-

lying fact or process that produces them both. From the point of view of Bayes's theorem, two different cases can be distinguished. One, much the more common, occurs when the underlying fact is the truth of one of the hypotheses being considered. Thus cancellation of all furloughs in the Russian army and recall of all Russian diplomats stationed in the United States "for consultations" are events that one well might expect to see more often simultaneously than at unrelated times, and yet their separate impacts on the hypothesis that Russia is about to start a war are not in any sense diminished by that correlation. Formally, the impact of each is given by a number of the form $P(D|H)$, and in this example $P(D_1|H, D_2) = P(D_1|H)$, and $P(D_2|H, D_1) = P(D_2|H)$. In this case, it is entirely appropriate to treat the data as independent, even though they are correlated.

The other and more difficult case arises when the correlation between D_1 and D_2 reflects something other than the truth of one of the hypotheses being considered. That the suspect in a murder investigation is left-handed and that the fatal blow was delivered from behind and to the left jointly have much more influence on the posterior probability of interest than might be expected from either item of information considered alone. The correlation reflects the fact that left-handed people find it easier to strike from the left than do right-handed people. In the example, the impact of the two items of information taken together is greater than that of both taken separately; it is also possible to construct examples in which the two items taken together have less impact, not more, than both taken separately. If observations are not independent in this way, then the effect of the nth item of information must be considered in the light of all the preceding information with which it interacts. Formally, the number that must be supplied to Bayes's theorem is not $P(D_n|H)$ but $P(D_n|H; D_{n-1}, D_{n-2}, \ldots D_1)$. When such a lack of independence among observations must be taken seriously, provision must be made to ensure that the nature of the dependence is taken into account—usually by human judgment. Fortunately, it often will be possible to ignore such dependencies, especially when one is interpreting data obtained from technical sensors like BMEWS and MIDAS or similarly quantitative data from nonmilitary sources.

Bayes's theorem integrates the output from different sources of information into a single, coherent, orderly picture of what is happening. If the question of interest to the system is an abstract one, such as whether or not the system is being attacked, and if so by whom, that picture will be correspondingly abstract, perhaps as abstract as a table of numbers, a bar graph, or a pie diagram. If the question of interest is more concrete, such as the current location of a missile-launching submarine, the display can be similarly more concrete; a map showing an ellipse within which the system is 90 percent sure the sub is located might be appropriate (Herman, Ornstein, & Bahrick, 1964). When the question is abstract, displaying some of the information on which the system's opinion is based, as well as displaying that opinion itself, may help intuition.

In addition to the central advantage of being able to accept and combine data from as many different sources as seems desirable, Bayes's theorem has several peripheral advantages. One is that it automatically screens information for relevance, filtering out noise, retaining useful information, and automatically weighting each item according to its relevance and importance. If the probability of a datum given any hypothesis with which the system is concerned is exactly equal to its probability given by any other hypothesis, that datum is irrelevant to the system's posterior opinion, which will then be exactly the same as its prior opinion. The extent to which the likelihood ratio $P(D|H_1)/P(D|H_2)$ differs from one is a measure of the effectiveness of D in changing opinion about the relative probabilities of H_1 and H_2. A second peripheral advantage is that Bayesian processing of information requires a minimum of record-keeping. Once a datum has been processed by equation 2, or 6, or some other version of Bayes's theorem, it may be discarded (so far as its use in Bayes's theorem is concerned); its impact on the hypotheses with which the system is concerned has been registered, and it is not required for any future calculations. As this statement suggests, the order in which different data are processed makes no difference to their impact on system opinion; the posterior opinion obtained by observing first D_1 and then D_2 is exactly the same as that obtained by observing first D_2 and then D_1. (These statements require modification if violations of conditional independence may occur.) A third peripheral advantage is that the output of the system is exactly what is required for choosing a course of action on the basis of maximizing expected value; this report returns to this point much later.

No elementary discussion of Bayesian statistics that highlights its relevance to processing qualitative information can be found, though Edwards (1962) has discussed the topic briefly. In 1962 Herman et al. presented an example of a military application and Dodson (1961) speculated on the topic. The only available elementary presentations of Bayesian statistics are two introductory texts by Schlaifer (1959, 1961) which discuss it from the point of view of applications to business problems. There is a small but rapidly growing literature in the statistical journals, but the only technically sophisticated book on the topic, by Raiffa and Schlaifer (1961), is a compendium of Bayesian distribution theory combined with an extensive discussion of Bayesian statistics applied to problems where economic issues are important. An expository paper by Edwards, Lindman, and Savage (1963) is intended primarily for psychologists and concerned primarily with the Bayesian version of null-hypothesis testing. The present report covers some of the same ground as papers by Edwards (1963) and Edwards and Phillips (1964).

2.1.4. A PROBABILISTIC INFORMATION-PROCESSING SYSTEM. Enough has been said above to specify abstractly how Bayes's theorem and Bayesian statistics could be used to process fallible military information. But it is instructive to consider a block diagram that might represent a system intended for that purpose. For specificity, a much-simplified version of a part of the task of the NORAD COC will be used as an example.

As has already been indicated, the prior probability for each Bayesian calculation after the first in any probabilistic information processing system (PIP) will simply be the output of the next previous calculation, and so need not be supplied to the system. Therefore the only number that must be supplied in order to permit application of equation 2 are the $P(D|H)$, likelihood ratios, or similar numbers. The heart of the problem of designing an effective PIP is the problem of obtaining these numbers for the data entering the system. Sometimes such numbers can be calculated, either on the basis of past experience with similar information, or on the basis of some model of the information-gathering process, or both. More often, however, such "objective" procedures are unlikely to be acceptable. Calculation might be entirely acceptable as a basis for inferring the probability that nothing is actually there when BMEWS reports 30 events with low reliability and one event with high reliability. But calculation alone is clearly inadequate to assess the probability that Russia would have launched 25 reconnaissance satellites in the last three days if she planned to attack within the next hour. Interpretation of such information is obviously a matter for human judgment; PIP is a proposal about how to obtain such judgments systematically and how to use them once obtained. Specifically, the hypothesis underlying PIP is that men can serve as transducers for $P(D|H)$; that is, they can be taught to estimate such probabilities (or rather, quantities related to them) with accuracy sufficient to serve as a basis for making decisions even when the probabilities cannot be calculated by any other procedure.

In effect, just such judgments must be made in any deterministic system. How can one judge the impact an item of information has on opinion about the truth of a hypothesis without asking, implicitly or explicitly, such questions as "How likely is it that I would see D if H were true? And how likely is it that I would see D if H were not true?" PIP in effect fragments the task of interpreting data into small, manageable pieces. The system must be provided with one number, $P(D|H)$ or a close relative, for each item of data and nearly each hypothesis considered. Actually, if likelihood ratios are being estimated when five hypotheses are being considered and 100 items of data reach the system, 400 likelihood ratios are needed. Making each such estimate, then, is a relatively small and hopefully a manageable task. It is, of course, a task for experts. They must be expert about the source of data, expert about the hypotheses being considered, and sufficiently expert about likelihood ratios that they know what it means to estimate one. It is an interesting question for experimental study whether one expert on a given source should process all data from that source, estimating its likelihood ratio for each pair of hypotheses with which the system is concerned, or whether an expert on a given hypothesis should process all the data reaching the system, judging its probability given that hypothesis, or whether some intermediate organizational principle might be better than either extreme. This question is an instance of a far more general question, pervasive in modern theory of organizations, about whether an organization should be structured around its inputs, around its outputs, or in some compromise configuration.

Once it is assumed that likelihood ratios will be supplied to the system by means of human judgment, the details of a sample PIP are easy to work out. Figure 1 presents an appropriate block diagram, one of many that might be considered.

The first group of human operators exists only to filter out obviously irrelevant information from the sensor returns, to put what is left into standard and economical format for further processing, and in general to perform low-level data-processing functions of a type familiar in present-day systems. There is no requirement that these functions be performed by men, if appropriate automatic techniques can be used instead. Nor should these functions be performed within the ultimate information processing system. In fact, filtering for irrelevancy should be performed at the sensor site in most instances, in order to reduce the demands on information-transmission facilities.

The second group of operators, called likelihood ratio estimators in figure 1, are the heart of the system. They provide the Bayesian processor (which in all probability would be a large digital computer) with likelihood ratios, or in some systems $P(D|H)$, for each incoming item of information and for each hypothesis with which the system is concerned. Of course, it is not necessary that they provide these judgments in explicitly numerical form; any form from which the computer can perform Bayesian calculations is perfectly acceptable. Identification of the best techniques for extracting such probabilities from operators is an important researchable problem, and some suggestive research is available.

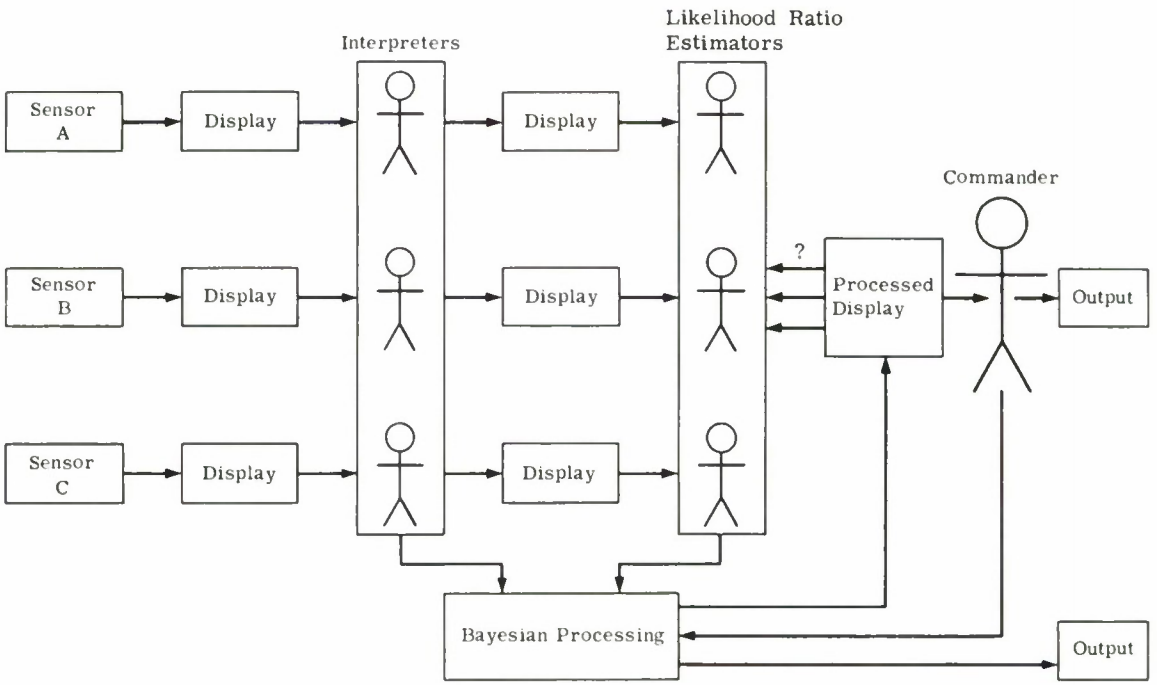


FIGURE 1. A PIP FOR THREAT EVALUATION

Estimates of likelihood ratios or $P(D|H)$, perhaps accompanied by specific information from the sensor itself for some PIPs, are the inputs to the Bayesian processor. The output of that block is a display of the system's current opinion about the hypotheses with which it is concerned—the processed display of figure 1. That display may be any of a number of things, depending on the natures of the system and of the hypotheses to be considered. If the set of hypotheses is finite and small, bar diagrams, pie diagrams, or direct displays of numbers are appropriate. If the set of hypotheses is a continuous, one-dimensional distribution, an ordinary graph is likely to be appropriate. If it is a two-dimensional distribution (location of submarines, computed impact areas), one or more contour lines of equal probability would be appropriate displays. Choice of the best display for a particular kind of hypothesis is clearly also a researchable problem.

The processed display is looked at by the likelihood estimators, so that they can interpret each new item of information in the light of the system's current opinion. This feedback loop makes it possible for the system to be unstable if the estimators pay more attention to the trend of the system's current opinion than to the specific content of the new information coming in. If so, breaking the feedback loop by preventing the likelihood estimators from looking at the processed display is perfectly feasible. In principle, they need only to observe D and to know what hypotheses the system is considering to estimate $P(D|H)$ or likelihood ratios. That is why figure 1 has a question mark above that feedback loop.

The other person who looks at the processed display is an officer of long experience, designated in figure 1 as the commander. It is assumed that he is expert about enemy intentions, tactics, and capabilities. If the system's opinions are patently absurd in the light of what he knows about the enemy, he can tell the Bayesian processor so and thus exercise a sort of veto power over the system's opinion. It is hoped, of course, that this veto power will seldom be used, if ever. Another way of thinking about this man would be to put him into the system as another source of information. In practice, his function in the system may be difficult or impossible to distinguish from the function of the intelligence system's input, and perhaps he should be a part of that block. At any rate, there must be some arrangement for overall human supervision of the system's opinions; he is it.

The output from the system might be the processed display itself, repeated to a location in which action selection is performed. Or it might be a set of decisions made by the commander. Or in very advanced systems, it might be a set of decisions or recommendations generated by the Bayesian processor. The question of exactly what the output is need not be important to the evaluation of a PIP. If the system makes if possible to keep close track of the hypotheses with which it is concerned, any action-selection technique based on its outputs is bound to reflect its effectiveness in processing information—though some reflections may make less use of the

special properties of a probabilistic system than others. At any rate, action selection need not be explicitly studied in research on PIP; it is a separable problem.

2.1.5. ARTICULATION OF DIAGNOSIS WITH ACTION SELECTION. For most risky decisions, the only formal principle that deserves to be taken seriously for the selection of an optimal act is the principle of maximizing expected value. (The minimax principle frequently encountered in the theory of games is so absurdly conservative, and therefore often so severely suboptimal, that it does not deserve serious consideration as a basis for real-life decisions even in games against a hostile opponent except under very unusual circumstances of a sort not worth examining here. For amplification of this controversial assertion, see Edwards, 1954.) Four kinds of information must be available in order to apply the principle of maximizing expected value. First, a well-defined set or list of acts to be considered must be specified. Secondly, an exhaustive set of well-defined mutually exclusive states of the world (hypotheses, diagnoses) must be specified. Thirdly, for each combination of a state and an act, a single number called a payoff must be specified; that payoff is simply the decision maker's judgment regarding the attractiveness of the particular combination of an act and a state. Fourthly, the probability of each state must be specified. If w_{ij} is the utility associated with the jth act when the ith state in fact obtains, then the expected value of the jth act is defined as

$$EV_j = \sum_i P(H_i | D) w_{ij} \tag{7}$$

If the set of states being considered is continuous, the equation is essentially the same; the summation over the H_i is replaced by an integration. Equation 7 is nothing more than the equation for an average, and the principle of maximizing expected value says nothing more complicated than that you should calculate the average payoff for each act and choose the act for which the average payoff is largest.

Equation 7 makes obvious the close relationship between the output of a PIP and the task of action selection. PIP considers the list of hypotheses relevant to selecting action, and it continuously provides as its output the probabilities of each of these hypotheses, on the basis of the data at hand. These probabilities can be combined with payoffs for each of a list of possible acts, and thus the best act can be identified.

Unfortunately, this formally simple picture conceals immense practical complexity. First, PIP cannot invent hypotheses, it can only evaluate their probabilities. The hypotheses it should consider must be supplied to it, presumably by unaided human judgment. Secondly, no mechanism other than unaided human judgment can invent acts to be considered. Of course, these problems of invention also exist with current systems; in these two respects, PIP is neither better nor worse than they. The third and most difficult practical problem is the determination of the entries in the payoff matrix. In typical cases, the problem is not one of too few numbers,

but rather of too many. In every interesting real-life case, the payoff for any particular combination of act and state will be complex and multidimensional, much more easily represented by many numbers than by one. It is a matter of judgment to reduce those many numbers to one, but the judgment need not be unaided. Yntema and Torgerson (1961) showed that a simple linear combination of the dimensions, which neglects their interaction, often will nevertheless be entirely adequate for the purpose of selecting action. When it applies, this simplification can make the task of filling out a payoff matrix relatively tractable.

The problem of action selection needs much research. Fortunately, that research can be separated substantially from research on diagnosis. For that reason, this report concentrates on diagnosis, and considers action selection only in a derived and secondary way. Still, it would be sterile to study diagnosis without keeping always in mind that diagnosis by itself is futile. It is worthwhile to figure out what is going on in the environment only if, at least under some circumstances, one contemplates doing something and needs the diagnosis in order to decide what to do.

2.1.6. RESEARCH REQUIREMENTS FOR PIPs. The research required to translate the general idea of PIP into a specific working system is large in amount and diverse in topic. It ranges from basic laboratory and theoretical research intended to advance the states of several arts on which the designer of a semiautomatic probabilistic system must draw, to applied simulation work intended to establish that such a system will work.

Some previous experiments bear on aspects of the idea. Robinson (1962) showed that in a very simple task in which subjects were required to track probabilities displayed by means of two flashing lights, they performed very well, almost as well as an optimal machine. Similarly, Shuford (1961) showed that in a simple static situation subjects are able to estimate relative frequencies very accurately. A review of much research on choices among bets has led Edwards (1954, and 1961, 1962b) to conclude that men perceive displayed probabilities very accurately, but fail to use them properly when making decisions. Among the most important reasons for this failure seems to be that men consistently prefer some probabilities to others and are willing to pay undue amounts to gamble at their preferred probabilities (see Edwards, 1953, 1954a, 1954b). This suggests the hypothesis that men should not be permitted to combine probabilities with payoffs in selecting action if that function can be automated.

Phillips, Hays, and Edwards (1966) have performed an experiment in which untrained college students were required to estimate the posterior probabilities of four mutually exclusive and exhaustive hypotheses on the basis of direct statements of prior probabilities, displayed evidence, and displayed numerical values of $P(D|H)$. All subjects produced estimates that were far indeed from the appropriate Bayesian posterior probabilities. Furthermore, all deviations from these

theoretical values were of the same general kind: subjects were unable to extract from the information all the certainty that was in it. Thus, when Bayes's theorem indicates that an estimate of 0.95 would be appropriate, subjects will produce estimates in the range 0.50-0.75; when 0.01 is right, the estimate may be 0.15. Phillips, Hays, and Edwards conclude that subjects are unable to extract from information all the impact on opinion that is in it, regardless of whether that information changes opinion in the direction of greater or less certainty. The conclusion implies that gains can be made by automating the task of reaching conclusions from uncertain evidence. (Furthermore, once $P(D|H)$ is available, it makes no sense not to automate the remainder of the task of finding posterior probabilities; only arithmetic is required, and arithmetic should never be done by men when there is a machine handy to do it.)

The Phillips-Hays-Edwards experiment is very complex, and it seemed possible that the conclusion might be different for simpler tasks. So Phillips and Edwards (1966) performed the simplest possible experiment that still contains the essence of Bayes's theorem. They told subjects that each of several bags contained 1000 poker chips. Some bags contained Np blue and $N(1 - p)$ red chips, where N is the number of chips in the bag and p is the proportion of blue chips; other bags contained $N(1 - p)$ blue and Np red chips. For example, five bags were 60-40 blue to red, five others were 40-60. A bag was randomly chosen, chips were drawn from it randomly and replaced, and after each draw the subjects were required to estimate the posterior probability that the bag was one in which blue chips exceeded red. In this very special case only the difference between the number of blue and the number of red chips drawn is important in modifying the prior probability that the true proportion of blue chips in the bag is p into the posterior probability of that hypothesis. Thus the task is exceedingly simple. Nevertheless, subjects performed little better than they did on the task of the Phillips-Hays-Edwards experiment, and made exactly the same kinds of responses as in that experiment. Even values of the difference between number of red and number of blue chips as large as 25, enough to change a prior probability of 0.5 into a Bayesian posterior probability greater than 0.9999, induced subjects to make estimates in the region 0.80-0.90.

Phillips and Edwards also found that various kinds of rewards for accurate probability estimation do improve performance over time, but not enough to eliminate the systematic failure to extract as much certainty as the data permit. Similarly, they compared posterior estimates expressed in probability with those expressed in odds. Estimates expressed as probabilities were more extreme when the response device displayed the probabilities with log odds rather than linear spacing. Odds estimates expressed either verbally or by setting a pointer on a log scale were more extreme than both kinds of probability estimates, but still less extreme than Bayes's theorem. The same kinds of findings appear in experiments conducted at the University of Colorado by Cameron Peterson and his associates (1965a, 1965b), and in much more complex

and system-like experiments performed in the Laboratory of Aviation Psychology at Ohio State University by Feallock and Briggs (1963), Southard, Schum, and Briggs (1964a, 1964b), and Schum, Goldstein and Southard (1966).

A number of experiments and studies bear more directly on Bayesian systems. By far the most extensive study of Bayesian systems to date is that now in progress at the Laboratory of Aviation Psychology by Feallock, Briggs, Southard, Schum and their colleagues.

The major differences between the kind of system they study and the kind discussed here is that they are interested in situations that repeat themselves often enough that it is possible to gather relative frequencies relating D to H for each D and H. In such situations there is no need for an operator to estimate $P(D|H)$, since the relative frequencies are the estimates. Clearly objective counts, when available, are preferable to subjective estimates; equally clearly such objective counts of repeated situations are unlikely to be widely available in military command and control settings. They are, however, widely available in medical diagnosis situations; there is some reason to argue that the results being obtained by Briggs and his colleagues are most directly relevant to medical diagnostic problems. Those results, incidentally, are highly favorable to Bayesian systems, though relatively complex and far too extensive to review here.

Both Ornstein, Herman, and their collaborators at North American Aviation in Columbus and Eady and his collaborators at Naval Electronics Laboratory, San Diego, have been concerned with Bayesian systems in anti-submarine warfare settings. The details are classified, but they too are producing results highly favorable to Bayesian systems. Finally, Kaplan, Newman, and Lichtenstein at System Development Corporation in Santa Monica have studied very simplified situations in which subjects estimate $P(D|H)$ or $P(H|D)$. They have confirmed that posterior probability estimates are conservative, and have found some data that suggests Bayesian calculations based on estimates of $P(D|H)$ produce better values of $P(H|D)$. But their data are equivocal, and they have a number of reservations about their own techniques (see Kaplan and Newman, 1963, and Kaplan, Lichtenstein, and Newman, 1963).

Research on the PIP idea is in progress in several other industrial, military, and academic settings.

2.2. EVALUATION OF AN EXPERIMENTAL PIP

The remainder of this chapter reports an experiment intended to evaluate the idea of PIP.

What constitutes good performance of a diagnostic system? The question is trickier than it appears. A diagnostic system is not directly responsible for making decisions, so it is not appropriate to apply the criterion that the system should make what later turns out to be the correct decision. (Nor would it be wise to apply such a criterion even if it were appropriate.

What action should be taken depends not only on diagnosis, but also on what is at stake. The diagnostic aspects of a decision-making system should be evaluated independently of evaluation of its methods for combining diagnosis with judged or measured values in order to select action.)

The next most obvious thought is that a doctor who says "Chicken pox" when looking at a case of chicken pox is doing a better job of diagnosis than one who looks at the same case and says "Measles." A diagnostic system should come up with the objectively correct diagnosis.

This idea has an appealing simplicity. Moreover, it is readily translatable into language appropriate for an explicitly probabilistic system like PIP: that system is best which gives the highest posterior probability for the hypothesis that eventually turns out to be correct.

But what does "objectively correct" or "eventually turns out to be correct" mean? How can one know that it is actually a case of chicken pox? This can mean only one or the other or both of two things. Either there is some other diagnostic system, presumably more effective than the one being studied, whose diagnoses are being taken as the criterion, or else later, more definitive information is obtainable that, for some diagnostic system or systems we are willing to trust, leaves the proper diagnosis beyond reasonable doubt. That is, in the last analysis all evaluation of diagnoses reduces to agreement between different systems given the same information, or agreement between diagnoses based on earlier and on later information, or both.

Moreover, this criterion of "truth" by agreement with the ultimate diagnosis is slippery. Consider a situation in which the subject must determine which of two bookbags, one containing 70 percent red and 30 percent blue poker chips and the other containing 30 percent red and 70 percent blue, has been selected by tossing a fair coin. Actually the predominantly red bookbag was chosen, but of the first six chips sampled (with replacement) five were blue and one was red. One subject, given the data, estimates that the probability that the predominantly red bookbag was chosen is 0.45, the other estimates that it is 0.05. Bayes's theorem would say that it is 0.03. Which subject is doing the more effective job of diagnosis? The 0.45 subject assigns higher probability to the truth, and yet few would argue that he is in fact making better use of the data available to him.

A more nearly sensible criterion would assert that that diagnostic system performs best which, in situations in which all systems end up in agreement about the right diagnosis, reaches that diagnosis on the basis of least information—provided that that system finds it no more difficult to change its opinions in response to a change in the trend of the data than do its competitors.

This criterion can be applied in situations in which definitive information is eventually available, and in situations in which it is not. But situations in which the evaluator establishes by fiat the truth of some hypothesis, and then evaluates diagnostic systems by the extent of

their agreement with his fiat, present a special problem. When the true-by-fiat hypothesis is linked with the data by means of some objectively specified and well-understood data-generating process (as in the bookbag and poker chip example), only the possibility of bad luck in the operation of that process need be worried about. But in the far more numerous and more interesting cases in which the data cannot be generated by an objectively specified data-generating process, the link that the evaluator supposes to operate between the true-by-fiat hypothesis and the data may not exist for any eyes but his — or, worse yet, may exist only in the opinions of those whose biases, theories, or other opinions happen to agree with those of the evaluator. This argument suggests that in the absence of an objectively specifiable data-generating process the criterion of agreement with the "truth" is a potentially misleading criterion of performance for judging a diagnostic system. Comparison of the diagnostic efficiency of competitive systems that eventually agree about which hypothesis is favored by the evidence is a more humble but less problematic and far more often applicable method for evaluating diagnostic systems.

These considerations interacted with the basic idea of PIP to specify the broad outlines of the experiment. We sought a situation of great complexity, in which data that could not be associated with any quantitative model of the data-generating process could be obtained in quantity. The data should be quite inconclusive; any diagnostic system worthy of the name should do well with good data, so to tell good systems from bad ones you should study how they do with poor data. We did not want to spend large amounts of effort simulating information-gathering systems; we were interested only in processing information. To avoid troubles with lack of conditional independence, we wanted successive observations to be relatively unrelated to one another; this implied that we wanted each scenario (sequence of data) to cover only a short period of simulated time. We wanted interpretation of data to depend on expert judgment — yet we could not use pre-existing experts, since in that case we could not control what they knew. Finally, we wanted to compare the best competitive systems we could dream up with PIP. We ended up studying three competitive systems, POP, PEP, and PUP (a small experiment added for comparison with PEP). We devoted at least as much attention to the design of each of them as to the design of PIP, in the attempt to ensure that each was the best representative of its philosophy of processing information that we could devise. The design of PEP, which was intended to be the nearest we could come to the way diagnostic information processing is done now, gave us a lot of trouble, and required almost six months of pretesting.

2.2.1. METHOD

2.2.1.1. Setting and Hypotheses. The experiment was set in 1975. The world of 1975 is much simplified compared with that of 1964, when the experiment was begun. Only six nations play significant political and military roles in it — China, Japan, North America, Russia, the

United Arab Republic, and the United Confederation of European States (UCES). China is an aggressively expansionistic Stalinist dictatorship. Russia has Stalinist and more peaceful factions; the latter currently but uneasily rule, but Premier Balinin has worked with both factions in the past. Russia and China fought a border war in 1969; China won, and took some Mongolian territory. Informal communications exist between Russian Stalinists and the Chinese government. During the 1969 war, the former Russian satellites joined with the rest of Europe (except Scandinavia and Ireland) to form the UCES, a loose economic and military confederation with a premier and parliament and a unified military command over nationally segregated units. The old national rivalries create continuous unrest within the UCES; only economic and military necessity keep it together. The military necessity results from the death of NATO in 1968, and the resulting need to keep the UCES together for protecting the former Russian satellites against Russia's possible reversion to a hard Stalinist line. Japan is a fat target: the richest trading nation in the East, with no military forces whatever. North America, an alliance of the U.S.A. and Canada, has unified military forces and is politically dominated by Washington and Republican President Goldneyfeller. The UAR, which reaches from the Atlantic to India, is dominated by the activities of Hadj Bey, a Moslem prophet, evangelist, and religious reformer, who has sparked a Moslem revival accompanied by pro-UAR civil unrest in the Islamic communities in countries surrounding the UAR, especially southern Russia. The UAR also covertly encourages semi-piratical interferences with UCES Mediterranean shipping.

The 50-page summary of the history of the world, 1964-1975, of which the preceding paragraph is a condensation, was designed to make eight hypotheses plausible, and to make it appropriate to treat them as mutually exclusive. They were:

H1. Russia and China are about to attack North America.

H2. Russia is about to attack the United Confederation of European states.

H3. Russia is about to attack the United Arab Republic.

H4. China is about to attack Japan.

H5a. Russia and China are about to embark on a major conflict with each other.

H5b. A revolution is about to break out in the United Confederation of European States.

H5c. The United Confederation of European States is about to attack the United Arab Republic.

H6. Peace will continue to prevail.

Actually, the subjects never heard of H5a, H5b, and H5c. Instead they were presented with H5: Some other major conflict is about to break out. Thus the scenario writers could prepare scenarios that looked markedly different from one another and yet that favored H5, the catchall hypothesis. In all of these hypotheses, "is about to" means that the event will occur within 30 days.

The information-processing system, located in the basement of the Pentagon, serves the Joint Chiefs of Staff. The six hypotheses are supplied to it by the Joint Chiefs, who specify that only these possibilities are to be considered and that they are to be treated as mutually exclusive. The subjects are the duty operators for the 3 PM to 6 PM shift on April 5, 1975.

Military technology in 1975 much resembles that of 1964, as a result of the depression of 1966-1970, which throughout the world nearly terminated research relevant to military purposes. For the same reason, armed forces are generally smaller in 1975 than in 1964, though not much. In particular, neither anti-missile missiles nor military applications of space technology have advanced since 1964, except for the development of a photo-reconnaissance satellite system.

2.2.1.2. Sensors and Data. Three sensors deliver data to the information processing system: the Ballistic Missile Early Warning System (BMEWS), the intelligence system, and the reconnaissance satellite system.

BMEWS is a very large computerized radar system with three sites, one at Clear, Alaska; one at Thule, Greenland; and one which is actually located at Fylingdales Moor, England, but which we moved to Drogedde, Ireland (not a part of the UCES) after the breakup of NATO. The overlapping coverage of these radars permits detection of any ballistic missile fired from anywhere inside Russia or inside the Arctic Circle toward any part of North America or western Europe. The real BMEWS is too reliable a system for our purposes. We degraded it, both in order to avoid revealing classified figures and in order to prevent it from giving information that was too conclusive.

A BMEWS report from Clear or Thule (Drogedde works on a slightly different basis, but produces the same sort of information) consists first of G (green), Y (yellow), R (red) or R-M (red-maintenance). Green means that the system at that site is working properly. Yellow means that the site is producing peculiar results; they may be attributable to environmental radio interference or to a malfunction within the system. In any case, the site is still producing somewhat meaningful, though less reliable, data. Red means that the data obtained from the site are worthless and will not be passed on, either because of severe environmental noise, enemy jamming, or malfunction. If a malfunction is detected, the classification is changed from red to red-maintenance until it has been corrected. BMEWS cannot distinguish between natural environmental noise and jamming, and it is not known whether or not the Russians are doing any jamming. The Russians have not attempted to spoof BMEWS (that is, to fool it into thinking that missiles are coming when they are not), but it is known that they have appropriate equipment to do so if they wish. Following a G or a Y, there is a number ranging from 0 to about 30 (there is no formal upper limit). This is the number of low-reliability events the system has detected in the last 15 minutes. A low-reliability event could be anything; meteors, satellites, (though

these are removed from the count if recognized as such), environmental noise, or an ICBM. Following the low-reliability events is a listing of high-reliability events, along with computed impact areas. A high-reliability event is an event that looks as if it is an ICBM, as the result of complicated radar sensing techniques and highly sophisticated information-processing. Actually, because we found in pretests that high-reliability events were completely disruptive to the simulation, we used none in the experiment itself. Pretest subjects, given a high-reliability event with New York at the center of the impact area, couldn't accept the fact that the next item from the intelligence system did not report the obliteration of New York.

A typical BMEWS report would look as follows:

I. G3
II. G0
III. G1

Such a datum would be essentially neutral with respect to all hypotheses, since while it indicates that nothing warlike is happening, it cannot preclude the possibility of a rain of missiles fifteen minutes from now. BMEWS cannot decrease the probability of any hypothesis except Peace, and cannot increase the probability of any hypotheses except H1 and H2.

The intelligence system consists of spies, military attaches in U. S. embassies abroad, readers of foreign newspapers, experts on foreign affairs, and the like. Each datum from this system is passed on in the form of a short paragraph.

Report of any event is usually accompanied by a brief qualitative statement about the degree to which the event is surprising and about what it might mean.

A typical intelligence system report might look as follows:

> Judging from a careful study by our agents of the production of Soviet parachute factories and military boot shops, our military panel estimates that Soviet paratroop units have been increased by about 20 percent in the last eight months.

The photo-reconnaissance satellite system consists of some large but unspecified number of satellites, renewed sufficiently often that every spot on the globe is photographed at least once every six hours. The photographs have admirable resolution: an automobile can be easily discriminated from the road on which it is moving. But the system cannot see through clouds and can only photograph lights at night. More seriously, it is plagued by a severe shortage of photo-interpreters. At least 99.5 percent of all photographs taken are never looked at. Almost all satellite reports include background information of the kind that might be obtained by comparing a recent photograph with previous ones, or that in some similar way might be available to a photo-interpreter. A great many of the satellite data items were reports of naval movements — mostly because the man who wrote them happened to be a former Navy officer.

A typical satellite system report might look as follows:

> At 0630 this morning, two squadrons of conventional submarines sailed from Vladivostok. They steamed in a southerly direction until they were clear of the harbor and then submerged. Evaluation: Probably routine exercises, though this is an unusually large force.

2.2.1.3. Scenarios. A total of 777 data items were prepared. Of these, 240 were from BMEWS, 250 were from the intelligence system, and 287 were from the satellite system. The item writers were instructed to make sure that no item was conclusively for or against any of the hypotheses, and to make every effort to ensure that any item could appear with any other in either order without violating logic. (Thus, for example, an intelligence item reporting assassination of President Goldneyfeller would have been excluded, since it might be followed by an item reporting a meeting between him and Premier DeBerry of the UCES.) All proposed items were screened for such contradictions, for plausibility, and for intelligibility of wording; at least a third of those originally proposed were discarded or drastically rewritten.

We set out to assemble 18 scenarios, each consisting of 60 items of data. Of course some items were used in more than one scenario, and some were never used; a training scenario was prepared from the latter. We wanted to have six nondiagnostic scenarios (that is, scenarios that did not strongly point toward any hypothesis), six mildly diagnostic ones, and six strongly diagnostic ones. We used our intuition to assemble them. It turned out that the nondiagnostic scenarios were indeed nondiagnostic; most of the mildly diagnostic ones seem, in the light of the data, to have been nondiagnostic also. The strongly diagnostic scenarios were, in the light of the data, mildly to strongly diagnostic, and in each case favored the hypothesis that we had intended to favor. We set out to make sure that no scenario was really conclusive; we succeeded too well. Later experiments should use less ambiguous scenarios, and perhaps fewer utterly meaningless items.

2.2.2. DESIGN OF THE FOUR INFORMATION-PROCESSING SYSTEMS. When the situation and data items were developed, we thought that the PIP subjects would simply estimate $P(D|H)$ for each D and H. So six of the experimenters tried it, for all 777 data items. We soon found that we could not do the task at all. The difficulty was that the level of detail with which a datum was specified was of greatest importance in controlling $P(D|H)$, regardless of whether the detail was diagnostically relevant or not. Thus, an intelligence system datum reading "The cake served at Mme. DeBerry's reception for members of the diplomatic corps last night was decorated with a UCES flag surrounded by candied flowers" would be more probable on any hypothesis than one that read "The cake served at Mme. DeBerry's reception for members of the diplomatic corps last night was decorated with a UCES flag surrounded by candied eidelweiss," even though the extra detail has no diagnostic significance whatever. Contemplation of the likelihood princi-

ple of Bayesian statistics should have led us to predict this fact, and did lead us to cope with it. That principle says that such diagnostically irrelevant information, though it controls $P(D|H)$, is irrelevant to Bayes's theorem. Since likelihood ratios are not affected by such irrelevancies, the obvious solution is to estimate them rather than $P(D|H)$.

For six hypotheses, there are fifteen pairs, and so fifteen likelihood ratios per datum. However, only five of these are independent of one another; given any five involving all six hypotheses, all others can be calculated. Obviously H6, Peace, has a special status in the list of hypotheses. So we tried estimating likelihood ratios comparing each war hypothesis with H6, found that this worked well, and so settled on it as standard procedure for PIP. The three other groups also worked with five pairs of hypotheses, in which each hypothesis was paired with H6.

The four groups had other characteristics in common. Each was a one-man system. Each made five responses per datum. Each started each scenario with odds of 5:1 in favor of Peace compared with each of the other hypotheses—or 16.7 and 0.1667 (the equivalent numbers) for PEP and PUP respectively. (PIP operators, of course, needed and had no prior distribution, but the computer required it in order to use the likelihood ratio judgments.) Each subject saw each datum by projection from behind the computer display and controlled the slide projector by means of a button. Any subject could look at any datum as long as he wished, but each slide remained on the screen for at least a minute. Thus, a 60-item scenario could not be finished in less than an hour; 80 to 90 minutes was typical. A subject completed one scenario per session, and had no more than one session per day.

2.2.2.1. PIP Task and Instructions. Subjects in the PIP group estimated five likelihood ratios per datum.

Training of course was extensive. The key points in about two hours of instructions were: Suppose you are at the moment considering a datum and are estimating the first likelihood ratio, which compares H1 with H6. First ask yourself whether this datum is more likely to have occurred if Russia and China were about to attack North America, or if peace were to continue to prevail. After you have decided that, then ask, in a ratio sense, how much more likely. These questions may be difficult if the datum is obviously linked with a third hypothesis. But when working with, say, H1 and H6 you must pretend for the moment that these are the only two possibilities. Thus, if the datum is that Russian troops have crossed the UAR border in force, and you are estimating the H1-H6 likelihood ratio, you must ask yourself if they would be more likely to have crossed the UAR border if Russia and China were about to attack North America, or if peace were continuing to prevail. The linkage between the datum and H3 is irrelevant until you are considering the H3-H6 pair. Remember, you are not interested in whether the hypothesis is unlikely or likely; you should never allow yourself to get confused and think about the relative

likelihoods of the hypotheses instead of the relative likelihoods of the datum in the light of the two hypotheses.

H5, the catchall hypothesis, presented a special problem for the PIP group. They had to be told qualitatively the quantitative fact that the current probability of each hypothesis within the catchall affects the likelihood of the datum given the catchall. Thus if the datum is relatively highly likely given some possibility included within the catchall, but that possibility is itself highly unlikely, then the datum is not very likely given the catchall. This took a lot of explaining, but the subjects seemed to catch on finally.

As had been anticipated, the other major problem within the PIP group was to teach the two points emphasized in the instructions: that they should not think of the probabilities of the hypotheses, and that they should treat the pair of hypotheses currently under consideration as the only possibilities while estimating the likelihood ratio associated with them. It took a fair amount of talk and practice, but both points did prove communicable.

The PIP group used as apparatus a PDP-1 computer with a cathode-ray-tube display. At the bottom of the display were five levers, each of which slid along a scale. Figure 2 shows the layout.

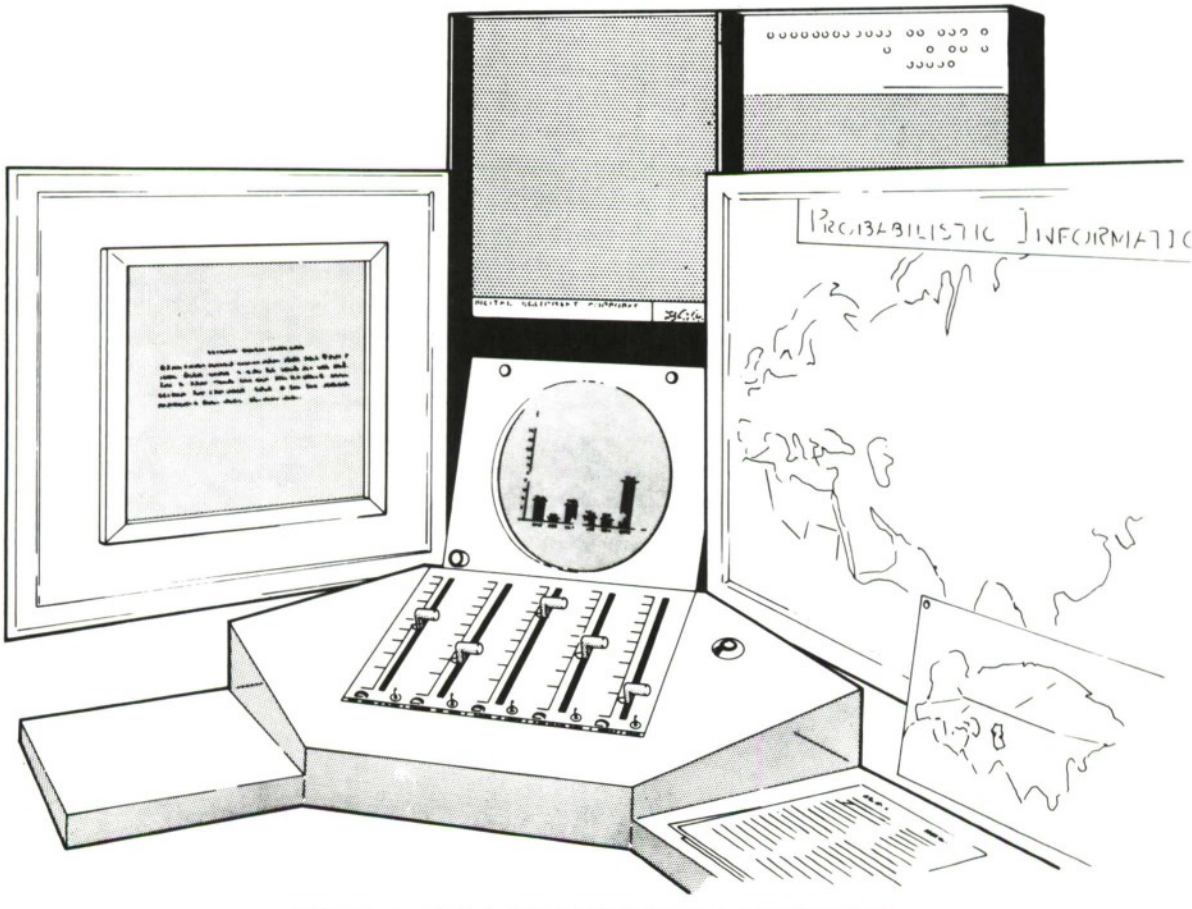


FIGURE 2. PDP-1 COMPUTER WITH A CRT DISPLAY

The scales are in likelihood ratio, taken as numbers equal to or greater than 1, and logarithmically spaced. For each scale, the subject could turn a knob to select one of six different ranges of likelihood ratios. The first range extended from 1:1 to 10:1. Five more logarithmically spaced ranges were available, running from 10:1 to 100:1, 100:1 to 1000:1, and so on up to 1,000,000:1. No subject in the PIP group, however, ever used any of the higher ranges. Associated with each lever was a switch. In the left position, it indicated that the datum was more likely on the other hypothesis than on H6; in the right position, it indicated the opposite. The subject pushed a button to the right of the levers to indicate to the computer that he was satisfied with his judgments; at that point, the computer recorded them and changed the slide, provided that a minute had elapsed since the last slide change.

The display on the cathode ray tube (CRT) above the levers was a display of posterior probability based only on the current estimates and equal prior probabilities. Thus it was simply a transformation on the current set of likelihood ratio judgments. It changed dynamically as the subject moved the levers, reset the switch indicating in which sense the likelihood ratio was to be interpreted, or changed scale range. A trial number also appeared on the screen. The subject compared that with the number that appeared on the slide showing the current datum in order to make sure that computer and slide projector were in agreement. Of course nothing in the basic idea of PIP requires such a display; it served primarily to give the subject two different ways to visualize the meaning of his responses.

2.2.2.2. POP Task and Instructions. The POP task somewhat resembled the PIP task, and used the same apparatus. Subjects were instructed that after each slide they were to estimate the posterior odds in favor of each hypothesis on the basis of all the data in the scenario so far. However, they left their levers in place after each judgment, so that instead of actually re-estimating the odds, they revised them up or down after each new datum. This was more nearly in the spirit of PIP than of POP, but was the only reasonable way to run such a system, since otherwise the subjects would simply remember or note down their previous judgments. Subjects started each scenario by setting levers (and so the display) to 5:1 in favor of peace compared with each of the other hypotheses. Since they did not reset levers to zero between trials, the display was always of the system's current posterior probabilities.

Two problems stood out in training POP subjects. One, as with PIP subjects, was to emphasize that when working with H1 and H6, say, they were to pretend for the moment that those were the only two possibilities, and estimate the odds between them. The other was to make clear to subjects that data sufficient to change odds from 1:1 to 2:1 would also be sufficient to change them from 100:1 to 200:1; that is, that data affect odds in a multiplicative way. (From this statement alone, a mathematically minded subject could have deduced Bayes's theorem; none did.) The logarithmic spacing of the scales helped communicate this idea; it was easy to

understand that a given item of evidence should produce the same amount of change in lever position regardless of where that position started. POP subjects did make extensive use of the 10:1-100:1 scale, but not of any of the higher ranges.

2.2.2.3. PEP Task and Instructions. The PEP group was by far the hardest to design. We wanted it to be as near as we could come to a system in which the commander looks at the data and decides what to do. Originally we intended to present subjects with a payoff matrix, have them choose an act, and from that infer their probability distribution over the hypotheses. Unfortunately, choice of an act from a payoff matrix is a very information-destroying transformation on a vector of probabilities. We calculated that in order to recover six probabilities with acceptable accuracy we would need to have subjects rank-order several hundred bets—an absurd task. Our thinking about the problem went through many phases and many pretests; some of this work is detailed in section 5 of this report. The task we finally used was relatively simple. Subjects were told to imagine that any war, if it broke out, would cost them 100 in some arbitrary unit of value. Peace costs nothing. What would the subject consider to be a fair price for an insurance policy that would pay the 100 in the event of a particular war, and nothing in the event of peace? That number was his response. Of course he made five such responses per datum, and was required when making each one to pretend that either the war he was considering at that moment or else peace were the only two possibilities. The PEP and PUP subjects did not use the computer; they wrote their responses on pieces of paper and had no display.

The only new problem that arose during the instruction and training of the PEP group was to explain that a datum that would produce a relatively large change of price for insurance when the prior price was near 50 would produce far less change when that prior price was near 0 or 100. No actual numbers were given to illustrate the extent of the difference, but enough practice was given to ensure that the principle was qualitatively clear.

2.2.2.4. PUP Task and Instructions. The PEP response, of course, is a disguised probability estimate: it is 100 times the probability of war, given that a particular war or peace are the only two possibilities. For comparison a fourth group, PUP, was run, more or less as an afterthought. PUP subjects had the same task as PEP subjects, except that it was presented as a probability estimation task rather than as an insurance pricing task. Their training and procedures were essentially identical to those of PEP subjects. Table II summarizes the characteristics of the four systems.

2.2.3. SELECTION AND TRAINING OF SUBJECTS. Our goal was to have subjects who were intelligent experts about the world, the sensors, and their own information-processing systems. We started by obtaining (by advertisement) 75 male juniors, seniors, or graduate students,

TABLE II. CHARACTERISTICS OF PIP, POP, PEP, AND PUP

System	Response	Display	Aggregation by
PIP	5 likelihood ratios per datum	Posterior probabilities based on uniform prior and current datum only	Computer
POP	5 posterior odds per datum	The system's current posterior probabilities	Man
PEP	5 fair prices for insurance against specific war per datum	None	Man
PUP	5 probabilities of war given that only war or peace are possible per datum	None	Man

with at least B averages, who were not majors in mathematics, experimental psychology, history, or political science. (We paid at the rate of $2 per hour throughout the experiment, so that we could be very selective about subjects.) We spent about ten hours training them in the history and current political and military characteristics of the world of 1975, and in the characteristics of the sensor systems. After that, we administered the most difficult two-hour objective examination that we could devise, covering all aspects of their training. However, some of the subjects did not complete the training and did not take the test. We reduced the number of subjects to the 36 top scorers on the examination. We later lost two subjects from the PIP group, so the final number of subjects was 34. We assigned ten subjects each to the PIP, POP and PEP groups, and six to the PUP group. Each subject received about six hours of additional instruction in the characteristics of his own information-processing system, including the opportunity to work through one 43-item training scenario. For all subjects, much effort went into assuring them that in our experiment as in real life most data are just about meaningless. For example, the PIP subjects were told that a likelihood ratio as big as 2:1 is produced only by a rare and relatively diagnostic datum. All subjects were taught to expect that by far the most frequent judgment would be that the datum does not change the status of the pair of hypotheses.

Of the many problems that arose during training, perhaps the most severe concerned the lack of sequential structure in the data. Subjects found this unnatural and unfamiliar. They found it helpful, however, to be reminded that on a news wire the successive stories are typically unrelated, and it is necessary to watch such a sequence of stories for a long time before one comes along that develops out of an earlier one. The shortness of the period of simulated time during which they would be working was emphasized to them.

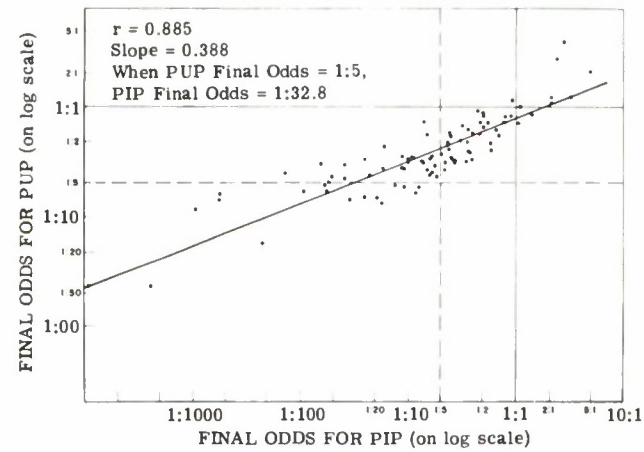
All subjects found the situation and task interesting and demanding in the early stages. After they had become familiar with their tasks, the minute wait between one datum and the next became annoying to them; 30 seconds or less would have been better.

2.2.4. RESULTS. The basic finding of the experiment can be presented in three scatterplots. They show the final odds in favor of each war; that is, the five odds after the 60th datum in each scenario. For the POP, PEP, and PUP groups, each individual's response after the 60th datum was converted from whatever metric it was in to log odds; here and elsewhere in this experiment, all means of probability-like numbers over subjects, scenarios, or data items within scenarios were taken in log odds or log likelihood ratios, regardless of the nature of the numbers being averaged. Figures 3a, 3b, and 3c are scatterplots comparing final odds for PIP with final odds for the other three groups. Though they differ in detail, the main finding is the same for all three comparisons. Note first that the correlation between final odds for PIP and final odds for each of the other groups is 0.85 or higher. This means that the qualitative agreement between PIP and the other groups is excellent; evidence that leaves a hypothesis favored or unfavored for PIP does so for POP, PEP, and PUP also. Next note the slopes of the regression lines. The largest of them is 0.422. This means that quantitatively, PIP is responding more vigorously to the scenarios than is POP, PEP, or PUP. Evidence simply moves PIP more than it does the other groups. In other words, as was predicted, PIP extracts more certainty from the data than do the other groups.

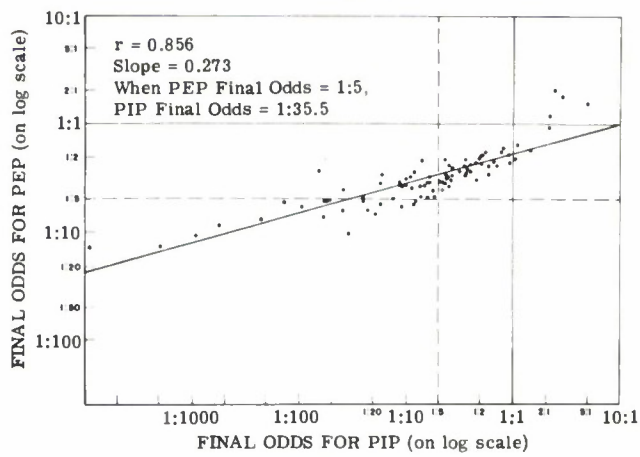
The final point to be made from figures 3a, 3b, and 3c has to do with the origin. Actually, the origin of each graph is at 5:1 in favor of peace, since those were the prior odds. But none of the regression lines pass through that point; instead, they all pass to the left of it. In other words, by comparison with the other groups, PIP has a bias in favor of peace. There is no way of knowing whether this is a result of individual differences, or whether when the hypotheses are as value-laden as are these, PIP and the other groups will exhibit bias to different extents. Indeed, in the absence of some criterion for objective correctness, it is impossible to tell which group is most nearly unbiased. At any rate, it is obvious from inspecting the graphs that although PIP is biased in favor of peace, its efficiency is so much greater than that of the other groups that it has already passed them by the time they reach 1:1.

That the axes of figure 3 are logarithmic implies that the effects seen there are very substantial. Table III is calculated from the regression equations. Suppose a scenario led PIP to give certain odds for or against war; what odds would the other groups give? (Remember that the origin is at 5:1 in favor of peace, so that 99:1 in favor of war is a much larger distance from the origin than is 99:1 in favor of peace.)

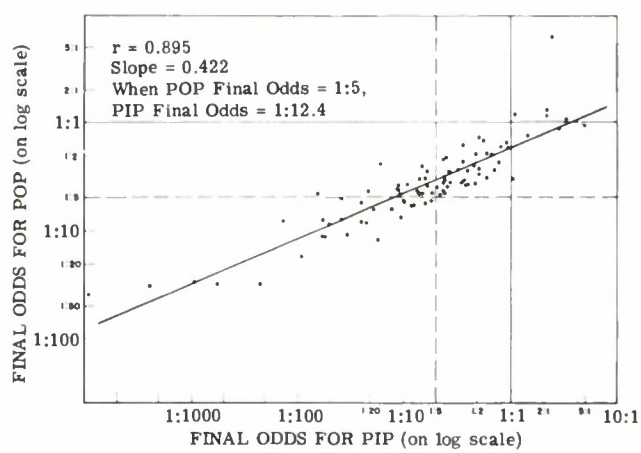
Figure 3 and table III combine to point out that POP is next in efficiency to PIP—a fact slightly concealed because POP has a peace bias relative to PUP. PUP is almost as efficient as POP. PEP, our attempt to simulate the way diagnostic information-processing is done now, is a poor last. For completeness, table IV summarizes the correlations, regression coefficients, and intercepts for the POP-PEP, POP-PUP, and PEP-PUP scatterplots. Each regression coefficient uses the first-named group as the X, or predictor, group.



(a) PUP vs PIP



(b) PEP vs PIP



(c) POP vs PIP

FIGURE 3. FINAL ODDS FAVORING WAR

TABLE III. ODDS OF EACH WAR HYPOTHESIS TO THE PEACE HYPOTHESIS

PIP	POP	PEP	PUP
99:1	4.0:1	1.9:1	4.6:1
19:1	2.0:1	1.2:1	2.4:1
1:1	1:1.7	1:1.9	1:1.3
1:5	1:3.4	1:2.9	1:2.4
1:19	1:6.0	1:4.2	1:4.0
1:99	1:11.9	1:6.6	1:7.7

TABLE IV. COMPARATIVE EFFICIENCIES OF THE FOUR SYSTEMS

	Correlation	Regression Coefficient	Intercept
POP-PEP	0.93	0.63	1:1.36
POP-PUP	0.91	0.84	1.15:1
PUP-PEP	0.93	0.67	1:1.64

A total of 206 items of data were used more than once. Of them, 82 were repeated once, 46 twice, 30 three times, 21 four times, 26 five times, and 1 six times. As a measure of reliability, standard deviations of the log likelihood ratios for each item were calculated. Using logarithms to the base 10, those standard deviations ranged from 0.0031 to 0.0236. For 27 of the 34 subjects, the mean standard deviation was less than 0.0115. These numbers are all very small, and indicate a relatively high degree of test-retest reliability for all groups.

A natural question for any sequential information processing task is whether the current state of the odds (or price of insurance, or whatever) influences judgments. To examine this question, we calculated regressions of log likelihood ratios associated with each datum on the log odds preceding presentation of that datum. The range of these regression coefficients is from -0.147 to 0.082. For 28 of the subjects, that range was between -0.03 and 0.03. These numbers are gratifyingly close to zero. Not surprisingly, these regression coefficients were closer to zero for PIP, which had no display of current odds, than for the other groups.

The way in which PIP produces its superiority over other groups is most clearly exhibited by the distribution of likelihood ratios. Table V shows that distribution. For all groups, likelihood ratios near (in fact, at) one are in the overwhelming majority. But PIP produces six to ten times as many likelihood ratios above two as do the other groups.

Table V also highlights an experimental flaw. We set out to produce relatively undiagnostic scenarios—and we succeeded far too well. As a result, we know little about how these groups might perform in the presence of highly diagnostic data. While it seems unlikely that PIP would be actually inferior to other groups in the presence of such data, we have no evidence to exclude that possibility.

TABLE V. DISTRIBUTION OF LIKELIHOOD RATIOS AVERAGED OVER SUBJECTS WITHIN GROUPS

	Value of the Likelihood Ratio				
GROUP	1.1:1 to 1:1.1	1.1:1 to 1.4:1 and 1:1.1 to 1:1.4	1.4:1 to 2.0:1 and 1:1.4 to 1:2.0	2.0:1 to 2.5:1 and 1:2.0 to 1:2.5	Over 2.5:1 Under 1:2.5
PIP	4875.0	274.9	145.1	56.2	49.0
POP	5018.0	302.4	63.5	9.5	6.6
PEP	5006.1	331.0	52.9	4.7	5.3
PUP	4979.6	340.6	65.5	8.2	6.2

We calculated intercorrelations of log likelihood ratios for individual items between all pairs of groups. As figure 3 would lead us to expect, the groups agreed very well, in a qualitative sense, about which way a datum pointed. If all data are considered, the lowest intercorrelation is 0.68 and the highest is 0.84. If only those data items for which at least one subject gives a likelihood ratio different from 1:1 are considered, the number of items entering into the intercorrelation is reduced from 5400 to about 4800 for comparisons including PIP and about 3600 for other comparisons, but the intercorrelations remain unchanged.

2.2.5. DISCUSSION. The data say overwhelmingly that the original expectation that PIP would extract more certainty from information than do its competitors is correct. They show the effects of some minor bias, but of course cannot show whether PIP, the other groups, or all of them are biased. They also show that the other two probabilistic groups, POP and PUP, are next to PIP in efficiency, while PEP, the attempt to approximate what is done now, is in last place.

The magnitude of the difference between PIP and the other groups surprised all experimenters. We had hoped that PIP might be 10 percent or 20 percent more efficient; instead, it is more than 100 percent more efficient than its nearest competitor, and close to 400 percent more efficient than PEP. (These percentages are obtained by finding the percentage by which the slope of the regression line would have to be increased to make it one when PIP is compared with another group). On no clearly defined basis, some of the experimenters believe that with scenarios that are more diagnostic the difference between PIP and the other groups should be even larger. This prediction seems intuitively likely for PEP and PUP, since they have difficulty with prices of insurance, or probabilities, near one or zero. It seems less likely for POP, which has no similar boundaries on its response scale.

The two points at which this experiment most severely departs from realism are in the arbitrary listing of hypotheses and the nonsequential nature of the data. The listing of hypotheses,

though arbitrary, is quite plausible as a structuring procedure in a military setting. Of particular interest is what should be done when the catchall hypothesis achieves high posterior probability. A later experiment should explore the problem of differentiating specific hypotheses out of the catchall and then reprocessing the data in terms of these specific hypotheses. Of course, the nonsequential nature of the data is by far the least realistic feature of this experiment. Moreover, an argument can be made that PIP, a system built around nonsequential procedures, is especially well suited to processing nonsequential data. The experiment should be repeated with progressively unfolding scenarios. If it were, evidence collected in a PIP simulation at Ohio State University suggests that PIP would still look best, though perhaps not by so overwhelming a difference as in the present experiment (Schum et al., 1966).

A lot of somewhat arbitrary decisions went into the design of our experiment. Among them are the logarithmic response scales for PIP and POP, the use of probability as a display, the decision not to let the PIP subjects know the current state of the system's opinion, and so on. None of these decisions are discredited by the result of the experiment or by experience gained in running it, except the decision to require at least one minute between successive items. Similar comments apply to the essentially ad hoc training procedures used. But all of these variables could be studied experimentally, and some of them certainly deserve to be. The next section reports one such study.

3
PROBABILISTIC INFORMATION PROCESSING SYSTEMS WITH CUMULATIVE AND NONCUMULATIVE DISPLAYS*

In figure 1 the feedback loop from the processed display of the system's current opinions to the likelihood estimator was marked with a question mark, to indicate uncertainty about whether or not such feedback is a good idea. The argument in its favor is simply that operators are curious about the meaning of their responses. With a real system, it might be difficult to prevent their obtaining access to information about the system's current opinions. The argument against such feedback is that it might turn the PIP task into a POP task. That is, the operator might change his likelihood ratios until he got the effect on the system's opinion that he wanted, rather than simply estimating likelihood ratios. The purpose of this experiment was to compare a noncumulative PIP like that studied in the preceding section with a cumulative PIP in which the system's current posterior opinions were displayed to the subject.

3.1. METHOD

3.1.1. TASK AND DISPLAY. The noncumulative PIP was in all respects like that described in the previous section. The response mechanism, instructions, and training for the cumulative

*Research reported in this chapter was conducted under Contracts AF 19(628)-2823 and AF 19(604)-7393.

PIP were also like those described. Only the display that appeared on the CRT was different. Before trial 1, the display showed a posterior probability of 0.10 for each of the war hypotheses and 0.50 for the peace hypothesis. When the subject moved the levers, the display changed dynamically, as for the noncumulative group. When the subject was satisfied with his estimates, he pressed the button, and then reset the levers to 1:1. After that, the display reappeared, with the posterior distribution implied by the combination of his estimates with the prior odds. This in turn changed dynamically with new estimates of likelihood ratios. Thus the display at any moment combined the probabilities prior to the current lever settings with these settings.

3.1.2. SUBJECTS. Eleven of the original 34 subjects used in the earlier experiment were used in this one; three PIP, four POP, three PEP, and one PUP. They were assigned to the new groups equally, so far as possible; six went into the cumulative and five into the noncumulative group. They were all re-instructed in the characteristics of the information-processing system they were to operate; this involved some relearning for those who had not originally been in the PIP group. They worked through the 43-item training scenario again.

3.1.3. SCENARIOS. The nine most diagnostic of the original 18 scenarios were used in this experiment. Two of the nine were left unchanged. The remaining seven scenarios were altered. Each one was to favor a different hypothesis. Thus some peaceful items and some nondiagnostic items (no more than ten) were removed from the original scenario, and items that favored the particular hypothesis were inserted. These items varied in diagnosticity and either were obtained from the original set or were generated for particular scenarios.

3.2. RESULTS

The basic finding of this experiment is given by the scatterplot in figure 4. As usual, the points are geometric mean final odds for each scenario. The correlation coefficient is 0.907, indicating high qualitative agreement between the two groups. The regression coefficient is 0.647, indicating that the noncumulative display group does indeed respond more sensitively than does the cumulative display group. For comparison, the PIP-POP regression coefficient from the previous experiment was 0.422, indicating that the PIP with cumulative display is intermediate in sensitivity between the optimal PIP without cumulative display and POP. This does suggest that the subjects are to some extent using their levers to set the display to the values they desire, rather than directly estimating quantities on the levers and letting the display fall where it will.

As in the previous experiment, the same conclusion obtained by looking at final odds can be reached also by looking at individual likelihood ratios. The intercorrelations in mean log likelihood ratio for each datum and pair of hypotheses between the two groups was 0.80. The distri-

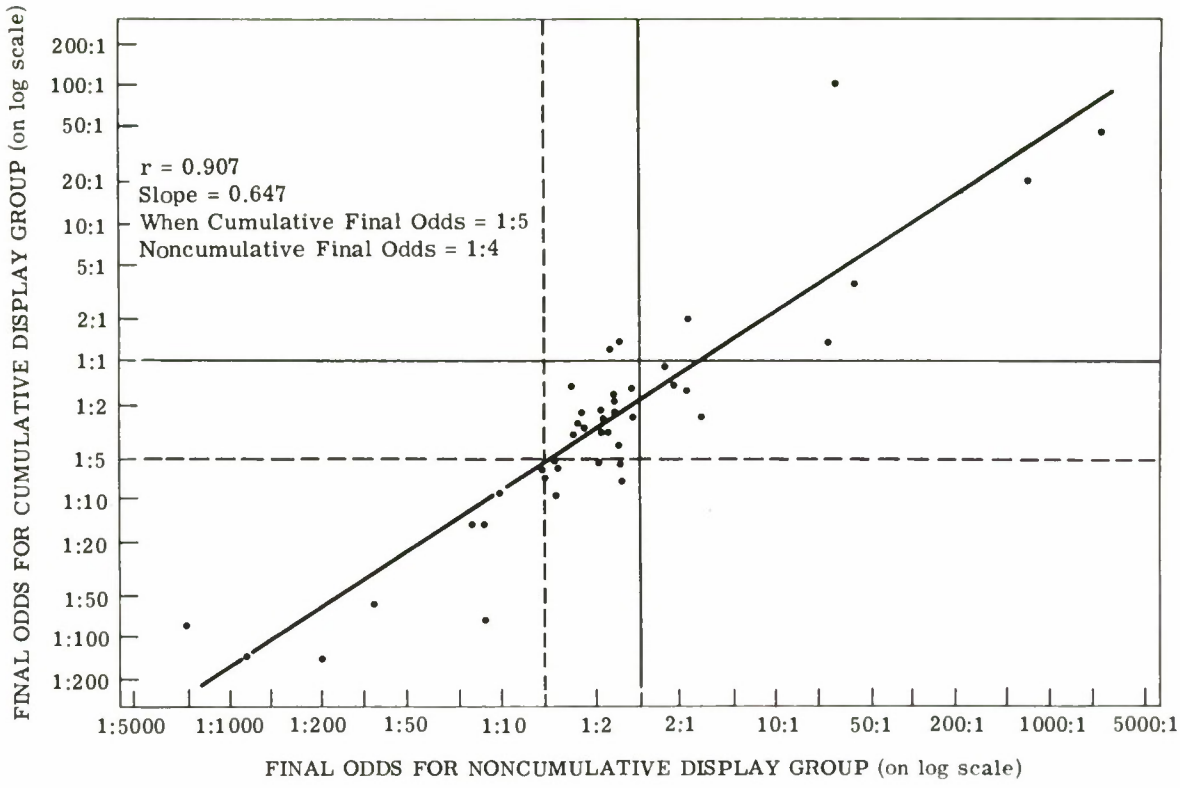


FIGURE 4. FINAL ODDS, CUMULATIVE DISPLAY GROUP vs NONCUMULATIVE DISPLAY GROUP

bution of likelihood ratios is given in table VI and looks like the corresponding distribution in the preceding study. It is apparent that the main reason the cumulative group is less effective than the noncumulative group is the smaller number of likelihood ratios of modest size. This fact highlights the point that diagnosis in these experiments is entirely a matter of putting together items that individually are fairly meaningless.

TABLE VI. GEOMETRIC MEANS OF DISTRIBUTION OF LIKELIHOOD RATIOS

	1.1:1 through 1:1.1	1.1:1 through 1.4:1 and 1:1.1 through 1:1.4	1.4:1 through 2.0:1 and 1:1.4 through 1:2.0	2.0:1 through 2.5:1 and 1:2.0 through 1:2.5	Over 2.5:1 Under 1:2.5
Cumulative Display	2462.0	150.2	52.5	12.7	22.6
Noncumulative Display	2393.2	204.6	73.8	14.0	14.4

It is instructive to examine these distributions subject by subject, as they are displayed in table VII. It is also interesting to examine the correlations between each subject's performance on the training scenario during the first experiment, and his performance, using his new response mode, during the second experiment. With two exceptions, those correlations are satisfyingly high. But Subject Two in the cumulative group had a correlation between first and second run-throughs of the training scenario of only 0.119. And he is responsible for by far the largest

TABLE VII. INDIVIDUAL LIKELIHOOD RATIO DISTRIBUTIONS

Subject	Group in 1st Experiment	Group in 2nd Experiment	Practice Scenario Correlation Coefficient	1.1:1 to 1:1.1	1.1:1 to 1.4:1 and 1:1.1 to 1:1.4	1.4:1 to 2.0:1 and 1:1.4 to 1:2.0	2.0:1 to 2.5:1 and 1:2.0 to 1:2.5	Over 2.5:1 Under 1:2.5
				DISTRIBUTION OF LIKELIHOOD RATIOS				
1	PIP	Cumulative	0.919	2502.0	141	42	10	5
2	PIP	Cumulative	0.119	23.6	123	116	32	113
3	POP	Cumulative	0.994	2660.0	38	2	0	0
4	POP	Cumulative	0.445	2515.0	157	28	0	0
5	POP	Cumulative	0.802	2541.0	137	16	4	2
6	PEP	Cumulative	0.949	2238.0	305	111	30	16
7	PIP	Noncumulative	0.895	2640.0	42	6	0	12
8	POP	Noncumulative	0.999	2230.0	341	98	17	14
9	PEP	Noncumulative	0.827	2457.0	143	73	17	10
10	PEP	Noncumulative	0.977	2244.0	246	141	33	36
11	PUP	Noncumulative	0.810	2395.0	251	51	3	0

number of large likelihood ratios in the cumulative group. Thus this erratic subject worked against the conclusion that the cumulative group extracted less certainty from the data than did the noncumulative group; exclusion of his data from the group's would have led to a larger disparity in favor of the noncumulative group than was in fact found.

3.3. DISCUSSION

It seems clear that the noncumulative display is preferable to the cumulative one. Subjects who have access to the system's current opinions take those opinions into account in estimating likelihood ratios. They thus turn the PIP task into something like a POP task, to its detriment. The moral for designing systems is obvious, though it may prove hard to enforce in real PIP systems.

Retraining of subjects originally trained in other systems to use the PIP response turned out to be easy. Apparently most of the effort devoted to original training was devoted to establishing general familiarity with situation and task. Actual estimation of likelihood ratios is not particularly difficult.

4
CONSERVATISM IN ESTIMATING PROBABILITIES*

This section reports one experiment that is addressed to two issues that arose in the previous section. All of the laboratory studies on inference show that subjects are conservative; they fail to extract from the data all the certainty that is theoretically available. The experiment reported here attempts to determine whether conservatism is a failure to combine optimally all the data that are available, or is a failure to understand properly the probabilistic relationship between the hypothesis and the datum. To make these distinctions clearer, recall the odds-likelihood ratio form of Bayes's theorem:

*This section was a doctoral dissertation by Lawrence D. Phillips.

$$\Omega_1 = \Omega_{0_j} \Pi L_j \tag{8}$$

This form of Bayes's theorem suggests that the process of inductive inference can be broken down into two major components, deductive inference and data combination. The deductive component is the process of determining likelihood ratios. Usually the ideal person assumes that some mathematical model is descriptive of the process that generated the data, as, for example, one would assume that a binomial probability law describes the outcomes of successive flips of a coin. The second component is the combining of individual likelihood ratios with themselves and with the prior odds. Strictly speaking, determining the prior odds should be included as a third component of inference. However, in nonlaboratory instances of inductive inference, prior odds are usually the posterior odds of some previous inference, and so need not be given the status of a separate component. Thus, in order to understand the process of inductive inference, it is necessary to understand both the deductive process that leads to individual likelihood ratios and the process that combines likelihood ratios and prior odds.

The second issue is that of validation. As was mentioned in section 2, comparison of diagnostic systems that process probabilistic information would be much easier if some separate criterion of truth were available. No such criterion was available for the experiments reported in sections 2 and 3, so alternative systems were compared with one another. This approach has been taken by a number of experimenters; their experiments are reviewed here briefly.

The first evaluation of a Bayesian man-machine system for making inferences was reported by Kaplan and Newman (1963). They told their subjects that enemy bombs are being aimed at exactly one of three possible targets. Some bomb impacts are shown as data, and subjects have to decide which target is the intended one. In the "PIP" condition, subjects estimated for each datum three values of $P(D|H)$, one for each hypothesis: values of $P(H|D)$ are then calculated by applying Bayes's theorem. In the "non-PIP" condition, subjects estimated $P(H|D)$. Comparing the estimated values of $P(H|D)$ with the calculated values showed PIP superior to non-PIP in that it achieved higher posterior probabilities for the correct hypothesis more quickly. Also, performance of non-PIP subjects deteriorated as the sequences of data became more ambiguous. No differences in performance were seen among the PIP group as a function of task difficulty.

The next study by this group of experimenters (Kaplan, Lichtenstein and Newman, 1963) complicated the basic bombing model to the point where subjects apparently were forced to adopt a much simpler model than the experimenters had intended, so that the effects of the independent variables became obscured. PIP performance was not superior to, and in fact, was slightly worse than non-PIP performance.

In the meantime, the Laboratory of Aviation Psychology at Ohio State University was conducting some Bayesian experiments. The first report of some of their work (Southard, Schum

and Briggs, 1964a) did not investigate human ability to estimate likelihoods but did give some support to Edwards' contention (1962a) that computer-generated posterior probabilities are superior to subject-estimated posterior probabilities when the probabilistic relationship between datum and hypothesis is specifiable in terms of a mathematical model.

In the next three studies (Schum, Goldstein and Southard, 1966), conducted at Ohio State University, posterior odds calculated from Bayes's theorem using the subjects' estimates as a source of $P(D|H)$ were compared to the subjects' estimates of $P(H|D)$. Here again, calculated posterior odds were higher for the correct hypothesis than were the estimated posterior odds.

In another study, Fox and Hackett (1964) examined the question of conditional independence of data. They presented subjects with two sequences of data; in one sequence the items were identical to the items in the other sequence, but occurred in reverse order. When the investigators found subjects' estimates differing for identical items of data even after one sequence had been mathematically reversed to make it comparable to the other sequence, they rejected the hypothesis that subjects make conditionally independent estimates. Unfortunately, the authors did not realize that they were placing their subjects in a task that presented a null hypothesis and a diffuse alternative hypothesis. As Edwards, Lindman and Savage (1963) have shown, data cannot be conditionally independent when point and diffuse hypotheses are considered simultaneously. Thus the behavior of Fox and Hackett's subjects is not inconsistent with the Bayesian model.

In a recent study by Schneider (1965) a Bayesian approach to medical diagnosis is tested. Nurses are first given items of data (signs, symptoms, and other information) and asked to make likelihood estimates for pairs of hypotheses that concern the possible post-operative states of a patient. After the subject observes a datum and makes likelihood estimates, she makes posterior probability estimates for the two hypotheses being considered. The data show that if the estimated likelihood ratios are multiplied according to equation 6, the resulting posterior odds are considerably greater than the posterior odds calculated from the estimated posterior probabilities.

In the first task, the nurses make only deductive inferences—they estimate likelihood ratios. In the second task, the subjects perform a deduction and then combine the resulting likelihood ratios in order to complete the inductive process. In other words, only one component of inference is involved in the first task, but both components are involved in the second. If it is assumed that the deductive inferences are the same in both, then the discrepancy in the posterior odds for the two tasks must be related to the added requirement for combining data in the second task. Imposing this additional requirement on subjects may cause them to extract less certainty from the data than they would if they only had to estimate likelihood ratios.

This finding supports Edwards' basic contention that in a Bayesian diagnostic system men should serve as transducers for likelihoods or likelihood ratios, leaving to the computer the

task of combining these likelihoods according to Bayes's theorem. However, in most studies of systems it is not possible to assess the amount of conservatism shown by subjects because an adequate criterion for validation is absent. Recall that conservatism is defined as the reluctance of subjects to revise posterior probabilities as much as is prescribed by Bayes's theorem. In other words, Bayesian revision is the criterion by which conservatism is judged. If to calculate the amount of Bayesian revision the experimenter assumes an arbitrary model for determining the likelihoods, then the finding of conservatism is only as valid as is the arbitrarily chosen model for $P(D|H)$.

In the experiments conducted by Kaplan and his associates, a plausible but debatable model was chosen for the bombing task. The Kaplan, Lichtenstein, and Newman study (1963) was unsuccessful because subjects chose a different and simpler model than the experimenters adopted. The Ohio State University experiments are based on a paradigm that emphasizes counting processes in arriving at $P(D|H)$, so their work is perhaps not applicable to systems having to deal with unique events for which no relative frequencies are available. In the Fox and Hackett studies (1964), $P(D|H)$ was arrived at by judgments from a panel of "experts." Schneider (1965), however, makes no attempt to find a criterion; he does no more than compare two different inferential processes, so nothing can be said about conservatism except in a relative sense. In fact, all of these studies can make only relative statements about conservatism because conservatism is relative to the validity of the model for $P(D|H)$. In a sense, of course, all models for $P(D|H)$ are of relative validity, but, to use the words of Edwards, Lindman, and Savage (1963), "some models are more public than others." Most people would agree that a binomial probability law describes the outcomes of successive flips of a coin, but any bombing model is open to considerable debate. In the experiment described in the next section, a public model for the likelihoods is used, and while the parameters are known to the experimenter they are not known precisely to the subject. Thus, like the bookbag-and-poker-chip experiments, a standard for validation is available, but, as in the PIP studies, the values of $P(D|H)$ are not displayed to the subjects. This experiment, then, serves to bridge the gap between bookbag-and-poker-chip studies and the studies of systems, for it allows PIP and POP conditions to be compared to optimal or Bayesian performance.

4.1. THE EXPERIMENTAL PROBLEM

Psychological studies of probabilistic inference use models of $P(D|H)$ obtained from models on which mathematicians agree, and the parameters of these models, are displayed to the subjects. No separate measurements are made of the subjects' perceptions of the displayed values of $P(D|H)$ for single items of data. Consequently, it is not possible to determine whether conservatism is caused by failure to combine the data adequately or failure to understand the probabilistic relationships between the data and the hypotheses.

The experiments to be described attempt to cope with these problems. The subjects are given both inductive and deductive inferential tasks. In the inductive task, the subjects are presented with two hypotheses, H_b and H_e. One hypothesis is chosen at random with equal probability, and data are sampled. After each observation, the subject is asked to state his current opinions, in odds, about the truth of each hypothesis. This condition is identical to the POP condition described in section 2. In the deductive task, the subject is asked to give estimates of the likelihood ratios for each of the possible items of data, and these likelihood ratios are combined by using Bayes's theorem to give posterior odds. This condition is identical to the PIP condition in section 2.

The deductive task can be described by a multinomial probability law; on this there should be no disagreement. Thirty different data are possible, so there are thirty parameters of the multinomial law. The data are two-letter combinations (bigrams), and the parameters are based on the subject's use of these bigrams in his writing. The parameters are known precisely to the experimenter because the frequencies with which the bigrams were used was determined by counting. No such counting process is available to the subject so he must rely on his intuition about how frequently he uses particular bigrams. Thus, the parameters of the multinomial process are not known precisely to the subject. Each subject is asked to perform the task so that the experimenter can assess the degree to which the subject's understanding of the parameters of the multinomial process agrees with the veridical values. ("Veridical" is used in this report in the sense of "objective" or "criterion.")

The effects on conservatism of the two components of inference are determined in several ways, but the most important is based on comparisons among three sets of posterior odds. In the first case, the posterior odds are calculated from Bayes's theorem, where the values of the likelihood ratios (LRs) are based on the veridical values of the parameters of the multinomial probability law (these veridical likelihood ratios will be abbreviated VLRs). In the second case, Bayes's theorem is again used, but the LRs are based on the subject's estimates of them (these estimated likelihood ratios will be abbreviated ELRs). In the third case, posterior odds will simply be those values estimated by the subjects in the inductive inferential task. Notice that each of these three cases involves a different combination of the two components of inference; these combinations are summarized in table VIII.

The logic of the experiments to follow can be seen by comparing the cases in table VIII. Comparing veridical odds with ELR-based odds enables determining the effects on conservatism of nonveridical understanding of the multinomial probability law (the data-generating, or d-g, process). The comparison between estimated odds and ELR-based odds will allow the effects from combining the data to be assessed. The effects from both components of inductive inference can be determined by comparing estimated odds with veridical odds.

TABLE VIII. REFLECTION IN DIFFERING POSTERIOR ODDS OF OPTIMAL COMBINATION OF DATA AND/OR VERIDICAL DEDUCTIONS REGARDING THE DATA-GENERATING PROCESS

Case	Basis for Posterior Odds	Optimal Combination of Data?	Veridical Understanding of d-g Process?*
I Veridical odds	Bayesian calculation; veridical likelihood ratios	Yes	Yes
II ELR-based odds PIP	Bayesian calculation; estimated likelihood ratios	Yes	No
III Estimated odds POP	Estimates	No	No

*The term d-g process is an abbreviation for data-generating process—in this use, the sampling process described by the multinomial probability law.

4.1.1. METHOD

4.1.1.1. Subjects. Eight men on the editorial staff of the daily student newspaper at The University of Michigan were asked to participate. One declined and one was unable to serve beyond the training session because his workload increased when he was promoted to a more responsible position on the editorial staff. The subjects were chosen solely on the basis of the quantity of their editorial output in the fall term, 1964; only editors who had written at least 4000 words were asked to participate.

4.1.1.2. Procedure and Design. This experiment consists of four parts: training, deductive task, inductive task, and replication of the deductive task.

The purpose of the training session was to familiarize the subjects with the response device used in the inductive task. With this device the subjects could express their uncertainty in odds, the range of possible values extending from 1:1 to 1,000,000:1. Each subject was told that the device was a general-purpose apparatus and not designed solely for this experiment. Consequently, the range might be too great, and so not all of it need be used, or it might be too small, and so could be extended by verbal estimates.

Odds, rather than probabilities, were chosen as the response mode because Phillips and Edwards (1966) found that the subjects' estimates were more nearly veridical when the subjects made verbal estimates of odds or responded on a logarithmically-spaced odds scale than when they responded in probabilities. Although the numbers may not be the same, the form of poste-

rior odds is identical to likelihood ratios, namely, x:1, where $x \geqq 1$. Thus, training in the estimation of posterior odds is also relevant to the estimation of likelihood ratios.

The training task was similar to Phillips and Edwards' experiment III (1966). The subject was told to imagine two bags, each of them containing 100 poker chips, with red chips predominating in one bag and blue chips predominating in the other. He was shown a bag and told that it was just as likely to be the predominantly blue bag as the predominantly red one. To indicate that each of the two bags was equally likely to be the chosen one, the subject was told to set his response device at 1:1.

The subject was told that the predominantly red bag contained the percentage p of red chips and percentage q of blue chips, while the predominantly blue bag contained the inverse percentages, p blue chips and q red ones. The percentage values of p and q were either 85-15, 70-30, or 55-45. Ten chips were then shown one at a time; the subject was told that the sequence of chips was the result of random draws, with replacement, from the chosen bag. After each new chip was shown, the subject stated which bag he thought was the more likely to be the chosen one. He then revised, on the odds apparatus, his previous intuitive estimates of the odds in favor of the stated bag. This process of selecting one bag at random from two and then drawing ten chips from the bag was repeated 15 times, five for each of the three pairs of bag compositions.

After the subject made each estimate, the experimenter told the subject what his payoff would be if the predominantly blue bag were the chosen bag, and what it would be if the predominantly red bag were the chosen bag. The subject then recorded both his odds setting and his two possible payoffs on a data sheet, a separate sheet being used for each sequence.

After each sequence of ten draws, the subject was told which hypothesis was correct. The subject then read the ten payoffs corresponding to the correct hypothesis to the experimenter who determined the total on a calculator. This total was shown to the subject and it was converted to money at the rate of five cents per 100 points. The subject was told the running total of winnings, in cents, after each sequence was completed. Total winnings after the 15 sequences were the only pay the subjects received for the training task. Winnings varied from \$2.63 to \$3.86 for a session that lasted from 1-1/2 to 2 hours.

The use of payoffs follows the suggestion of Phillips and Edwards (1966) and of Edwards (1961b) that payoffs serve not only as motivators, but also as instructions. Payoffs with a logarithmic relationship to $P(H|D)$, and so to $\frac{\Omega}{1 - \Omega}$, were used because Phillips and Edwards found that the subjects' probability estimates were more nearly veridical with fewer overestimations under this payoff scheme than when linear, quadratic, or no payoffs were used.

Specifically, when Ω is set equal to the odds estimated in favor of the correct hypothesis, and $v(\Omega)$ equal to the payoff for the estimate Ω,

$$v(\Omega) = 100.00 + 332.19 \log\left(\frac{\Omega}{1 + \Omega}\right) \tag{9}$$

This is the payoff, in points, the subject received if he named the correct hypothesis. Since $\lim_{\Omega \to \infty} v(\Omega) = 100$, the most the subject could win was 100 points. If the subject was wrong he received payoff $v\left(\frac{1}{\Omega}\right)$, where

$$v\left(\frac{1}{\Omega}\right) = 100.00 + 332.19 \log\left(\frac{\Omega}{1 + \Omega}\right) \tag{10}$$

for which the maximum loss is $-\infty$. As a practical limitation, payoff calculations were rounded off to two decimal places. At odds of about 29,000:1, the winning payoff is precisely 100.00; so several subjects reported they did not feel it was worthwhile to exceed odds estimates of 29,000:1. (The possible loss at this value of odds is -1382.39.)

For this payoff scheme the optimal strategy is for the subject to estimate his subjective odds rather than any others. This strategy is optimal in the sense that it maximizes subjectively expected value (SEV). Specifically, letting ω represent the subjective odds in favor of the correct hypothesis, the SEV function, as given by

$$\text{SEV} = \left(\frac{\omega}{1 + \omega}\right) v(\Omega) + \left(\frac{1}{1 + \omega}\right) v\left(\frac{1}{\Omega}\right) \tag{11}$$

has its maximum at the point where $\Omega = \omega$, i.e., when the estimated odds are equal to subjective odds. Further discussion of this class of payoffs can be found in Toda (1963), van Naerssen (1962), and Phillips and Edwards (1966).

Prior to the first deductive task, the subject was told that all his editorials that appeared on the left side of the editorial page (which had undergone minimal editing) during the fall semester had been typed into a computer. The computer had counted the number of times he began his written words with a particular bigram (a two-letter combination) and the number of times his words ended with the bigram. The computer ignored single-letter words and counted two-letter words as both a beginning and an ending bigram. The following example was given. In the sentence " 'My hero!' he roared," these bigram counts would be obtained:

		No. Times Begun	No. Times Ended
Bigram	MY	1	1
	HE	2	1
	RO	1	1
	ED	0	1

The subject was not given the counts on his editorials. For a selection of 30 bigrams he was asked to specify whether a given bigram was more likely to have occurred in his editorials at

the beginning of words or at the end of words, and then how much more likely (in a ratio, not a difference, sense). In other words, based on his experience with using the language, the subject was asked to make a probabilistic deduction concerning a specific bigram. The subject wrote all responses in this task.

For the inductive task, the subject was told to imagine that all the beginning bigrams appearing in his fall editorials had been placed in one bag, Bag B, and all the ending bigrams of his editorials in another bag, Bag E. The number of bigrams in each bag depended on the frequency counts made previously. Thus, if the computer counted beginning/ending frequencies of 20/40 for the bigram MY, then 20 MY bigrams were placed in Bag B, and 40 in Bag E. Next, the subject was told to imagine that one of the two bags had been chosen by flipping a fair coin. The contents of the bag were mixed and one bigram was drawn at random. It was recorded, returned to the bag and the mixing-and-drawing process was repeated nine more times. This process of selecting one bag at random from two and then drawing ten bigrams from the bag was repeated 40 times.

The subject was shown the results of each draw and asked to tell the experimenter which bag was the more likely to have been the one from which the sample was drawn, and to estimate on the log odds device how much more likely. After each sequence of ten draws, the subject was told which bag was correct. Actually the bags contained the bigrams of a composite subject, who will be explained later, so that the same 40 sequences could be shown to each subject.

After all 40 sequences were completed, the subject was asked to repeat the probabilistic deductive task.

4.1.1.3. Apparatus. Sequences used in the training task were displayed by lighted bulbs mounted on an upright panel (this device is fully described by Phillips and Edwards, 1966). Bag compositions were shown on a separate display.

In the training and inductive inference task, each subject estimated odds by first selecting a scale that contained his odds estimate and then moving a pointer to the desired value on the selected scale. Six 13 1/4-in. scales were available to the subject; they encompassed the range 1:1 to 1,000,000:1 in six $\log_{10}$ cycles. Each cycle was mounted on one side of a six-sided bar. The subject could choose any cycle by rotating a knurled knob attached to the bar.

Payoffs for the training task were displayed on a 6 × 6-cycle graph 20 in. square. A black line indicated winnings in points as a function of estimated odds, and a red line indicated losses. In addition to this display, whose accuracy was limited to two or three significant figures, the experimenter used a table to determine the exact payoff to two decimal places, so that the subject could record it on the data sheet.

4.1.1.4. Stimulus Sequences. The 15 sequences used in the training task were generated by a repeated Bernoulli process but were constrained to have the proper error characteristics. This means that if a perfect Bayesian subject selected the more probable hypothesis as the correct hypothesis after n draws (or flipped a fair coin to choose between equally probable hypotheses), he would be wrong the expected number of times over each block of sequences generated with the same Bernoulli probability.

The 30 bigrams used in the deductive and inductive tasks were selected from the 576 possible bigrams by an informal heuristic procedure. A large number of bigrams was chosen so that the subjects would see many different data and so the correlation analyses based on these data would have a reasonably large n. The particular bigrams used, shown in table IX, occurred in each subject's writing as both beginning and ending bigrams at least ten times. Thus, no bigram was chosen that was used very infrequently by any subject.

TABLE IX. BIGRAMS USED IN THE INFERENCE TASKS AND THE GEOMETRIC MEANS (ACROSS SUBJECTS) OF THEIR VERIDICAL LIKELIHOOD RATIOS.
The direction of the VLRs is beginning to ending

Bigram	VLR	Bigram	VLR	Bigram	VLR
ad	1.21	hr	0.10	of	1.10
al	0.50	ho	1.63	on	0.45
an	2.34	if	0.99	or	0.30
ar	2.04	in	1.52	re	0.95
as	0.61	is	0.62	se	0.72
at	0.46	it	1.15	so	3.97
be	2.13	le	0.46	st	1.14
by	0.94	me	0.61	te	0.30
ch	0.34	ne	1.21	th	12.25
en	0.43	no	4.29	to	1.09

After frequency counts for the seven subjects who started the experiment had been made by the computer, a composite subject was invented whose frequency counts for a given bigram were defined as the sum of the frequencies for the seven subjects. Hypothetical beginning and ending bags were then filled on the basis of the frequencies for the composite subject. Forty sequences of ten draws each were then generated by random draws (with replacement) from the two bags, 20 sequences per bag. Thus, the data-generating process is described by a multinomial probability law, the values of $p_1, p_2, \ldots, p_{30}$ being determined from the frequency counts of the composite subject.

In order to compute posterior odds for each sequence of data, it was first necessary to determine veridical likelihood ratios (VLRs) for each subject. The VLR for a given bigram was obtained by dividing the beginning frequency count by the ending count for each subject. The

geometric mean of the VLRs of the six subjects who completed all experiments is shown for each bigram in table IX.

The VLRs for a given bigram are generally very similar from one subject to the next. This can be seen by examining the intercorrelation matrix shown in table X. Each cell represents the linear correlation of the logarithm of the veridical likelihood ratios (LVLRs) of subject i with those of subject j where $i \neq j$. Log VLRs rather than VLRs were correlated in this and subsequent analyses because the log transformation preserves the symmetry about VLRs of 1:1.

TABLE X. CORRELATIONS OF EACH SUBJECT'S LVLRs WITH ALL OTHER SUBJECT'S LVLRs

Subject	1	2	3	4	5	6
1	--	0.95	0.83	0.91	0.94	0.94
2		--	0.87	0.92	0.96	0.96
3			--	0.84	0.81	0.89
4				--	0.90	0.91
5					--	0.93
6						--

Posterior odds were computed, by applying Bayes's theorem, for each data sequence for each subject. Median posterior odds for each subject were then determined separately for beginning and ending sequences. Because the correlations given in table X are very high, the between-subject differences in median posterior odds are fairly small.

4.1.2. RESULTS FROM THE TRAINING TASK. Performance of the subjects in the training task was analyzed by using the power function model suggested by Phillips and Edwards (1966). Assuming that the sampling is best described by a binomial process, define a success as the drawing of a chip with the same color as the predominant chips in the chosen bag, and a failure as the drawing of a chip of the other color. The variable s - f is defined as the difference between the number of successes and failures in a given sample. Then, let $\tilde{\Omega}_1$ represent the posterior odds estimated by a subject for a given value of s - f, Ω_0 represent the stated prior odds, and L represent the corresponding theoretical likelihood ratio for all the data presented to the subject since the prior odds were stated. The power function model states that

$$\tilde{\Omega}_1 = L^c \Omega_0 \tag{12}$$

where c is a fitted parameter dependent only on the bag composition, and where

$$L = \left(\frac{p}{q}\right)^{s-f} \tag{13}$$

The constant c has been termed an accuracy ratio (Peterson et al., 1965).

Accuracy ratios were determined for each subject and for each bag composition by the following procedure. First, median posterior odds were computed as a function of s - f. Then, a

regression line for predicting log estimated posterior odds from s - f was computed under the constraint that it had to pass through the origin. (Since this analysis was done graphically, it was not convenient to use the least-squares criterion in fitting the line of regression. Instead, the signed deviations of the medians from the regression line add to zero.) Taking logs of equation 12 and rearranging gives

$$c = \frac{\log \tilde{\Omega}_1 - \log \tilde{\Omega}_0}{\log L}$$

or, since $\Omega_0 = 1$ for this experiment,

$$c = \frac{\log \tilde{\Omega}_1}{\log L} \tag{14}$$

Thus, equation 14 can be solved for c by substituting the regression-line prediction, at any value of s - f, for $\log \tilde{\Omega}_1$, and the corresponding theoretical value for log L. Accuracy ratios calculated from equation 14 are a function of both s - f and p, but Phillips and Edwards (1966) found that c is a constant function of s - f.

Accuracy ratios computed from the training data are shown in table XI. Values less than one indicate that the subjects revised their odds less than the amount prescribed by Bayes's theorem; this is conservative performance. Values greater than one result when the subjects revise their odds more than does Bayes's theorem. To facilitate comparison with the Phillips-Edwards data, the performance of their group of 12 subjects who estimated odds on a log-odds device is shown in the right column. The numbers shown are the values of c averaged over s - f and subjects. All of the subjects except one were generally conservative. The estimates from subjects in the present experiment were more veridical than were those from subjects in the Phillips-Edwards experiment.

TABLE XI. VALUES OF C FOR THE TRAINING TASK AND FOR THE DATA OBTAINED BY PHILLIPS AND EDWARDS (1966)

		S1	S2	S3	S4	S5	S6	Mean	Phillips-Edwards
Bag Composition	0.55	0.55	0.49	1.00	1.35	1.68	0.27	0.89	1.42
	0.70	0.79	0.87	0.89	2.27	0.74	0.38	0.99	0.52
	0.85	0.89	0.77	0.63	1.20	0.40	0.40	0.72	0.32

Accuracy ratios were also determined for the inductive task to facilitate comparison with performance in the training task. Median veridical posterior odds were determined for each subject separately for beginning and ending sequences. Then a group's veridical performance was determined by computing the medians, across subjects, of these medians. With the number of draws as the independent variable, a regression line was computed for the beginning sequences and another one for the ending sequences, using the same procedure employed in the training

task. This procedure was repeated for the estimates of posterior odds. Equation 14 was then solved to determine accuracy ratios for beginning and ending sequences. The accuracy ratio for beginning sequences was 0.29, and for ending sequences, 0.08. (Data to be reported later in this paper suggest that the inequality of the accuracy ratios reflects a bias.) No subject's estimates were more extreme than the theoretical values, though this was not true in the training task. It is apparent that the subjects were very conservative in estimating posterior odds.

4.1.3. RESULTS FROM DEDUCTIVE INFERENCE TASKS. In order to determine the extent to which the subjects misunderstood the d-g process, and, more importantly, the nature of this misunderstanding, analyses of linear regression were made on the data from the deductive task. For convenience, the estimated likelihood ratios in the first and second deductive tasks will be designated by ELR_1 and ELR_2, respectively; their logs will be referred to as $LELR_1$ and $LELR_2$. Veridical likelihood ratios are again abbreviated VLR; their logs, LVLR. Linear correlations between $LELR_1$ and LVLR and between $LELR_2$ and LVLR were obtained for each subject and the slopes and intercepts of the regression lines (for predicting estimates from veridical values) were computed. These correlations and regression parameters are shown in table XII and the corresponding scatterplots are shown in figure 5. The mean $LELR_1$ by LVLR correlation is 0.50; the mean $LELR_2$ by LVLR correlation is 0.60. The increase of 0.10 suggests that performance improved slightly, perhaps reflecting the experience the subjects gained in the inductive task. A Bayesian analysis was carried out to estimate the true difference between the correlations obtained in the first deductive task and those in the second.

An independent normal process, with neither mean nor variance known, was assumed to have generated the random variables $d_1, d_2, \ldots d_6$, where $d_1 = z_{i1} - z_{i2}$, with z_{i1} representing the Fisher z-transformation of the $LELR_1 \times$ LVLR correlation coefficient and z_{i2} representing

TABLE XII. COEFFICIENTS AND REGRESSION PARAMETERS OF THE LINEAR CORRELATIONS BETWEEN LOG ESTIMATED LIKELIHOOD RATIOS (DEPENDENT VARIABLE) AND LOG VERIDICAL LIKELIHOOD RATIOS (INDEPENDENT VARIABLE) FOR THE FIRST AND SECOND DEDUCTIVE INFERENCE TASKS

	$LELR_1$	x	LVLR	$LELR_2$	x	LVLR
Subject	r	slope	intercept, in odds, at 1:1	r	slope	intercept, in odds, at 1:1
S1	0.56	0.55	1.20	0.57	0.76	1.32
S2	0.61	1.36	3.68	0.80	1.49	1.08
S3	0.50	0.66	3.04	0.60	0.74	1.72
S4	0.22	0.14	1.21	0.36	0.25	1.00
S5	0.44	0.38	1.38	0.51	0.44	1.18
S6	0.69	0.44	1.02	0.68	0.33	1.00

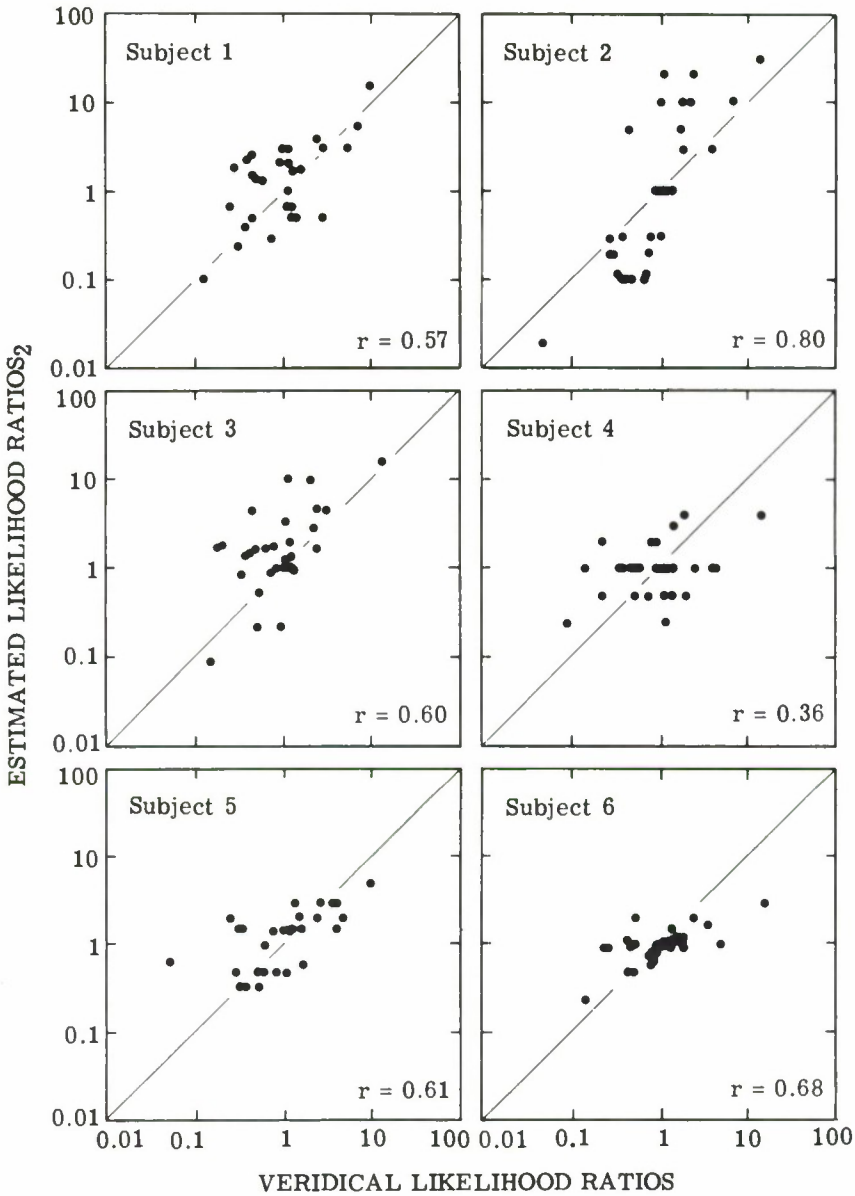


FIGURE 5. SCATTERPLOTS OF ESTIMATED LIKELIHOOD RATIOS FROM THE SECOND DEDUCTION TASK

the same quantity for the second deductive task. For purposes of the Bayesian analysis the process mean, μ, and variance, σ^2, were treated as random variables. The prior joint distribution over the two variables was taken as a normal-gamma distribution because it was reasonably descriptive of the form of the author's prior opinion and because, as a natural conjugate to the independent normal process, it simplifies analysis. Table XIII presents the parameter (m', n', u', v') of the prior distribution, the sufficient statistic (m, n, u, v) of the sample, and the parameter (m'', n'', u'', v'') of the posterior distribution. The parameters m' and m'' are equal to the means of the prior and posterior marginal distributions of μ, m is the sample mean, n is the number of observations in the sample, and u and v are related to the variances and degrees of freedom of their corresponding distributions. (The parameters and statistics are defined in Raiffa and Schlaiffer, 1961, Chapter 11C.)

TABLE XIII. PARAMETERS AND STATISTICS OF THE BAYESIAN ANALYSIS DESCRIBED IN THE TEXT

	Prior	Sample	Posterior
m	0.05	0.155	0.112
n	4.16	6.00	10.16
u	0.52	0.0198	0.221
v	4.00	5.00	10.00

Since Phillips and Edwards (1966) observed a modest amount of learning in their experiments, it was assumed that performance would improve in the present experiment. This would be shown in the higher correlations between estimated and veridical LLRs on the second deductive task than obtained in the first task. However, the learning in probability inference experiments is usually small, so the prior mean of the marginal distribution of $\tilde{\mu}$ was made to be small but positive (specifically, 0.05). The variance assigned was large, indicating uncertainty about $\tilde{\mu}$. The 95 percent prior credible interval centered on 0.05 extends from -0.92 to 1.02. (Correlations of 0.72 and 0 on the first and second deductive task, respectively, will yield a difference in z-coefficients of -0.92. Correlations of 0 and 0.77 will give a z-difference of 1.02.) The mean of the posterior marginal distribution of $\tilde{\mu}$ is 0.11, with a 95 percent posterior credible interval of $-0.33 \leqq \tilde{\mu} \leqq 0.55$.

But what does this analysis say about the hypothesis that performance improves from the first deductive task to the second? The hypothesis "performance improves" can be characterized by values of μ from 0 to ∞ while "performance degrades" is characterized by values of μ from $-\infty$ to 0. Prior odds favor the hypothesis "performance improves" by about 1.23 to 1 (as calculated from the prior marginal distribution of μ), while the posterior odds favor the same hypothesis by about 3.28 to 1. This result moderately favors the hypothesis that performance improves from the first to the second deductive task.

The nature of the subjects' deductions concerning the d-g process can be inferred from the results of the LELR by LVLR regression analyses. Table XII shows the slopes of the linear regression and also the intercept (expressed as an LR, not LLR) at VLR = 1.0. For five of the six subjects, the slope of the regression line is less than one. One interpretation of this is that these subjects interpret the d-g process as being less diagnostic than it really is, an underestimation. One subject does not show the underestimation effect; he overestimates, i.e., he interprets the d-g process as being more diagnostic than it really is.

Another kind of nonoptimal deduction of the d-g process can be inferred from the intercept of the regression line on the ordinate that corresponds to an LR of one. For all subjects this intercept is at a value of LLR for which the LR is greater than one. Since the LR is defined with the beginning bigram count in the numerator, intercept values for which the LR is greater than one indicate a bias in favor of the beginning bookbag.

To summarize, these data show the effects of three subcomponents on the deductive component of inductive inference: an error subcomponent, as measured by the correlation coefficient; an underestimation or overestimation subcomponent, as measured by the slope of the regression line; and a bias subcomponent, as measured by the intercept of the regression line. The subjects show a modest ability (r = 0.60) to estimate the parameters of the d-g process, they usually underestimate the parameters, and they do so with a bias in the beginning direction.

4.1.4. RESULTS FROM THE INDUCTIVE INFERENCE TASK: INFERRED LIKELIHOOD RATIOS. A hypothetical d-g process was inferred from the estimates the subjects gave in the inductive task. Inferred likelihood ratios were computed for each estimate from

$$L = \frac{\tilde{\Omega}_n}{\tilde{\Omega}_{n-1}} \tag{15}$$

where $\tilde{\Omega}_n$ is the posterior odds estimated by a given subject at the nth draw of a given sequence, $\tilde{\Omega}_{n-1}$ is the previous odds estimate and $\tilde{\Omega}_0 = 1$. The 400 values of L were reduced to 30, one for each bigram, by taking the geometric mean of the repeated measures of L on a single bigram. Since equation 15 is simply a version of Bayes's theorem, likelihood ratios inferred from its application are determined under the assumption that the subjects are combining data optimally. The strength of this assumption can be assessed, then, by comparing the mean log inferred LR (LILR)-by-LVLR correlations and the regression parameters with the LELR-by-LVLR correlations and the regression parameters. This can be done by referring to table XIV. Scatterplots of the LILRs by LVLRs are shown in figure 6. The major difference between the LELR by LVLR analyses in table XII and the LILR by LVLR analyses in table XIV is the set of values for the slopes of the regression lines. Underestimation is considerably more marked in the LILR-by-LVLR analyses.

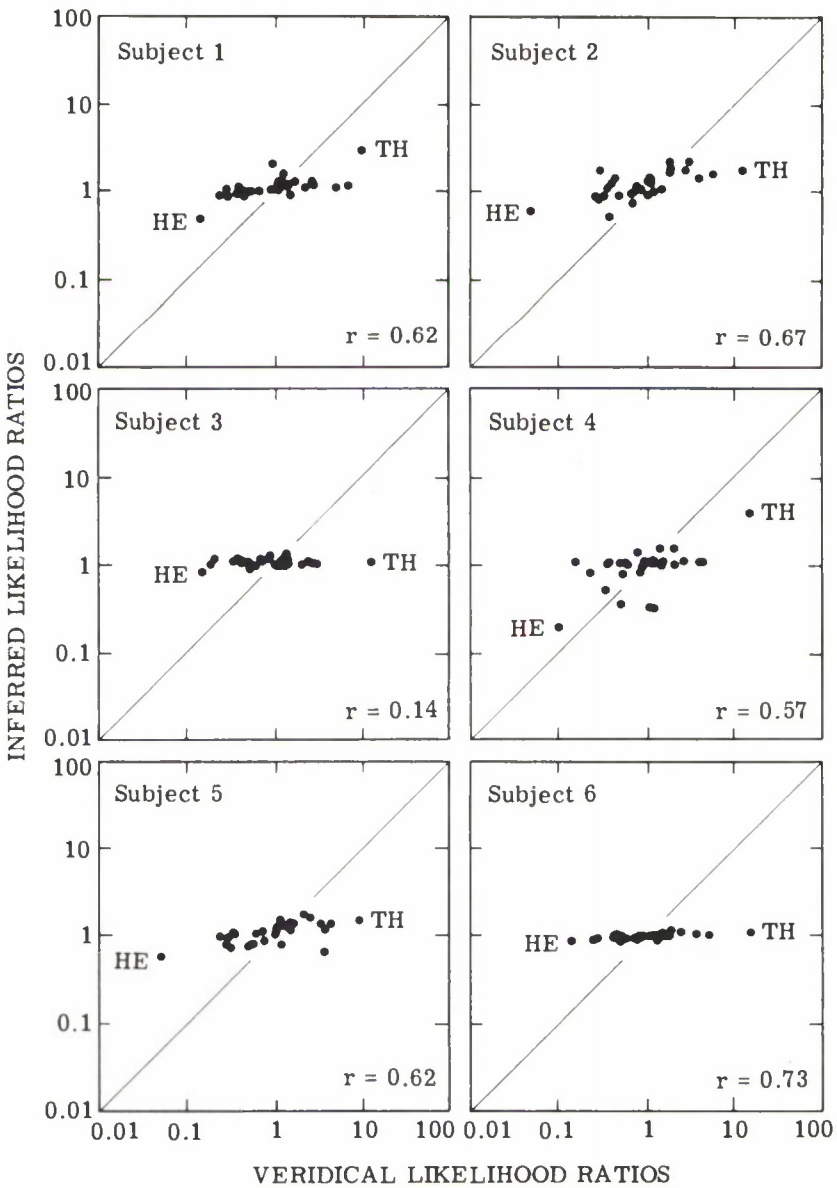


FIGURE 6. SCATTERPLOTS OF INFERRED LIKELIHOOD RATIOS. The ratios are geometric means across repeated measures on the same datum.

TABLE XIV. COEFFICIENTS AND REGRESSION PARAMETERS OF THE CORRELATION BETWEEN LOG INFERRED LIKELIHOOD RATIOS AND LOG VERIDICAL LIKELIHOOD RATIOS

	LILR × LVLR		
Subject	r	slope	intercept, in odds, at 1:1
1	0.62	0.20	1.14
2	0.67	0.22	1.27
3	0.14	0.01	1.06
4	0.57	0.31	0.91
5	0.62	0.16	1.10
6	0.73	0.05	1.00

The difficulties with a linear regression analysis of the LILR data can be seen by examining the scatterplots in figure 6. It is obvious that for Subjects One and Four a straight line is not the best fit to the data. The bigrams TH and HE do the most violence to the straight-line fit.

These data confirm the effect that was predicted by Edwards and Phillips (1964); when the subjects estimate posterior odds and are confronted with many different data, they will tend to ignore all but the most highly discriminative data. Figure 6 shows that the subjects treat most data as being relatively undiagnostic: most ILRs are near one. The word "THE" is used frequently and words beginning with TH are used much more frequently than words ending in "HE." Thus, "TH" appears frequently in beginning sequences, and "HE" frequently in ending sequences. All the subjects reported that they looked for these bigrams as important cues. Figure 6 shows that three subjects tended to assign higher LRs to these bigrams than to any others.

4.1.5. RESULTS FROM THE INDUCTIVE INFERENCE TASK: POSTERIOR ODDS COMPARISONS. Median posterior odds estimated by the subjects are shown separately for beginning and ending sequences in figure 7. These medians are actually medians of medians: first, medians were determined across the 20 beginning sequences and also across the 20 ending sequences for each subject, then the medians of these medians were determined across the six subjects. This same approach was applied to the posterior odds calculated from Bayes's theorem using VLRs; the resulting median posterior odds are shown by the "veridical" plots in figure 7. Medians based on posterior odds calculated from Bayes's theorem using ELRs are shown by the ELR-based plots in figure 7. The group data are representative of individual data for only the estimated and veridical plots; "ELR-based" plots show very great individual differences, some being higher than the veridical plots, and others being lower.

Estimated posterior odds for the beginning sequences are about one-twentieth as large as the veridical odds, while posterior odds for the ending sequences are less than one-thousandth

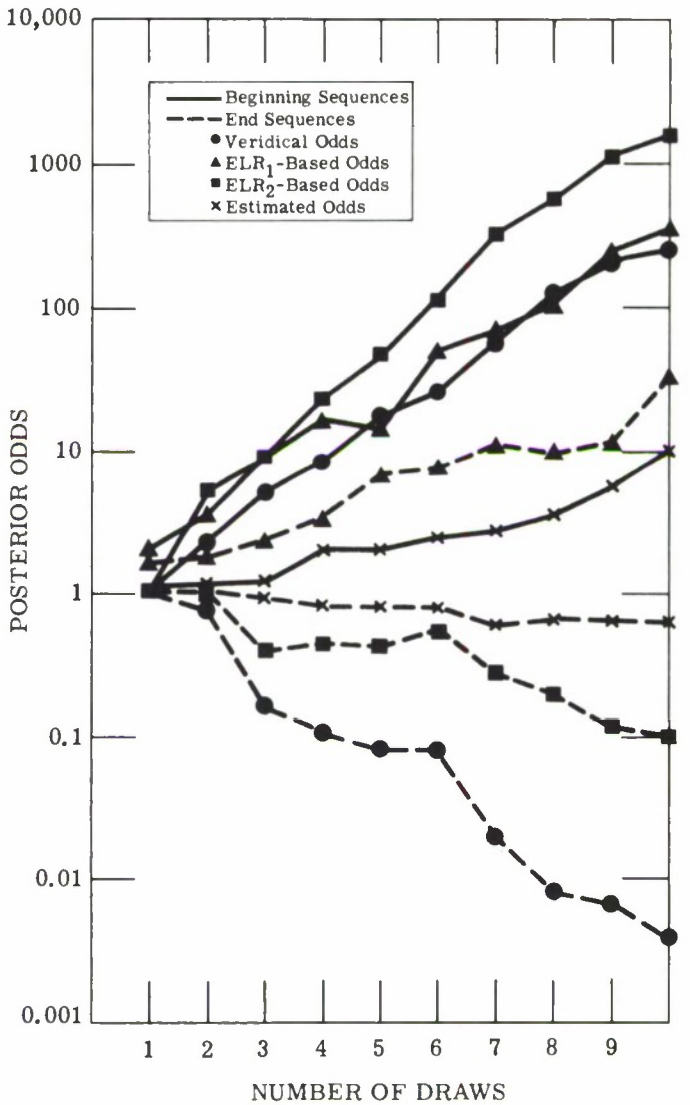


FIGURE 7. POSTERIOR ODDS AS A FUNCTION OF NUMBER OF DRAWS (MEDIANS ACROSS SUBJECTS). Posterior odds are in the beginning-ending direction.

as large as the veridical odds. Comparison of the ELR-based odds with the veridical reveals some strange effects. One plot shows odds that are in the wrong direction, and another plot shows odds that are more extreme than are the veridical odds.

These effects can be attributed to the bias subcomponent of the likelihood-ratio estimates. In order to show the effects of only the other two subcomponents (error and underestimation-overestimation), the ELR-based odds were corrected for bias in the following manner. Each ELR was considered to consist of two multiplicative parts, in symbols:

$$ELR = ELR' \times b \tag{16}$$

where ELR' represents the estimated likelihood ratio uncontaminated by bias and b represents the bias. The value of the bias was set equal to the value of the intercept given in table V. This procedure assumes that for each subject, bias is the same for all ELRs given within one of the deductive tasks, but differs in the two deductive tasks. Substituting equation 16 into equation 8 gives

$$\Omega_1 = \Omega_0 b^n \prod_{j=1}^{j=n} ELR'_j \tag{17}$$

Thus, bias in the posterior odds was corrected for by dividing the posterior odds at the nth draw by the nth power of the intercept. This was done for the median ELR-based odds. The medians, across subjects, of these corrected posterior odds were determined and are plotted in figure 8. Now the ELR odds fall between the veridical and the estimated odds.

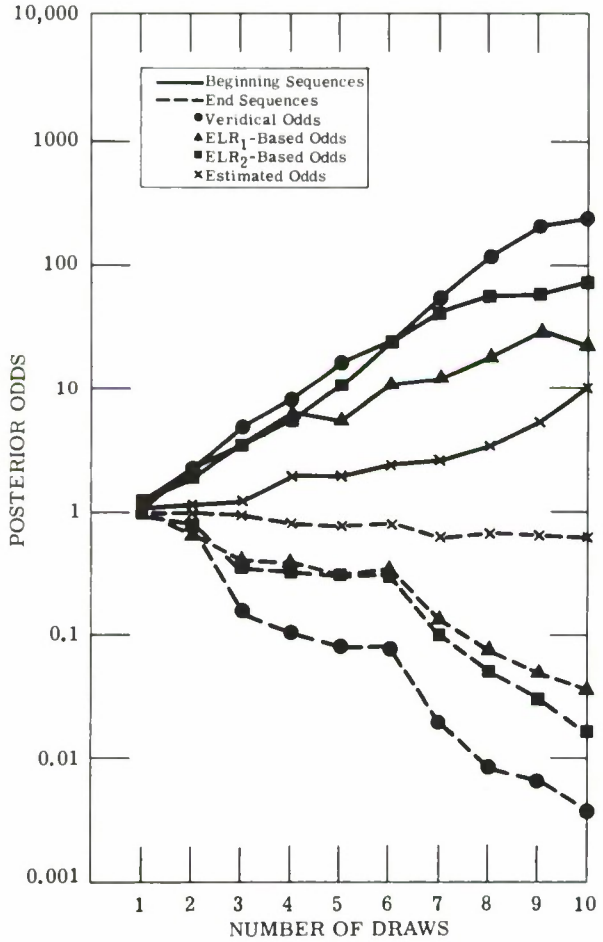


FIGURE 8. POSTERIOR ODDS AS A FUNCTION OF NUMBER OF DRAWS (MEDIANS ACROSS SUBJECTS), WITH ELR_1 AND ELR_2 PLOTS CORRECTED FOR BIAS. Posterior odds are in the beginning-ending direction.

The comparisons suggested by table VIII can now be made using the data in figure 8. Comparing the ELR-based odds with the veridical shows that some conservatism can be ascribed to the error and underestimation subcomponents of the deductive inferences. Comparing the estimated odds with the ELR-based odds shows that some conservatism is attributable to the requirement that the data be combined. In short, conservatism can be found in both the deductive and combinative components of inductive inference.

4.2. DISCUSSION

Because the subjects in this experiment were not randomly chosen from the student population at The University of Michigan, there is a question of the extent to which these results may be generalized. The data from the training task indicate that the performance of these subjects is less conservative than that of the more representative subjects in the Phillips-Edwards (1966) study. However, most of the subjects in the Phillips-Edwards study were conservative, as were the subjects in the present experiment. (The nonconservatism shown by the Phillips-Edwards subjects in the sequences where $p = 0.55$ is typically found when the data are not very discriminative. Most of the veridical likelihood ratios in the present study are greater than the low-discrimination likelihood ratios that lead to nonconservatism.) Thus the conservatism effect is a general one, but the degree of conservatism shown by the subjects is not necessarily representative of the student population. In fact, it is possible that the great amount of conservatism shown in the inductive inference task is an underestimate of the amount of conservatism that would be shown by a more representative sample of students.

The most striking finding of this experiment is the large amount of conservatism shown by all subjects in the inductive inference task. One reason for more conservatism in the inductive task than in the training task may be the greater number of possible data (30 instead of 2). The only experiment (C. R. Peterson, personal communication with the author) that has investigated this variable shows that conservatism increases as the number of possible data increases. Phillips, Hays, and Edwards (1966) also make this suggestion, though their data are not conclusive on the point. The greater vagueness of the d-g process in the inductive task than in the training task may also account for the greater conservatism in the inductive task.

Two components of conservatism emerge from the present study. One, nonoptimal deductions regarding the d-g process, has three subcomponents whose effects can be seen in the LELR-by-LVLR regression analyses: random error, bias, and underestimation-overestimation. The other component, the combining of data, apparently imposes an additional information-processing load on the subjects beyond that required in a deductive task, causing the data to lose some of their discriminative impact for the subjects.

4.3. IMPLICATIONS

These findings have several implications for the controversy in psychology concerning clinical vs. actuarial prediction, for medical diagnosis, for the design of military diagnostic systems — in short, for any attempt to systematize the inferential process. The implications are seen best in the clinical vs. actuarial controversy, for here the battle lines have been drawn most clearly. Prior to 1954, lines were not so clearly drawn; Meehl's book (1954) changed this situation. To him the question was this: Is clinical or actuarial prediction best? He surveyed the literature on the topic and found 20 studies were relevant. In all but one, actuarial predictions were equal to or better than clinical ones. His book was valuable in raising some basic issues that had not been previously well-understood — such questions as what is a datum, how should hypotheses be formulated, and how should data be combined.

Later, Gough (1962) claimed that the locus of the controversy was in how the data were to be combined. He reviewed actuarial methods, clinical approaches, and briefly surveyed the history of the controversy. He concluded that the problem of validating either procedure is a major drawback to resolving the controversy, that neither procedure has done very well, and that the proper use of a clinician's skills could be a supplement or addition to the actuarial methods.

Sawyer (1963) has attempted a further formulation. He suggests that a distinction has to be made between methods of data collection and methods for combining data. He applies this distinction to 40 clinical-statistical studies and finds that mechanical methods are still generally superior. However, he suggests that "the clinician's potential contribution to increasing validity may be not so much as a combiner of data, but rather as a sensitive measuring instrument. "To devise measurement methodology capable of more fully capturing, in an objective fashion, the broad range of subtle behaviors which the clinician perceives, should be a prime goal of those interested in improving prediction." (Italics his.)

The present study suggests that clinicians may be able to act as sensitive measuring instruments when they perform deductive inferential tasks on individual observations, and that when these inferences are combined optimally the resulting inductive inferences may be of better quality than the inductive inferences made by the unaided clinician. Thus, the real issue is not clinical vs. actuarial prediction, but rather the proper roles for the clinician's intuition and the actuary's statistics or mechanical methods when both work together in making inferences, a point that is implicit in Edwards' (1962) suggestions for a man-machine diagnostic system.

One implication for system designers is that training should aim to familiarize estimators of likelihoods with the d-g process. One type of training would be to expose the subject to many, many instances of data generated by a particular d-g process. The efficacy of this approach

has been amply demonstrated in studies conducted at the Laboratory of Aviation Psychology at Ohio State University (Southard, Schum, and Briggs, 1964; Schum, Goldstein, and Southard, 1965). Another type of training might rely on plausibility arguments. For example, Warner, Toronto, Veasey, and Stephenson (1961), in a study on diagnosis of congenital heart disease, estimated values of $P(D|H)$ for the rare diseases, on which adequate statistics were not available, by considering the pathologic physiology of the symptom. In a similar vein, Edwards and his associates, in the study reported in section 2, trained subjects to estimate likelihood ratios by first familiarizing their subjects with the logical structure of the experimental environment and then giving practice data to their group of subjects. Likelihood ratios were estimated for these practice data and the values were discussed by the group. Arguments of plausibility were used by the experimenters to convince the subjects that their estimates should have been greater or lesser.

Let us return, now, to the question of validity. How can one assess the degree of validity in the conclusions reached by one method or the other? This same problem is, of course, also found in scientific inference, and, ultimately, there is only one answer: it is not possible to ascertain with certainty which of a set of possible theories is true. This philosophic answer, however correct, is unsatisfactory, and so various devices, discussed in section 2, have been invented that substitute some kind of reliability for the unattainable goal of validity.

The methodology of the experiments reported here deals with the problem of validation by employing an experimental task in which the experimenter knows which hypothesis is true. Furthermore, the d-g process was completely specified mathematically, and was known precisely to the experimenter but imprecisely to the subject. Imprecise knowledge of the d-g process is characteristic of most real-life situations in which people must make inferences on the basis of inconclusive data. But when the d-g process is known imprecisely to an experimenter (as has been the case for most of the studies cited in the clinical vs. actuarial controversy), then the only way to check on validation is to count the number of times the hypothesis that eventually turns out to be correct (if ever this can be determined) is identified as the true hypothesis. But this approach may lead to erroneous conclusions, as was indicated in section 2. For the inductive inferential task of the experiment, both the d-g process and the correct hypothesis are known to the experimenter; for the deductive inferential task, the d-g process is known to the experimenter, thus allowing an analysis of the components of the inferential process. However, this approach trades one set of problems for another. While it is true that the experimenter knows the parameters of the d-g process precisely, this is only possible because the d-g process was defined as relevant to the fall editorials of the subjects, not to all their written material. It is quite likely that the subjects relied heavily on their knowledge of all their written editorials and news articles, or even on their general use of English in both writing and speaking. If this is true, then the VLRs are really only estimates, though probably good estimates,

of the LRs that characterize each subject's general use of English. Looked at this way, reliability has been substituted for validity in this study, too. The experiment reported here, then, can be considered as another, and successful, attempt at the convergent validation of PIP.

5
CHOICE AMONG BETS AND REVISION OF OPINIONS*

Men usually fail to extract as much information from inconclusive data as is latent in those data; compared with Bayes's theorem, men are conservative. This conservatism is a function of several variables. It depends on response mode; subjects are more conservative when estimating probabilities than when estimating odds. Knowledge of results and appropriate payoff matrices decrease conservatism. Conservatism increases with the diagnostic impact of the data; that is, the larger the likelihood ratio, the greater the conservatism. This is true whether the likelihood ratio is increased by making the populations that might be sampled easier to discriminate, and thus affecting the diagnostic impact of the individual datum, or by increasing the number of data included in a single observation. (For data supporting these generalizations, see Edwards, 1965; Edwards and Phillips, 1964; Peterson and Miller, 1964; Peterson, Schneider, and Miller, 1964; Phillips and Edwards, 1966; Phillips, Hays, and Edwards, 1966.)

The dependence of conservatism on response mode introduces into decision theory a problem familiar to psychophysicists. When several different kinds of responses that ought to be tapping the same internal process disagree, which most faithfully reflects that internal process? In the case of subjective probabilities, the situation is still more complicated because there are two other classes of experiments to consider. A number of psychophysical experiments on the direct estimation of relative frequency have found excellent agreement between actual and estimated relative frequencies (Shuford, 1961; Robinson, 1962). It is not entirely clear what relation, if any, such experiments have to the estimation of probabilities; at any rate, they clearly indicate that circumstances can be found in which men can estimate numbers between zero and one that closely match some normative external criterion. Perhaps most important, however, is the rather large and rather inconsistent body of experiments in which subjective probabilities have been inferred from choices among bets (for reviews of this class of experiment, see Edwards, 1954 and 1961a; for the best recent examples see Tversky, 1965, and Lindman, 1965). These experiments agree unanimously that subjective probability so inferred is not linear with relative frequency (or whatever other external standard seems appropriate to the particular experiment) but disagree about the form of the nonlinearity. None of these experiments, however, has been conducted in a setting to which Bayes's theorem is relevant.

This study compares several response modes in a Bayesian setting, including one that is intermediate between estimating probabilities and choosing among bets. Its original motivation

*This section represents work done by Andries F. Sanders.

was an attempt to find a choice-among-bets response from which probability estimates could be recovered with precision. Such a response was needed for use in the PEP group of the first PIP experiment. It was apparent that to recover probabilities from choices among bets with precision, it would be necessary either to have the subject rank-order a very large group of bets, or else to have him pick one bet from an even larger group. The former was impractical, so various versions of the latter were explored. In order for the subject to be able to pick one from a large group of bets and have the response be meaningful, the bets must be arranged in an orderly way; the natural ordering principle is, of course, the probability that the choice of each bet would imply. But the task of choosing a bet from a list ordered in probability is very similar to the task of estimating a probability; indeed, it can be made identical by appropriate choice of the payoffs and form of the bets.

Toda (1963) and van Naerssen (1962) have pointed out that the most natural payoff scheme for probability estimation, linear payoff, is completely inappropriate. In such a scheme the subject would win $0.75 if he estimated the probability of the event that turned out to be the truth as 0.75, would win $0.90 if his estimate was 0.90, and so on. In such linear schemes, the optimal strategy is to estimate the probability of the most likely event as one, all others as zero. There are, however, nonlinear payoff schemes that make it optimal to give your true opinion as your estimate. The one with the most attractive mathematical properties, the logarithmic payoff scheme (whereby your payoff is proportional to log p), unfortunately implies the possibility of losing near-infinite amounts of money, which makes it impractical in many contexts. For two-hypothesis situations a quadratic payoff scheme (whereby your payoff is proportional to p^2) is more satisfactory. Moreover, it turns out to be possible to develop choice-among-bets responses that look linear with probability but are in fact quadratic. Such a response looks somewhat like a probability estimate, and yet is a choice among bets, and has the formal properties of a quadratic payoff scheme; one such response was studied in this experiment, and a very slightly different version was used in the first PIP study (without the real-gambling feature).

5.1. THE INFERENCE TASK

Subjects were told that there were two urns, one with 60 percent red and 40 percent blue poker chips in it, and the other with 40 percent red and 60 percent blue. (The actual number of chips was not specified, but was implied to be large. Since samples were drawn with replacement, it was formally irrelevant.) The subject was told that one of the urns had been chosen by tossing a fair coin, and a sample had been drawn from it, randomly, with replacement. The subject was told the number of reds and blues in the sample, and then made a response indicating which urn he thought likely to have been chosen. Then the subject was asked to imagine that a

new choice between the urns had been made, was given a new sample from the chosen urn, and made a new response.

5.1.1. RESPONSE MODES. Three response modes were used in the four subexperiments. The simplest is the verbal odds mode. In this mode, the subject simply states verbally the odds in favor of the more probable urn; odds are always taken as a number equal to or greater than one. In the quadratic gain mode, the subject is presented with a table of bets like the one illustrated in table XV, and must choose one. To encourage subjects to take the task seriously, they were informed that their pay would depend on the number of points earned as a result of what bet was chosen and what urn had been used to generate the sample. In the table presented to the subject, there were 100 bets rather than ten, and the column headed "Implied Probability of Red Urn" was absent. The final response mode was a bidding mode, using a pseudolinear payoff scheme. The subject stated how many points he would bid for a bet that paid 100 points if the urn sampled was the red urn, and zero points if it was the blue one. They were told that their payoff would be determined as follows. The experimenter would draw a random number between zero and 100. If the number drawn was higher than the bid, the subject earned a number of points equal to the number drawn. If the number drawn was not higher than the bid, the subject earned 100 points if the red urn had generated the sample, and zero points otherwise. It can be shown that this is formally equivalent to a quadratic payoff scheme, and that the subject's best strategy is to bid a number of points equal to his probability that the red urn was chosen.

TABLE XV. QUADRATIC GAIN BETS

Red Urn	Blue Urn	Implied Probability of Red Urn
100	0	1.0
99	19	0.9
96	36	0.8
91	51	0.7
84	64	0.6
75	75	0.5
64	84	0.4
51	91	0.3
36	96	0.2
19	99	0.1
0	100	0.0

5.1.2. SUBJECTS AND EXPERIMENTAL CONDITIONS. All subjects were male University of Michigan undergraduates. There were four subexperiments. In subexperiment 1, seven subjects used the quadratic gain and bidding modes. There were two blocks of 40 trials each; stimuli in the second block were identical with stimuli in the first except that the colors were reversed. Four of the subjects used the quadratic gain mode in the first block and bidding in the second; the other three reversed this order. The stimuli in the first block are shown in table XVI.

TABLE XVI. SAMPLES IN FIRST BLOCK OF SUBEXPERIMENT ONE

Blue	Red	Blue	Red	Blue	Red	Blue	Red
13	3	12	11	6	10	16	12
6	10	10	2	6	8	11	9
2	9	7	5	18	12	11	10
13	7	10	10	12	12	10	1
7	9	11	12	13	16	4	11
6	11	9	6	12	11	5	9
12	7	17	11	11	13	13	14
3	11	10	3	14	11	7	12
12	10	14	15	10	15	10	11
4	7	11	13	6	6	10	8

In subexperiment 2, quadratic gain was compared with verbal odds. Of the six objects, three used quadratic gain first and then verbal odds; the other three reversed this order. Table XVII shows the 30 samples in the first block. An attempt was made to include more relatively unbalanced (highly diagnostic) samples than had been used in subexperiment 1.

TABLE XVII. SAMPLES IN FIRST BLOCK OF SUBEXPERIMENT TWO

Blue	Red	Blue	Red	Blue	Red
6	10	7	5	12	11
2	9	14	5	16	8
13	7	17	11	14	11
7	9	3	10	10	15
6	11	6	15	16	12
12	7	7	9	10	1
3	11	6	10	4	11
4	7	6	8	5	9
12	11	18	12	7	12
10	2	13	16	12	10

In subexperiment 3, the response modes were again quadratic gain and verbal odds. However, in this experiment the samples were not summarized. Instead, they were presented one poker chip at a time to the subjects. Of the seven subjects, four saw six sequences of 20 chips each using quadratic gain, and then three more using verbal odds. The remaining three subjects got the two conditions in reverse order.

In subexperiment 4, the urns were 70-30 and 30-70 instead of 60-40 and 40-60 (this changes only the instructions to the subjects, since no actual urns were used). Of the eight subjects, four saw five sequences of 20 chips each using the quadratic gain mode, and then saw them again with reversed colors using the verbal odds mode; the other four subjects used the same response modes in reverse order.

5.2. RESULTS

In symmetric binomial Bayesian tasks like this one, the number of chips in a sample is not diagnostically relevant; the only relevant aspect of the sample is the difference between the number of red and the number of blue chips in it. This is the appropriate independent variable against which to plot the data. Moreover, if there are no response biases in favor of red or blue (and there were none), it does not matter whether the data favor the red or the blue urn. Consequently the independent variables will be called successes minus failures, or s - f, where a success is a chip of the color predominant in the sample.

Bayes's theorem can be written

$$\log L = \log \Omega_1 - \log \Omega_0 \tag{18}$$

where Ω_0 is the prior odds (in this case, one), Ω_1 is the posterior odds, and L is the Bayesian likelihood ratio (for this symmetric binomial task, the logarithm of the likelihood ratio is (R - B) log (p/q), where R and B are the number of red and blue chips in the sample and p is the probability of drawing a red chip from the predominantly red urn).

Peterson, Schneider, and Miller (1965), Phillips and Edwards (1966), and others have found that if the correct prior odds and the posterior odds calculated from a subject's estimates are used to infer a log likelihood ratio according to equation 18, that a log likelihood ratio typically will have an approximately constant ratio to the Bayesian log likelihood ratio. The constant by which the Bayesian log likelihood ratio should be multiplied to obtain the subject's inferred log likelihood ratio has been called the accuracy ratio. It is a useful dependent variable for sequential Bayesian experiments. Typical accuracy ratios in previous experiments have ranged from 0.2 to 0.6, for 70-30 urns. The accuracy ratio is the dependent variable used in data analysis in this experiment.

Mean accuracy ratios for all values of s - f used in subexperiment 1 are presented in table XVIII, along with 1 percent confidence intervals. It is apparent that performance was considerably more Bayesian for the bidding response mode than for the quadratic gain mode. The variability appears to be somewhat larger for the bidding mode, but this is probably a byproduct of the fact that the transformation from bid to log odds expands the upper end of the probability scale far more than the region near 0.5, and thus apparently larger variability is bound to be associated with a larger mean. (Note: separate log posterior odds were calculated for each observation and then averaged; it is very important not to average probabilities, since a mean probability is usually uninterpretable.)

Table XVIII also shows the results of subexperiment 2, comparing quadratic gain with verbal odds. There are no major differences between quadratic gain for the two subexperiments, or between quadratic gain and verbal odds in subexperiment 2, until an s - f of 7 is reached. Then

TABLE XVIII. ACCURACY RATIOS AS A FUNCTION OF s - f

	Subexperiment One		Subexperiment Two	
s - f	Quadratic gain	Bidding	Quadratic gain	Verbal odds
1	0.39 ± 0.26	0.67 ± 0.33	0.25 ± 0.19	0.30 ± 0.20
2	0.53 ± 0.17	0.63 ± 0.31	0.37 ± 0.27	0.41 ± 0.23
3	0.62 ± 0.20	0.67 ± 0.27	0.45 ± 0.26	0.62 ± 0.22
4	0.59 ± 0.19	0.72 ± 0.31	0.53 ± 0.22	0.63 ± 0.19
5	0.60 ± 0.14	0.70 ± 0.29	0.65 ± 0.18	0.72 ± 0.23
6	0.57 ± 0.17	0.69 ± 0.33	0.56 ± 0.17	0.56 ± 0.18
7	0.91 ± 0.34	1.12 ± 0.45	0.88 ± 0.26	0.67 ± 0.21
8	0.98 ± 0.35	1.00 ± 0.32	0.86 ± 0.24	0.62 ± 0.15
9	1.10 ± 0.28	1.16 ± 0.17	0.92 ± 0.21	0.62 ± 0.20

all other response modes rise to very Bayesian levels of performance, while verbal odds stays down in the 0.6 region.

The results of subexperiments 3 and 4 are not worth presenting in detail. They show good agreement between verbal odds and quadratic gain for the sequential procedure at all levels of s - f. They also show remarkably good agreement between the results of the verbal odds mode in this experiment and the results of the same mode in the Phillips and Edwards (1966) experiment. They thus establish comparability between this and previous experiments. A number of other data analyses concerned with the order of presenting response modes and similar issues produced no interesting results.

5.3. DISCUSSION

All response modes and all groups were conservative, as is usual in such experiments. There seemed to be little important difference between verbal odds and quadratic gain as response modes, except for the anomalous behavior of the quadratic gain groups at s - f of seven and higher. It remains to be seen whether that finding is reliable; nothing comparable occurred in subexperiments 3 and 4. The bidding method produced considerably more nearly Bayesian performance than any other studied; it is clearly the method of choice when a choice-among-bets response mode is needed. However, the bidding method is complex and hard for subjects to understand. Had this experiment been longer, so that subjects could have the chance to become more experienced with it, it seems entirely possible that they might have become more conservative, not less, in using it.

These data offer no support whatever for the idea that choices among bets lead to probability estimates different from those obtained by direct estimation. However, these choices among bets are of rather special form. The possibility remains that in other choice-among-bets situations, differences not found here might emerge.

On the basis of these data, the bidding method (or rather, a variant of it) was adopted for the PEP group of the first PIP experiment.

6
ESTIMATION OF PROBABILITIES IN MILITARY AND ABSTRACT SETTINGS*

Phillips, Hays, and Edwards (1966) have found in a complex task requiring evaluation of simulated threats that subjects are unable to extract as much certainty from data as is implied by Bayes's theorem. Indeed, no amount of evidence seemed to be able to induce their subjects to estimate high posterior probabilities.

The aims of the present study were various. One major interest was to see if subjects could be induced to give a high-probability assessment to one hypothesis by presenting sufficient data to indicate overwhelmingly that this hypothesis was true. A second was to compare a task with a military setting comparable to the air-defense threat-evaluation task, with the same problem in an abstract setting.

The remaining aims were derived from consideration of the way a subject should behave if he were a perfect Bayesian operator. Two features of the perfect Bayesian that were selected for examination were the irrelevance of order, and the ability to combine data from different sources. Irrelevance of order means that the total influence of a number of data should be unaffected by the order in which these data were received. A subject may weigh early data more heavily than recent or vice versa. No prior expectations were held except that a subject may well differ from the perfectly Bayesian. The combination of data from different sources is a most useful feature of the Bayesian process and is clearly demanded in many practical situations. It was studied for this reason. The basic strategy was to compare the efficiency of subjects when handling data from two sources with their efficiency in handling data from one. To make this comparison as general as possible the two sources were made to behave in one case as though their indications were complementary while in another case they appeared to be discordant. The detailed manner in which these different conditions were achieved is described below.

6.1. METHOD

6.1.1. MATERIAL. A conditional probability distribution relating each of five possible events (e1-e5) to four possible states of the world (S1-S4) was constructed (see table XIX). This distribution has certain special properties: it neatly fits the military cover story; and if S1 and S3 are considered, it is possible to fractionate the conditional probabilities so that evidence from two sources can appear to have either similar or dissimilar implications (see table XX).

6.1.2. DESIGN. All subjects received 60 items of evidence. These came from six blocks of ten items. The events represented in each block corresponded exactly in their relative frequency to the conditional probability distribution of either S1 or S3. The sequence within an S1 block was always: e4, e1, e3, e1, e1, e3, e2, e5, e1, e4; and that for S3 was always: e1, e1,

*This section represents work done by Harold Dale.

TABLE XIX. THE CONDITIONAL PROBABILITIES AND CODING OF EVENTS

(a) Conditional Probabilities

Possible states of the world and corresponding hypotheses	Possible Events e_1	e_2	e_3	e_4	e_5
S1 (H1) (Frontal Attack)	0.40	0.10	0.20	0.20	0.10
S2 (H2) (Flank Attack)	0.10	0.40	0.10	0.10	0.30
S3 (H3) (Parachute Attack)	0.20	0.10	0.10	0.40	0.20
S4 (H4) (Pincer Movement)	0.30	0.30	0.10	0.10	0.20

(b) Coding of Events

		Possible Events e_1	e_2	e_3	e_4	e_5
Military Task	Agent reports he has sighted . . .	Tank	Armored Personnel Carrier	Light Artillery	Aircraft	Heavy Artillery
	Radio activity detected in area . . .	B	D	A	E	C
Abstract Task	Color of matchstick	Red	Green	Black	Yellow	Blue
	Number of stripes on matchstick	2	4	1	5	3

e4, e5, e4, e4, e5, e3, e4, e2. The blocks were in order A (Q = 1) S1, S3, S1, S1, S1, S1; or B (Q = 2) S3, S1, S3, S3, S3, S3, so that for Q1 at the end of the run the evidence indicated that S1 was true, whereas for Q2, S3 was true.

The items were attributed to either or both of two sources, referred to as 01 and 02. When both were employed, the items were allocated either symmetrically (so that the evidence from both sources appeared to be similar) or asymmetrically (so that evidence from both sources conflicted). Since with the asymmetrical split more items were attributed to one source than the other it was desirable to counterbalance and remove any bias this inequality might introduce. This led to a basic design employing 16 subjects (each subject being given only one problem):

Order of blocks	A (Q1)						B (Q2)					
Sources	01	02	Both				01	02	Both			
Split	-	-	Sym.		Asym.		-	-	Sym.		Asym.	
Counterbalancing	-	-	c	c^1	c	c^1	-	-	c	c^1	c	c^1
No. of Subjects	2	2	1	1	1	1	2	2	1	1	1	1

The task was given either as a military or an abstract setting. Two sets of 16 subjects were given each.

TABLE XX. SPLITTING THE CONDITIONAL PROBABILITY DISTRIBUTIONS FOR S1 AND S3

		Number of Events in each Block of Ten Events			
		Asymmetrical Split (Evidence is conflicting)		Symmetrical Split (Evidence is consistent)	
State	Event	Resembles S1	Resembles S3	Resembles S1	Resembles S1
S1	e_1	3	1	2	2
	e_2	1	0	1	0
	e_3	2	0	1	1
	e_4	0	2	1	1
	e_5	0	1	1	0
		Resembles S1	Resembles S3	Resembles S3	Resembles S3
S3	e_1	2	0	1	1
	e_2	0	1	0	1
	e_3	1	0	0	1
	e_4	1	3	2	2
	e_5	0	2	1	1

6.1.3. APPARATUS. For the military setting, a 2 ft × 3 ft map of the enemy territory within the battle area was provided (fig. 9). It was covered with plexiglass and clamped to a mill board. Large charts, in the form of histograms, showed the relative probabilities of the different possible reports coming from the two sensors. A response board was provided on which the subjects displayed their assessment of the probability of each form of attack. This 3-1/4 in. × 19 in. × 1 in. board had four 14 in. pegs (1/2 in. in diameter), one corresponding to each state. One hundred 1-1/2 in. square (by 1-1/8 in.) anodized aluminum washers were provided and the subject had to distribute these on the four pegs so that the height of each pile corresponded to the probability of the associated state. A back board had a scale marked on it so that the number of washers on each peg could be read off readily.

A pad of paper, lead pencil, and colored grease pencils were provided so that the subject could make notes or calculations and keep records. The grease pencils could be used to mark the map.

Printed cards, 1-3/4 in. × 2 in., were used to provide evidence. Those from the Agent were white and carried messages of the form "AGENT REPORTS LIGHT ARTILLERY." Those indicating detections of radio activity were blue and were of the form "RADIO ACTIVITY REPORTED IN AREA A." The general setup is shown in figure 10.

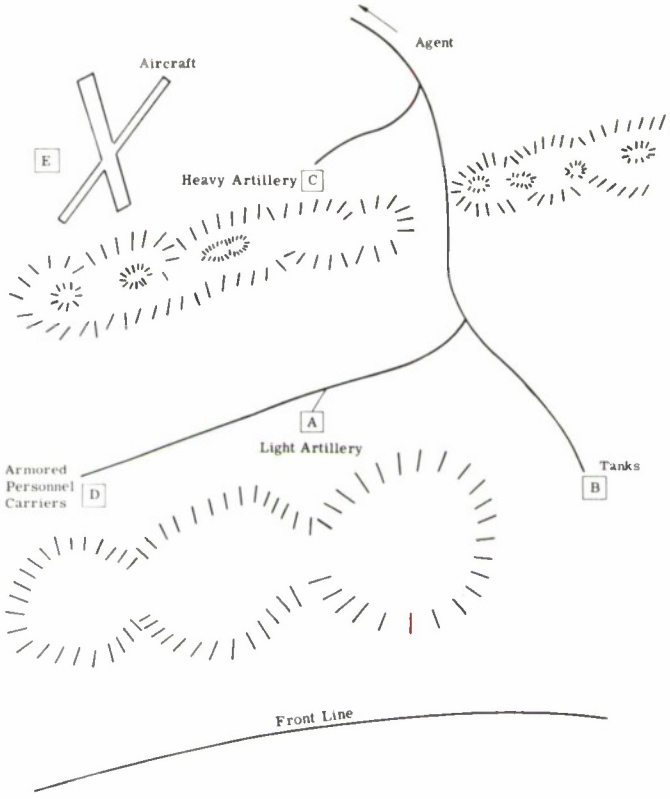


FIGURE 9. MAP OF ENEMY TERRITORY WITHIN BATTLE AREA

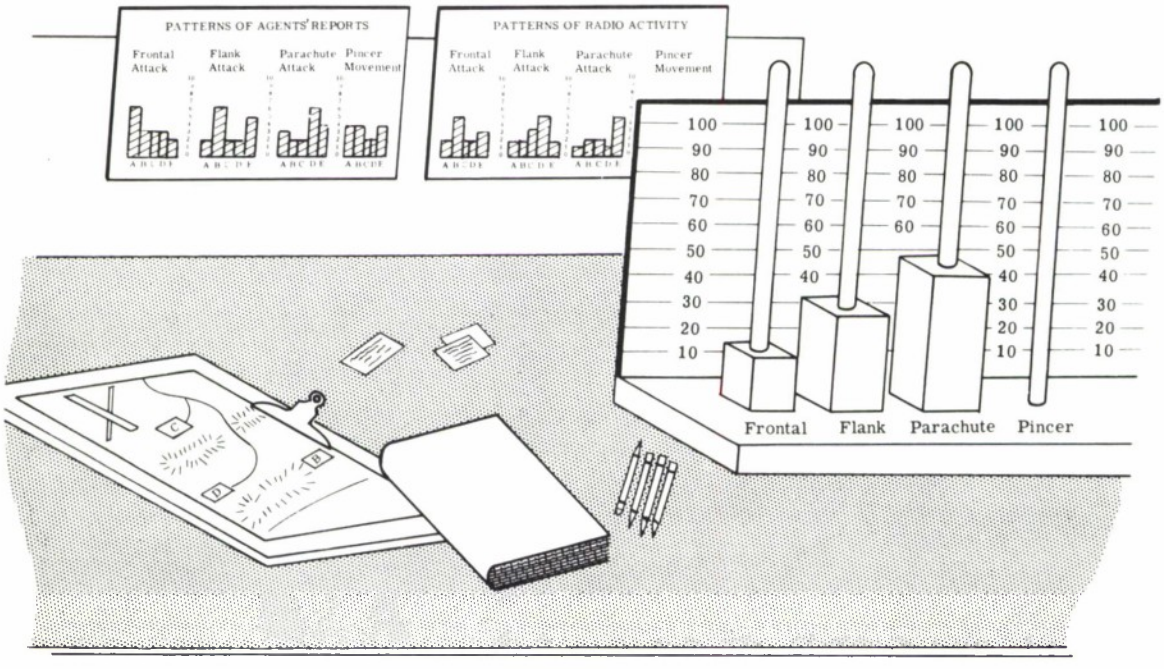


FIGURE 10. EXPERIMENTAL SETUP

For the abstract setting, the conditional probabilities were displayed in the form of histograms as they were for the military setting. The same response board was also employed, and pad and pencils were again available. Colored and/or dotted matches were used as evidence.

6.1.4. SUBJECTS. 64 male students of The University of Michigan between the ages of 18 and 24.

6.1.5. PROCEDURE. For the military setting, the subjects were settled in the experimental room and given the following instruction.

> This task has been designed to see how well persons can handle information in a military intelligence assessment situation. You have to imagine you are an intelligence officer in a unit which is defending this front line. Thus most of your land is off the map. The enemy is in this country shown on the map. (The features of the map are then described.)
>
> The enemy is about to attack you. This you know for certain. What you are unsure about is the form the attack will take. However, the constraints of the situation limit the possibilities to four. These are:
>
> (i) Frontal attack. The enemy will marshall a force of tanks in area B. He will then soften up your front line areas with light artillery which he will deploy behind these low hills and by shooting you up with pursuit planes which he will fly from this field (E). He will then move across here (i.e., between B and the front line) with his tanks. Infantry will ride on the tanks and run behind them and will infiltrate once a breakthrough has been made.
>
> (ii) Flank attack: For this the enemy will marshall a force of armored personnel carriers in area C, his aim being to strike around your left flank. (I must explain that the terrain on this flank is good enough for wheeled vehicles whereas on the right hand side of the front line it is suitable only for tanks.) Before he strikes he will soften up your rear areas with his heavy artillery which he will deploy behind the higher ridge of hills (area C).
>
> (iii) Parachute attack. For this the enemy will build up a force of transport planes on the field. Then he will soften up the dropping area (which is behind the line in the center of the front) with heavy artillery. He will also build up his tank force to some extent, to be ready to join up with the airborne forces once they are established and to worry the front line troops before them.
>
> (iv) Pincer movement. This is rather like a combination of the first two possibilities. The enemy will strike round both flanks simultaneously, using armored personnel carriers on the left and tanks on the right. Remember that the terrain prevents the use of faster vehicles on the right. His two forces will aim to link up with each other somewhere in your rear. Before moving off you can expect bombardment in your rear areas—those he will try to occupy—from his heavy artillery.
>
> (It is stressed that the order in which these four possibilities are described is of no significance.)
>
> You will get information about what the enemy intends to do from two sources. Firstly, there is an agent. He is hiding just off your map on a position where he

can observe movement on the road, and this is the enemy's sole supply route, and also aircraft over the airfield. He cannot communicate at any great length about what he sees or he would be detected. In fact, he is restrlcted to a rather crude form of communication. He has an equipment with five keys whlch connect with five different tones in your receiving equipment. If he sees a tank moving up the road he presses one key and you hear the corresponding tone. If he sees an armored personnel carrier, he presses a second and you hear a second tone and so on. While these messages are crude, the system ls reliable.

The second source of information is provided by a mechanism operated from your side of the line which can detect enemy radio transmissions. If some tanks are assembled here (area B) and one uses radio to talk with another, this device will detect the transmission and pinpoint the position of the transmitter. It will not be able to pick up the message. This detection provides a way of judging the size of forces deployed in any particular area. The more frequent the reports of radio activity the larger the forces deployed in any particular area. The more frequent the reports of radio activity, the larger the forces. You can expect the enemy command to impose some restrictions on the use of radios, but the relationship will still hold. If enemy forces are preparlng for attack they will have to communicate with each other.

The charts (i.e., the conditional probability displays) show you what to expect with each form of attack.

(E then ran right through the conditional probability distributions.)

You can see that regardless of the form the attack will take some reports of every kind can be expected. This is because the enemy will have all kinds of units deployed in the area and day-to-day replacements will be needed for those becoming defective. In determining the nature of his preparations you have to pick out those movements which are extra to this day-to-day activity.

A point which must be emphasized is that as intelligence officer you are offering a service to the commander. It is his job to decide what defensive action to take and how to deploy his forces. Your job is solely to tell him as best you can what the enemy is dolng.

In the real-life situation you could expect the general to call you at any moment demanding to know what the current situation is. When he does this he won't want to know what reports have come in. He will want to know the chances that a frontal attack is being launched rather than a flank or parachute or pincer attack. In other words, he will want to know the relative probabilities associated with each.

We do not have a general on the telephone here. Instead we have this device (the response board) on which I want you to display the probabllities. You see there are four pegs, one corresponding to each form the attack might take, and exactly one hundred washers. I want you to distribute the washers among the pegs so that the helght of the plle corresponds to the probability you assoclate with that alternative. To give an example of the way you should use this: If at some point the evidence led you to become absolutely certain that a pincer attack would take place, the other three possibllitles being completely out of the question, then you should have all one hundred washers on this peg. Contrariwise, if you felt from the evidence that pincer was impossible, the remaining three being equally likely, then these washers should be removed from here (pincer) and distributed equally among these other three columns.

You have to imagine the settings of these columns represent the estimates which would be reported to the general if he were to call.

The informatlon will come to you in the form of printed cards. (These were then briefly described.)

In this experiment you will work under rather ideal conditions. Whereas in a real military situation reports might come in quickly at times and you might be expected to make a very rapid assessment of them, here you will have as long as you like to consider each one. I shall hand you a card, which you will keep. You will then decide in your own time what it implies and then take any necessary adjustment to the heights of these four columns of washers. When this has been done I will record the levels and hand you the next report.

You have pad and pencil in case you care to make any notes. There are also grease pencils. You may use these to write on the map.

Any questions?

One final point must be settled before we begin. As you can see, there are 25 washers on each column. This implies that all four forms of attack are equally likely. If either from your interpretation of the military situation, or from your expectations because this is an experiment, you feel that one possibility is more likely than another, I want you to display this fact by adjusting the piles accordingly. I am not trying to induce you to have hunches, I simply want to know what they are if you've got them.

Once any necessary adjustments were made, the time was noted and the first report was handed to the subject. The session was run without interruption. At the end comments were invited and if the subject wanted to know what he should have done the normative probabilities were shown to him. The time was recorded at each tenth trial, so a rough indication of his rate of working was obtained.

For the abstract setting, four large imaginary urns were substituted for the four forms of attack the enemy could employ in the military task. Each urn supposedly contained a large number of matchsticks, which were coded by being painted at one end with one of five colors or by having from one to five stripes around them. Each matchstick corresponded to an intelligence report, colored ones corresponding to agent's reports while striped ones corresponded to reports of radio activity. The content of each urn was mixed according to the conditional probability distribution in table XIX. The subjects were handed matchsticks which they were told all came from one urn and their task was to assess the probability that each urn was the source, adjusting their assessments as each additional matchstick was added to their sample.

The design of the experiment was precisely the same as that of the military task. When two sources of information were required subjects were told the urns contained both colored and striped matchsticks. With one source, they were all colored or else all striped. Conditional probability distributions were displayed according to the condition being run, and paper and pencils were available to the subjects. At the beginning the subjects were questioned about their a priori beliefs but without the elaborate cover story of the military task there were no grounds for expecting one urn rather than another to be used as the source.

6.2. RESULTS AND ANALYSIS

6.2.1. THE MEASURES EMPLOYED. The subjects were required to make trial by trial adjustments to the four probabilities. Their efficiency can be assessed by computing the discrepancy between their settings on any given trial and the settings that would result if Bayes's theorem had been applied. Root mean square errors have been calculated on this basis. The normative solutions properly depend upon the a priori settings the subjects made before receiving any data. In fact all but nine of the 64 subjects began with equal settings. The prior probabilities for these nine are in table XXI. It can be seen that they were generally included to favor H2 and H4, but not very strongly. (Throughout this section reference is made to the four hypotheses H1-H4. H1 is the hypothesis that S1 is the state of the worlds, H2-4 similarly corresponds to S2-4.) Root mean square error has been calculated both with allowance for the subjects' prior settings (E_B) and without (E_L). In addition error has been calculated for each move separately. With E_B and E_L, the subjects' error on all but the first trial is the cumulative result of a number of adjustments. To assess the accuracy of each trial independently, E_N was computed by assuming the subjects' settings on trial N-1 to have been correct and comparing his adjustment on trial N with the appropriate adjustment calculated from Bayes's theorem.

To provide a simple graphic indication of performance, response to the critical hypothesis alone has also been singled out. The critical hypothesis is the one that the evidence eventually indicated to be true. For sequence 1 this was H1 (Frontal); for sequence 2 this was H3 (Parachute).

TABLE XXI. A PRIORI PROBABILITIES FOR THE NINE SUBJECTS WHO DID NOT RATE THE HYPOTHESES AS EQUALLY LIKELY

Subject	H1 Frontal	H2 Flank	H3 Parachute	H4 Pincer
1*	0.33	0.33	0.17	0.17
2*	0.30	0.20	0.30	0.20
3*	0.35	0.35	0.15	0.15
4*	0.10	0.25	0.25	0.40
5*	0.25	0.31	0.09	0.35
6*	0.20	0.25	0.19	0.36
7*	0.11	0.20	0.08	0.61
8*	0.15	0.30	0.15	0.40
9≠	0.23	0.23	0.23	0.31
Mean	0.22	0.27	0.18	0.33

*Military cover story

≠Abstract cover story

Finally, the time was recorded at the beginning of the test and subsequently after every 10 trials so that the rate at which subjects worked could be determined.

6.2.2. THE GENERAL PATTERN OF RESPONSE. Figure 11 shows the subjects' response to the critical hypothesis throughout the run. This follows the pattern of the normative response fairly closely but with diminished amplitude. The amplitude is reduced both when the probability of the hypothesis increases and when it decreases. This lack of response to the evidence also showed at times as a failure to make any adjustment whatever to the probabilities. During the

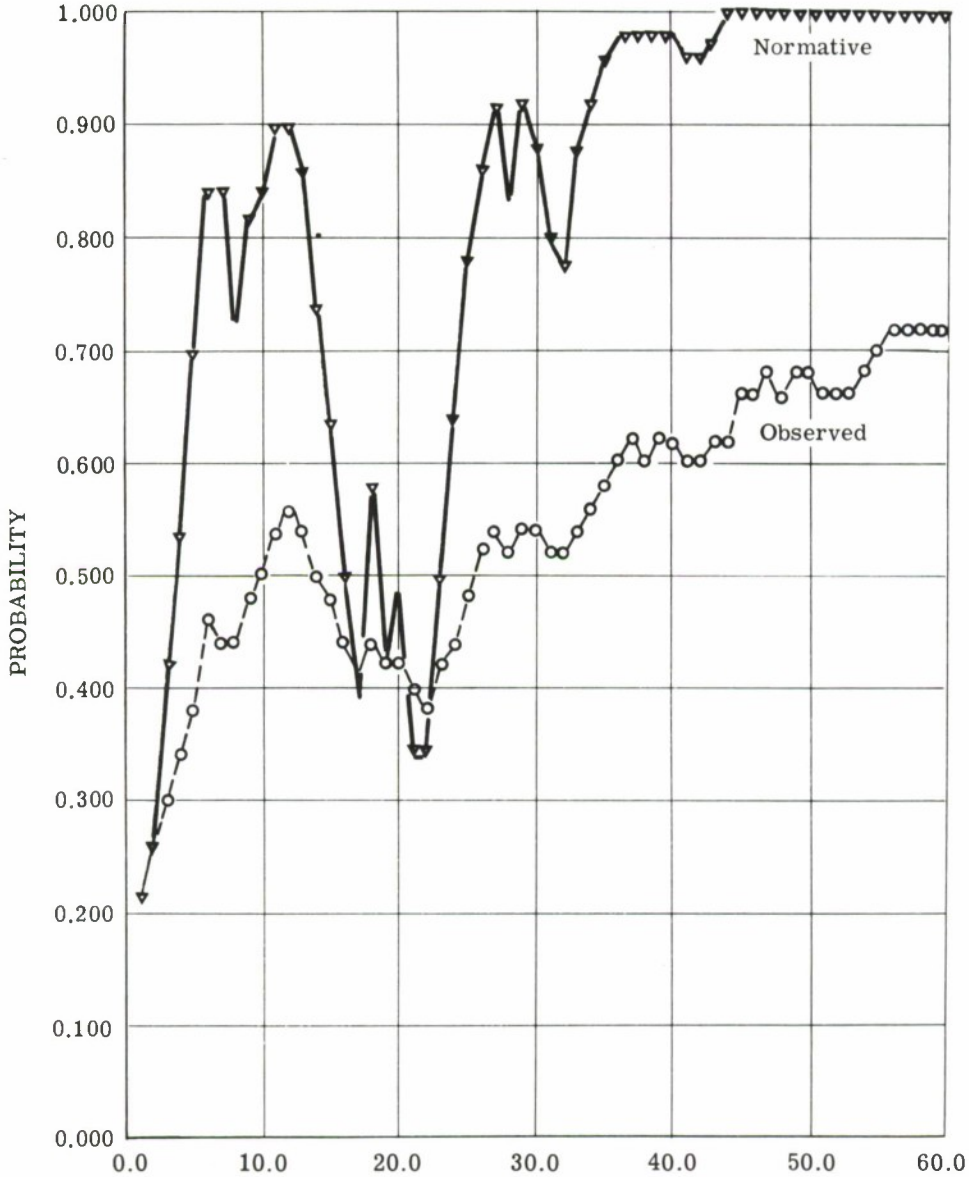


FIGURE 11. RESPONSE TO THE CRITICAL HYPOTHESIS: MEAN FOR ALL 64 SUBJECTS

first 20 trials substantial changes should have been made after each datum was received. Examinations of the protocols reveals that on average subjects made no response at all on 3.34 trials ($\sigma = 3.46$). The histogram in table XXII, which summarizes this analysis, shows that 73 percent of the subjects failed to respond on at least one trial and one made no adjustment on 15 of the 20.

TABLE XXII. FREQUENCY WITH WHICH THE EVIDENCE LED TO NO ADJUSTMENT OF THE POSTERIOR PROBABILITIES DURING TRIALS 1-20.

	Number of Trials with no Responses															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Subjects	17	11	7	5	3	1	5	4	4	3	2	0	0	0	0	1

Each subject's record has been examined on trial 10 and again on trial 20 to check the direction rather than the magnitude of their response to the critical hypothesis. On trial 10, 57 of the 64 correctly gave it the highest probability estimate. Six of the seven errors were made by subjects given the abstract task setting. On trial 60, 63 of the 64 were correct, the one error being made by a member of the "abstract" group.

The variability between subjects was considerable. Table XXIII shows the distribution of estimates of the probability of the critical hypothesis on trial 60, while figure 12 illustrates the range of variability by contrasting the behavior of one of the most responsive subjects with that of one of the least responsive. A total of 13 subjects put all 100 chips on the critical hypothesis on trial 60.

TABLE XXIII. FREQUENCY DISTRIBUTION OF SETTINGS OF PROBABILITY OF CRITICAL HYPOTHESIS ON TRIAL 60.

	Probability								
	0-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Number of Subjects	0	2	4	5	10	8	9	5	21

E_N decreased steadily throughout the run, the means for successive blocks of 10 trials being 0.093, 0.089, 0.088, 0.074, 0.067. This is largely because the normative calculations fairly rapidly indicated that the evidence supported one hypothesis and the subjects gradually accepted the same conclusion. In other words this decreasing error is a consequence of the lag illustrated in figure 11.

In addition to a failure to make a sufficiently high response to the critical hypothesis subjects showed lack of responsiveness in failing to adjust all four hypotheses when they should have done so. To demonstrate this failure each subject's response was examined on just one trial.

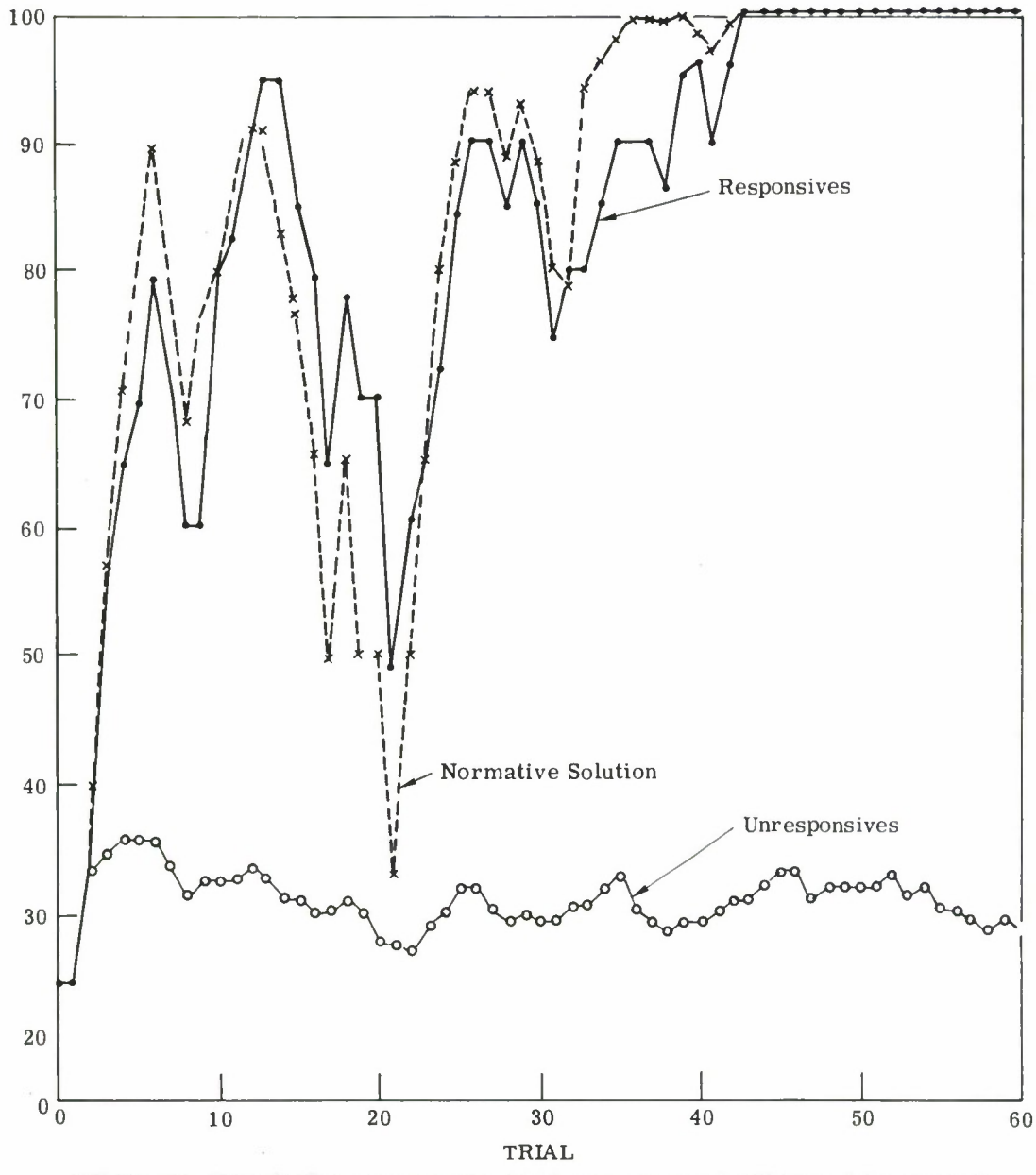


FIGURE 12. PROBABILITY OF THE CRITICAL HYPOTHESIS INDICATED BY THE MOST RESPONSIVE AND THE LEAST RESPONSIVE SUBJECTS

The data were sampled in the following way: For each subject one trial on which all four hypotheses should have been adjusted has been selected at random from trials 1-30. This meant that trials for which a particular subject's a priori probabilities included at least one value of zero or unity had to be excluded. The result of this count was as follows: 8 adjusted zero probabilities, 23 adjusted two, 10 adjusted three, and 23 adjusted all four.

Although by trial 10 nearly all the subjects correctly put the highest setting on the most probable hypothesis the relative settings on the remaining three could be in error. One problem

that arises in analyzing all four settings is that the probabilities of both H2 and H4 rapidly approached zero so that by trial 10 they did not differ appreciably. In order, therefore, to make the examination at a point where there were relatively large differences in all four normative probabilities only the very early trials are relevant. In fact, substantial differences between all four hypotheses existed only on trials 2 and 3. Of these, trial 3 has been selected for examination. The normative settings on this trial were: H1, 57; H2, 04; H3, 28; H4, 11 for sequences 1 and H1, 55; H2, 02; H3, 28; H4, 15 for sequence 2. Thus the correct rank order was the same for both sequences. On this trial 47 subjects correctly put most chips on H1. Of these 47, 27 correctly ranked H3 as second and 22 also ordered the remaining hypotheses correctly. Only 34 percent of the subjects, therefore, correctly ranked all four hypotheses whereas 73 percent had correctly selected the most likely one. In this aspect of performance the setting of the task had no effect. Data from those subjects given the military setting were almost identical to those from subjects given the abstract setting.

6.2.3. THE EFFECT OF THE EXPERIMENTAL VARIABLES. Analyses of variance were computed separately on E_L, E_B, and E_N for trials 1, 10, 20, and 50; on the means for each block of ten trials; and on the overall mean for all 60 trials. Apart from a significantly greater response ($p < 0.01$) leading to reduced error with the abstract setting on trial 1, these analyses failed to reveal any statistically significant differences for any of the variables. And this one finding is misleading, since the general trend throughout the 60 trials was for subjects to respond to the critical hypothesis more with the military setting. The trends for the other variables are shown in table XXIV. Where the response to the critical hypothesis was greatest, the root mean square errors were least.

TABLE XXIV. THE EFFECT OF EXPERIMENTAL VARIABLES ON RESPONSE TO THE CRITICAL HYPOTHESIS

Variable	Condition	Trial 10	Trial 60
Setting	Military	0.54	0.73
	Abstract	0.46	0.71
Number of Sources	1	0.54	0.78
	2	0.46	0.66
Division of Information Between Sources	Biased	0.39	0.59
	Normal	0.52	0.73

6.2.4. RECENT vs. OLD INFORMATION. Sequences Q1 and Q2 presented the same evidence in trials 1-20. With Q1 the first block (trials 1-10) were consistent with H1 and the second block was consistent with H3. With Q2 the order of these blocks were reversed. If subjects were in-

fluenced differentially by old data or by new data this should lead to differences in their settings on trial 20. Thus if they were influenced more strongly by the early data they would favor H1 with Q1 and H3 with Q2. The relevant results were as follows: with Q1, mean response to H1 on trial 20 was 0.4381 and to H3 it was 0.3425; with Q2 the mean response to H1 was 0.3775 and to H3 0.4094. Thus there is an indication that early data exerts more influence than recent. Analysis of each subject's bias shows that 39 were most influenced by early data, 17 by recent, while 8 showed no bias in either direction. A "t-test" shows the differences to be significant ($t = 27.2$, 63 df $p < 0.01$).

6.2.5. THE RATE OF WORKING. Most subjects worked rather slowly but the variability was considerable. Mean time for the 60 moves was 53 min., and σ, 23 min. The extremes were 19 min. and 137 min.

6.2.6. GENERAL FEATURES OF THE SUBJECTS' REACTION TO THE TASK. Most subjects arranged the reports they received to form a histogram and matched this against the conditional probability histograms. Their responses therefore indicated the relative similarity between the pattern formed by the data and each of the conditional probability distributions. With two sources of data they had two separate comparisons to make, but those given the military cover story marked the map with grease pencils as reports were received and in doing so most detected the equivalence between the "agent's" reports and the "radio activity" reports. This enabled them to behave as if only one source were involved. With the abstract task, collating the data from two sources caused obvious difficulty to some subjects. Thus one, who had to deal with the asymmetric division between sources, made this remark at the end of the experiment: "The numbers show it couldn't possibly be II. Colors show it couldn't possibly be IV and it certainly isn't I or III."

Many subjects were obviously at a loss to know what magnitude of response was appropriate, and some said so. To quote one such subject (trial 4), "It's hard to know what weight to give these various movements." This same subject on trial 37, when he had assessed the probability of the critical hypothesis at 0.38 (normative setting 0.98), remarked "Getting mighty high!" Reluctance to change away from the null setting at the beginning of the urn was correlated in some cases with remarks to the effect that the data were insufficient to lend support to any hypothesis. Thus subjects while making no adjustment to the probabilities made remarks such as "It could come from any one of them."

It has been noted in the formal analysis of the data that subjects were also unwilling to reduce their settings on a given hypothesis when contrary evidence arrived. Some made comments that illuminate this reluctance. Thus one who had the sequence that indicated parachute attack,

then frontal attack, then parachute attack, commented on trial 24 when the data were again beginning to indicate parachute, "Everything seems to concur with a frontal attack at the moment. May I have more information? I hate to change my opinions."

A solution that some subjects appeared to adopt to the general problem of how large an adjustment was appropriate on each trial was always to move a constant number of chips. Thus several usually moved five at a time, while others moved just one.

Fifteen of the subjects made some form of paper and pencil calculation. This number includes some who calculated on every trial and others who made occasional calculations. Some of these did no more than work out the percentages of reports of each kind to facilitate pattern matching. One subject, a mathematician who was planning to major in statistics, laboriously worked by hand a calculation very similar to the correct Bayesian procedure except that he added each prior probability to the appropriate conditional probability instead of multiplying them together. Sometimes abrupt changes of strategy were made by subjects. Some began by calculating and then gave up and proceeded thereafter intuitively. Others changed their system of calculation. In some cases this led, through reconsideration of all the data received, to a large change in their assessments. This change was not the response to the datum received on the particular trial when the reconsideration was made.

6.2.7. FALLACIES EXHIBITED. A combination of behavior and comments made by some subjects revealed two fallacies that at least one subject committed. One of these was a failure to respond when data were received that reinforced the conclusions already reached from previous data. Thus one subject in such a situation made no adjustment of the probabilities and remarked "This seems to concur with what I have so far."

The second was to infer that a perfect match between the data and our distribution of conditional probabilities implied that this distribution was definitely the source. Such an inference led some subjects to put all 100 chips on the critical hypothesis on trial 10 when the normative level was 0.80 for sequence 1 and 0.86 for sequence 2.

A related fallacy was committed by the experimenter in planning the experiment. The perfect match between data and conditional probability was used with the intention of presenting the strongest possible evidence in favor of the critical hypothesis. In fact the strongest possible indication of a frontal attack over ten trials would have been ten reports of light artillery. With these the normative probability would be nearer to 100 than to 0.99, whereas the normative probability given a perfectly matching distribution was 0.89.

6.2.8. WAYS IN WHICH INSTRUCTIONS WERE DISREGARDED. Some subjects found it difficult to evaluate evidence without making a decision on the basis of that evidence. With the

military task this was revealed in some cases by questions about the defensive forces. In other cases it was revealed by comments at the end of the session, such as "And now I suppose you want me to make a decision." Some thought a high assessment on one hypothesis meant that a choice had been made, as the subject who said at the end: "I guess I should have decided sooner — meanwhile 10,000 men would have died." Another reaction was in the opposite direction: another subject commented afterwards "I thought: what would I tell the general right now. What if I said 'parachute' and it turned out to be wrong?" Thus subjects are concerned with both types of statistical error.

Some subjects given the military task failed to consider the four hypotheses as exclusive. One commented specifically that he was not convinced that there were four discrete forms of attack. A related phenomenon was a refusal to use the conditional probabilities in a simple-minded way. Subjects believed the enemy would be elusive and that some movements he made would be deliberate attempts to upset the intelligence system. Thus one who was reluctant to make a substantial change in his assessments when the evidence was apparently strong said that he did not want to be more extreme in his settings because "an enemy wouldn't display all that clear a picture of his movements."

6.2.9. AFFECTIVE REACTION. Finally it might be added that the subjects enjoyed the experimental task — so much so that one remarked "It's more fun than chess" while another suggested it should be patented as a game.

6.3. DISCUSSION

Although the evaluation of uncertain evidence is an ingredient of much everyday behavior, as was pointed out in the Introduction, it would appear from the results of this experiment that most persons do it rather badly. The subjects generally underrated the significance of the data they received and were unable to rank the probability of the four hypotheses correctly. Thus they erred in both the direction and the magnitude of their probability assessments.

There are at least two explanations of this apparent contrast between the experiment and everyday behavior. One is that, although they are very inefficient, persons are not usually troubled by their inadequacies because everyday situations are undemanding. The other is that the experimental results are artifactual and persons are not necessarily as inefficient outside the laboratory.

Some hints that the experimental data could contain magnitude errors were given by the subjects in their comments. At times it was clear that they did not fully express their opinions by their settings because they were waiting for more information. In the experiment they expected rather a large number of reports, and although instructed to keep their assessments up to date,

they did not always do so. If the experiment had been stopped at an intermediate stage and the subject had been told that all the information which would ever be available was already before him, it is conceivable that he would have been willing to make more extreme settings of the probabilities.

Another exercise that might reveal the extraction of more information from the reports would be to demand that the subjects should base a decision upon it. But although this approach is superficially attractive, it would appear difficult to infer the subjective probabilities from the resultant choices with any precision.

From the comments made by some subjects given the military version of the experimental task it is clear that some felt they were in fact controlling a decision by their settings. It is interesting to note that in the comments mention was made of both type I and type II errors. Some subjects would not wish to precipitate too hasty a decision, whereas others would not wish to delay unnecessarily.

Despite the poor performance of most subjects, there were a few whose response to the critical hypothesis was extremely close to the normative solution. A question raised by their performance is whether or not it should be attributed to chance. The test, of course, would be to check their performance in other situations. Unfortunately this is not possible, but a separate study could be run with the sole purpose of checking on the consistency of individual differences with particular interest in those who perform well.

Another way of regarding the few good performances is to seek remedies for the poor performance of the majority of subjects — in other words, to train them. The way in which many were obviously at a loss to know the appropriate magnitude of response suggests that guidance on a few trials might provide the scale they need. But there were many subjects (66 percent on trial 3) who failed to set the four probabilities in the correct rank order of magnitude and these need more than a guide to the scale of response. Simultaneous adjustment of four hypotheses might exceed the information-processing capacity of the human subject. But it is conceivable that a way of breaking down the task, perhaps by successive binary comparions, could be evolved and the appropriate procedure successfully taught.

The variability between subjects was extremely great in this task. Training can be regarded as one way of reducing variability. Other less elaborate procedures could possibly contribute as well. The subjects varied noticeably in the way they tackled the task. To mention two facets of their behavior: they varied in the way they sorted out the reports they were given and also in deciding whether or not to attempt to work intuitively or to calculate on paper. All this variability could possibly have been reduced by additional instructions.

In allowing subjects paper and pencils the procedure employed here differed from that generally used in studies of intuitive statistics. The reason for providing these aids was primarily to optimize performance. If subjects thought some explicit calculation appropriate and endeavored to carry out the calculation in their heads, their performance would differ from the subject given paper and pencil only in the addition of an error introduced through their inability to carry out the mental calculation correctly. Viewed in this way, intuitive statistics are contaminated by unnecessary error. But there could be a case for insisting that subjects should work without calculating aids in that there is an interest in generalizing from artificial situations where conditional probability distributions can be specified precisely, to real-life intelligence tasks where the conditional probabilities are only known intuitively by the subject. In these there is no possibility whatsoever of formal calculation.

The nature of the calculations subjects carried out was always false. Not being statisticians, they were unaware of the appropriate procedure. This raises the point that without formal training or a great deal of thought the rules of probability theory are not apparent. This is not surprising, since the origins of probability theory lay in the awareness of the perceptive gambler that his intuitions were faulty and his consequent employment of a mathematician as advisor. In the present study it would appear that the wheel has been turned a full circle.

REFERENCES

Dodson, J. D., Simulation System Design for a TEAS Simulation Research Facility, Report AFCRL 1112 and PRC R-194, Planning Research Corporation, Los Angeles, November, 1961.

Edwards, W., "Probability-Preferences in Gambling," Amer. J. Psychol., Vol. 66, 1953, pp. 349-364.

Edwards, W., "The Theory of Decision Making," Psychol. Bull., Vol. 51, 1954a, pp. 380-417.

Edwards, W., "Probability-Preferences Among Bets with Differing Expected Values," Amer. J. Psychol., Vol. 67, 1954b, 56-67.

Edwards, W., "Behavioral Decision Theory," Annu. Rev. Psychol., Vol. 12, 1961, pp. 473-498.

Edwards, W., "Utility, Subjective Probability, Their Interaction and Variance Preferences," J. Conflict. Resol., Vol. 6, 1962a, pp. 42-51.

Edwards, W., "Dynamic Decision Theory and Probabilistic Information Processing," Human Factors, Vol. 4, 1962b, pp. 59-73.

Edwards, W., "Subjective Probabilities Inferred from Decisions," Psychol. Rev., Vol. 69, 1962c, pp. 109-135.

Edwards, W., Probabilistic Information Processing in Command and Control Systems, ESD-TDR-62-345, Report No. 3780-12-T, Institute of Science and Technology, The University of Michigan, Ann Arbor, February, 1963.

Edwards, W., Human Processing of Equivocal Information, ESD-TDR-64-601, Report No. 3780-23-F, Institute of Science and Technology, The University of Michigan, Ann Arbor, April, 1965.

Edwards, W., Lindman, H., and Savage, L. J., "Bayesian Statistical Inference for Psychological Research," Psychol. Rev., Vol. 70, 1963, pp. 193-242.

Edwards, W., and Phillips, L. D., "Man as Transducer for Probabilities in Bayesian Command and Control Systems," G. L. Bryan and M. W. Shelly, eds., Human Judgments and Optimality, Wiley, 1964, pp. 360-401.

Feallock, J. B., and Briggs, G. E., A Multi-Man-Machine System Simulation Facility and Related Research on Information-Processing and Decision-Making Tasks, Technical Documentary Rept. No. AMRL-TDR-63-48, Laboratory of Aviation Psychology, Ohio State University, Columbus, June 1963.

Fox, W. R., and Hackett, E., Empirical Techniques for the Design, Analysis, and Evaluation of Command and Control System Displays, Report No. TO-B 63-112, Technical Operations Research, 1964.

Gough, H. G., "Clinical vs. Statistical Prediction in Psychology," Leo Postman, ed., Psychology in the Making, Alfred A. Knopf, 1962.

Herman, L. M., Ornstein, G. N., and Bahrick, H. P., "Operator Decision Performance Using Probabilistic Displays of Object Location," I.E.E.E. Trans. on Hum. Fact. in Electronics, September 1964, pp. 13-19.

Kaplan, R. J., Lichtenstein, S., and Newman, J. R., PIP Study No. 2: Probabilistic Information Processing Under Conditions of Varying Payoff Task Difficulty, Report No. TM-1150/000/00, System Development Corporation, Santa Monica, Calif., 1963.

Kaplan, R. J., and Newman, J. R., A Study in Probabilistic Information Processing (PIP), Report No. TM-1150/000/00, System Development Corporation, Santa Monica, Calif., 1963.

Lindman, H., The Simultaneous Measurement of Utilities and Subjective Probabilities, unpublished doctoral dissertation, The University of Michigan, 1965.

Meehl, P., Clinical Versus Statistical Prediction, University of Minnesota Press, Minneapolis, 1954.

Peterson, C. R., and Miller, R. J., "Sensitivity of Subjective Probability Revision," J. exp. Psychol., Vol. 70, 1965, pp. 117-121.

Peterson, C. R., Schneider, R. J., and Miller, A. J., "Sample Size and the Revision of Subjective Probabilities," J. exp. Psychol., Vol. 69, 1965, pp. 522-527.

Phillips, L. D., and Edwards, W., "Conservatism in a Simple Probability Inference Task," J. exp. Psychol., Vol. 72, No. 3, September 1966, pp. 346-354.

Phillips, L. D., Hays, W. L., and Edwards, W., "Conservatism in Complex Probability Inference," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 7-18.

Raiffa, H., and Schlaifer, R., Applied Statistical Decision Theory, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Robinson, G. H., Continuous Human Estimation of a Time-Varying, Sequentially Displayed Probability, unpublished doctoral dissertation, The University of Michigan, Ann Arbor, 1962.

Savage, L. J., The Foundations of Statistics, Wiley, 1954.

Sawyer, J., Department of Psychology, University of Chicago, "Measurement and Prediction, Clinical and Statistical," unpublished paper, 1963.

Schlaifer, R., Probability and Statistics for Business Decisions, McGraw-Hill, 1959.

Schlaifer, R., Introduction to Statistics for Business Decisions, McGraw-Hill, 1961.

Schneider, R. J., Optimality of Subjective Probabilities in Clinical Inference, unpublished Master's thesis, University of Colorado, Boulder, 1965.

Schum, D., Goldstein, I., and Southard, J., "Research on a Simulated Bayesian Information-Processing System," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, 1966, pp. 37-48.

Shuford, E. H., "Percentage Estimation of Proportion as a Function of Element Type, Exposure Time, and Task," J. exp. Psychol., Vol. 61, 1961, pp. 430-436.

Southard, J. F., Schum, D. A., and Briggs, C., An Application of Bayes's Theorem as a Hypothesis-Selection Aid in a Complex Information-Processing System, Report No. AMRL-TDR-64-51, Wright-Patterson Air Force Base, Ohio, 1964a.

Southard, J. F., Schum, D. A., and Briggs, G. E., Subject Control over a Bayesian Hypothesis-Selection Aid in a Complex Information-Processing System, Report No. AMRL-TR-64-95, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio, September 1964b.

Toda, M., Measurement of Subjective Probability Distribution, Division of Mathematical Psychology, Institute of Research, State College, Pennsylvania, Report No. 3, 1963.

Tversky, A., Additive Choice Structures, unpublished doctoral dissertation, The University of Michigan, Ann Arbor, 1965.

van Naerssen, R. F., "A Scale for the Measurement of Subjective Probability," Acta Psychol., Vol. 20, 1962, pp. 139-166.

Warner, H. R., Toronto, A. F., Veasey, I. C., and Stephenson, R., "A Mathematical Approach to Medical Diagnosis. Application to Congenital Heart Disease," J. Amer. Med. Assoc., 1961, Vol. 177, pp. 177-183.

Yntema, D. B., and Torgerson, W. S., "Man-Computer Cooperation in Decisions Requiring Common Sense," I.E.E.E. Trans. on Hum. Fact. in Electronics, March, 1961, pp. 20-25.

Appendix I
PUBLICATIONS UNDER CONTRACT AF 19(628)-2823

Edwards, Ward

With Lindman, H., and Savage, Leonard J., "Bayesian Statistical Inference for Psychological Research," Psychol. Rev., Vol. 70, 1963, pp. 193-242.

Probabilistic Information Processing in Command and Control Systems (ESD-TDR-62-345), Report No. 3780-12-T, Institute of Science and Technology, The University of Michigan, Ann Arbor, February 1963.

With Phillips, L. D., "Man as Transducer for Probabilities in Bayesian Command and Control Systems," in G. L. Bryan and M. W. Shelly, eds., Human Judgments and Optimality, Wiley, 1964.

"Probabilistic Information Processing by Men, Machines, and Man-Machine Systems," Proceedings of the XVIIth International Congress of Psychology," North-Holland Publishing Co., Amsterdam, 1964.

"The Design and Evaluation of Probabilistic Information Processing Systems," Proceedings of the Fifth National Symposium on Human Factors in Electronics, Institute of Electronic and Electrical Engineers, New York, 1964.

With Lindman, H., and Phillips, L. D., "Emerging Technologies for Making Decisions," New Directions in Psychology II, Holt, Rinehart and Winston, 1965, pp. 261-325.

Human Processing of Equivocal Information (ESD-TDR-64-601), Report No. 3780-23-F, Institute of Science and Technology, The University of Michigan, April 1965.

"Information Seeking to Reduce the Risk of Decisions," Predecisional Processes in Decision Making: Proceedings of a Symposium, 1964, Report No. AMRL-TDR-64-77, Aerospace Medical Research Laboratories, Wright-Patterson AF Base, Ohio, March 1965, pp. 5-19.

"Optimal Strategies for Seeking Information: Models for Statistics, Choice Reaction Times, and Human Information Processing," J. Math. Psychol., Vol. 2, 1965, pp. 312-329.

With Slovic, S. P., and Lichtenstein, S., "Boredom Induced Changes in Preferences Among Bets," Amer. J. Psychol., 1965, pp. 208-217.

"A Tactical Note on the Relation Between Scientific and Statistical Hypotheses," Psychol. Bull., Vol. 63, 1965, pp. 400-402.

"Probabilistic Information Processing Systems for Diagnosis and Action Selection," Information System Sciences: Proceedings of the Second Congress, Spartan Books, 1965.

With Phillips, L. D., and Hays, W. L., "Conservatism in Complex Probabilistic Inference," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 7-18.

"Introduction: Revision of Opinions by Men and Man-Machine System," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 1-6.

With Tversky, A., "Information vs. Reward in Binary Choices," J. exp. Psychol., Vol. 71, No. 5, May 1966, pp. 680-683.

With Phillips, L. D., "Conservatism in a Simple Probability Inference Task," J. exp. Psychol., Vol. 72, No. 3, September 1966, pp. 346-354.

Beach, Lee Roy

"Accuracy and Consistency in the Revision of Subjective Probabilities," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 29-36.

Peterson, C. R., and Phillips, L. D.

"Revision of Continuous Subjective Probability Distributions," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 19-22.

Slovic, Paul

"Value as a Determiner of Subjective Probability," I.E.E.E. Trans. on Hum. Fact. in Electronics, Vol. 7, March 1966, pp. 22-28.

Phillips, L. D.

Some Components of Probabilistic Inference, Ph.D. thesis, The University of Michigan, 1966.

DISTRIBUTION LIST

Copy No.	Addressee
3	Electronic Systems Division Air Force Systems Command Laurence G. Hanscom Bedford, Massachusetts ATTN: ESRHT, Project Officer Contract AF19(628)-2823

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)* Institute of Science and Technology, The University of Michigan, Ann Arbor	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP N/A

3 REPORT TITLE

NONCONSERVATIVE PROBABILISTIC INFORMATION PROCESSING SYSTEMS

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final Report, September 1963 - April 1966

5 AUTHOR(S) *(Last name, first name, initial)*

Edwards, Ward

6 REPORT DATE December 1966	7a TOTAL NO OF PAGES viii + 85	7b NO OF REFS 40
8a CONTRACT OR GRANT NO. AF 19(628)-2823 b. PROJECT NO. 2806 c. d.	9a ORIGINATOR'S REPORT NUMBER(S) 5893-22-F 9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* ESD-TR-66-404	

10 AVAILABILITY/LIMITATION NOTICES

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY Electronic Systems Division Air Force Systems Command, USAF, L. G. Hanscom Field, Mass.

13 ABSTRACT This report is concerned with two large-scale simulation experiments on probabilistic information processing (PIP) systems. One, a very large and prolonged study of four systems, yielded the conclusion that PIP is indeed an efficient philosophy for information-processing systems — at least twice as efficient as its next-best competitor, and four times as efficient as a representative of current processing techniques. The second PIP experiment was concerned with whether likelihood estimators in PIPs should be allowed to know the state of system opinion; the data confirm the suggestion that it might be undesirable. These experiments required the use of an on-line computer system.

This comparison of PIP and its competitors clearly indicates that PIP is superior, but does not indicate how PIP compares with theoretically optimal performance since no objective model of the data-generating process was available. A smaller-scale laboratory experiment is reported that compares PIP with a posterior-odds estimation system (POP) in a task sufficiently complex to be difficult for subjects and yet allowing an objective standard of correct performance. PIP was far superior to POP. PIP and calculations of optimal performance were roughly comparable, with PIP sometimes more extreme than optimal performance and sometimes less extreme. Another small laboratory study, concerned with the development of a response mode in which subjects report on probabilities by making choices among bets, is reported. Its original purpose was to develop a response mode for one group in the first PIP experiment, but it proved to be considerably more important than that. A study is reported in which the fact of human conservatism in information processing, the fact with which PIP is designed to cope, is again demonstrated under conditions of realistic complexity that have a military flavor.

People are shown to be conservative information processors. To cope with this it is appropriate to design information processing systems in which human estimates of likelihood ratios are followed by computer aggregation of these into posterior distributions by means of Bayes's theorem. Such procedures extract information from data more efficiently than any other way of exploiting human judgment yet tried, and produce data roughly comparable with theoretically optimal calculations when such calculations are possible.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Information processing systems						
Probability						
Human engineering						
Theorem (Bayes's)						
Behavior						
Decision making						
Experimentation						
Design						
Game theory						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.