AMRL-TR-65-25

THE INFLUENCE OF EXPERIENCE AND INPUT INFORMATION FIDELITY UPON POSTERIOR PROBABILITY ESTIMATION IN A SIMULATED THREAT-DIAGNOSIS SYSTEM

DAVID A. SCHUM
IRWIN L. GOLDSTEIN
JACK F. SOUTHARD

OHIO STATE UNIVERSITY

APRIL 1965

BEHAVIORAL SCIENCES LABORATORY
AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

NOTICES

When US 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.

Requests for copies of this report should be directed to either of the addressees listed below, as applicable:

Federal Government agencies and their contractors registered with Defense Documentation Center (DDC):

DDC
Cameron Station
Alexandria, Virginia 22314

Non-DDC users (stock quantities are available for sale from):

Chief, Input Section
Clearinghouse for Federal Scientific & Technical Information (CFSTI)
Sills Building
5285 Port Royal Road
Springfield, Virginia 22151

Change of Address

Organizations and individuals receiving reports via the Aerospace Medical Research Laboratories automatic mailing lists should submit the addressograph plate stamp on the report envelope or refer to the code number when corresponding about change of address or cancellation.

Do not return this copy. Retain or destroy.

900 - May 1965 - 448-41-901

THE INFLUENCE OF EXPERIENCE AND INPUT INFORMATION FIDELITY UPON POSTERIOR PROBABILITY ESTIMATION IN A SIMULATED THREAT-DIAGNOSIS SYSTEM

DAVID A. SCHUM
IRWIN L. GOLDSTEIN
JACK F. SOUTHARD

FOREWORD

The research described was performed in the Laboratory of Aviation Psychology, The Ohio State University, Columbus, Ohio, under Air Force Contract No. AF 33(657)-10763 during the period 1 October 1963 to 1 June 1964. This research was performed in support of the Aerospace Medical Research Laboratories Project 7184, "Human Performance in Advanced Systems," Task 718403, "Man-Machine Systems Research." Dr. George E. Briggs was the principal investigator representing The Ohio State University. The contract was monitored by Capt. Karl L. Wiegand, Chief, Systems Research Branch, Engineering Psychology Division, Behavioral Sciences Laboratory. The two experiments being reported formed the basis for a doctoral dissertation presented by the senior author to The Ohio State University Graduate School.

The authors are pleased to acknowledge the assistance and guidance of Dr. George E. Briggs, Dr. William C. Howell, and Dr. James C. Naylor, all of The Ohio State University. Messrs. Edwin R. Lassettre and Lonnie D. Whitehead were primarily responsible for preparing the large amount of computer programming which the present research required. Miss Carolyn Black provided invaluable assistance in the collection of data and preparation of stimulus materials. Miss Barbara Lindig and Miss Janis Frye provided much editorial assistance in the preparation of the report manuscript.

This technical report has been reviewed and is approved.

WALTER F. GRETHER, PhD
Technical Director
Behavioral Sciences Laboratory

ABSTRACT

This report describes two experiments in which posterior probability estimates made by humans are compared with similar estimates made by a computer using a modification of Bayes' theorem incorporating human estimates of P(D|H). The task was to estimate, on the basis of intelligence data from a simulated threat-evaluation situation, the likelihood of various alternative hypotheses that could account for the observed data.

The purpose of the first experiment was to determine the effect of increased experience upon the human's ability to estimate posterior probabilities. With increased experience the subjects' performance improved. In terms of the size of the estimates placed in the correct-hypothesis category there were no overall statistically significant differences between the subjects' estimates and the Bayesian calculations. However, the Bayesian solution placed significantly more first-choice estimates in the correct hypothesis categories.

The purpose of the second experiment was to compare human and automated posterior probability estimates under several levels of input data fidelity. It was predicted that, under low fidelity conditions, human posterior probability estimates would become increasingly inferior to automated solutions. This hypothesis was only partially confirmed. In both experiments, but particularly in the second, the humans provided higher posterior probability estimates than the certainty in the data justified. Several reasons for these excessive estimates are discussed.

With respect to the design of diagnostic systems, the present research tends to confirm the feasibility of automated Bayesian hypothesis-selection incorporating expert human estimates of the conditional probabilities P(D|H).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

I. HYPOTHESIS SELECTION AND BAYES' THEOREM

In an effort to understand, describe, and predict various aspects of human performance, psychologists have made rather free use of formal mathematical statements or models originally developed for other disciplines such as physics, economics, and statistics. Information theory, servomechanism theory, game theory, and decision theory are examples of formal systems appropriated by psychologists in their search for orderly relationships in human behavior. In many cases these formal statements specify the requirements for optimal performance under a prescribed set of conditions. Unhappily for psychology, it is easier to specify mathematically optimal performance than it is to predict how humans will actually perform. For example, the various decision theory models discussed by Chernoff and Moses (ref. 1) are "canons of reationality" in the sense that they specify how a decision maker "ought" to behave in order to maximize his expected utility. In the face of human inconsistencies, however, these models fail to be adequate descriptions of actual human performance. Nevertheless, these imported theoretical systems have often suggested methodological innovations highly useful to psychologists in their investigations of a variety of human tasks. Shuford has recently mentioned that although information theory, game and decision theory, and dynamic programming (all of which he calls "purposive mathematics") are not exact models of human behavior, they can be used profitably by psychologists in their investigations of the logical structure of behavioral tasks (ref. 2).

One task area of current interest concerns situations in which humans are required to ascribe causes to effects or, more specifically, to estimate the likelihood of various alternative hypotheses in the light of observed data and to revise these estimates as new data become available. The diagnosis made by a physician on the basis of a set of observed symptoms provides a convenient example of the task of ascribing causes (hypothesized states of nature) to effects (observed data). Similarly, there are military agencies in which intelligence data are meticulously processed and evaluated in an attempt to discern the intent of some real or potential adversary whose activities are under surveillance. It happens that there is a formal quantitative statement which describes, subject to the acceptance of certain conditions, how one ought to revise his opinions about the probability of some hypothesis in the light of new data or experience. The Reverend Thomas Bayes seems to have been the first to develop an exact and quantitative statement of inductive inference. This statement, subsequently called "Bayes' theorem," appeared in an article published postumously in the Philosophical Transactions of the Royal Society in 1763 (ref. 3). Bayes' theorem, which logically follows from the notion of conditional probabilities and from the fact that probabilities assigned across some hypothesis set must have unit sum, states that the probability of a hypothesis given an observed datum (the a posteriori probability of the hypothesis) is equal to the normalized product of the a priori probability of the hypothesis and the probability of the datum given that the hypothesis is true. In modern symbols this statement is described as follows:

$$P(H_i|D) = \frac{P(H_i)\ P(D|H_i)}{\sum_{k=1}^{n} P(H_k)\ P(D|H_k)} \qquad \text{(Eq. 1)}$$

where

$P(H_i|D)$ = the probability of hypothesis i in the light of datum D (the a posteriori probability of hypothesis i).

$P(H_i)$ = the a priori probability of hypothesis i (or the probability of hypothesis i before the observation of the datum).

$P(D|H_i)$ = the probability of the datum if hypothesis i is true.

n = the number of hypotheses in the mutually exclusive and exhaustive set of hypotheses.

$\sum_{k=1}^{n} P(H_k)\, P(D|H_k)$ = the normalizing constant which assures that $\sum_{k=1}^{n} P(H_k|D) = 1.0$.

The major impediment in the path of universal practical application of this theorem results from the fact that there is no universal agreement about the definition of probability. Good (ref. 4) has recently described five different explicit definitions of probability that have been formulated throughout the years, and he suggests that there may be others. From one point of view probabilities are defined in terms of long-run frequencies. Those who favor this frequentistic definition generally reject Bayes' statement as being self-evident or axiomatic on the grounds that the prior probability required in Bayes' theorem is not specifiable in terms of relative frequencies. R. A. Fisher, for example, relates that the advocates of inverse probability (Bayes' theorem) are ". . . forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes" (ref. 5, p. 6). An even more extreme position might be taken, namely that a priori probabilities are unknowable or do not exist. Regarding this view Uspensky observes, "To admit a belief in the existence of certain unknown numbers is common to all sciences where mathematical analysis is applied to the world of reality. If we are allowed to introduce the element of belief into such 'exact' sciences as astonomy and physics, it would be only fair to admit it in practical applications of probability" (ref. 6, p. 70).

Uspensky's comment leads us into consideration of "subjective" or "personal" probability, a notion which has provided impetus to a current reconsideration of Bayes' theorem. Recent interest in this interpretation of probability stems primarily from the work of Savage (ref. 7). According to this view, probability measures the confidence which a "reasonable" person has in the truth of some proposition. This view allows for probabilistic estimates to be applied to propositions about processes which are not repetitive and which, from a "frequentistic" point of view, seem unspecifiable in terms of probabilities. Advocates of personal probability are quick to point out, however, that personal probabilities measure consistent or orderly opinion (ref. 8). Good, for example, relates: "A subjective probability is a degree of belief that belongs to a body of beliefs from which the worst inconsistencies have been removed by means of detached judgments" (ref. 4, p. 446). Savage (ref. 9) and Edwards, et al. (ref. 8) have demonstrated

several types of meaningful propositions which are subject to personal but not frequentistic interpretations of probability. With respect to Bayes' theorem, advocates of personal probability are inclined to believe that prior probabilities, far from being meaningless or unspecifiable, can indeed be estimated by humans though not without considerable variability and vagueness (ref. 8). The extent to which vague and variable prior probabilities adversely affect estimations of posterior probabilities is a matter for some conjecture. One statement concerning the effects of prior probabilities upon posterior probabilities is the so-called "principle of stable estimation" (ref. 9). The essence of this argument seems to be that if the prior distribution of some hypothesis in question is not exceedingly divergent with respect to the conditional distribution of some datum given this hypothesis [P(D|H)], then the posterior probability [P(H|D)] can be reasonably approximated by K[P(D|H)], where K is the normalizing constant. To use Savage's expression, if the prior distribution behaves "gently" with respect to the P(D|H) distribution, then the influence of the prior distribution will be overcome by an accumulation of data. As far as human judgments are concerned, this means that although two individuals may hold initially divergent prior beliefs, they will, if they are open-minded, be forced into arbitrarily close agreement about future beliefs upon the accumulation of data (ref. 8).

If one can accept the notion of personal probabilities and the argument about prior opinions, then one can see that Bayes' theorem provides a formal statement about how one ought to revise these opinions in the light of new data (provided, of course, that expressions about data and hypotheses are cast in probabilistic terms). The usefulness of such a formal statement in the fields of medicine, business, and military affairs has been apparent for quite some time. In the present study the major concern is with applications of the Bayesian paradigm in a certain military context. Hopefully, however, the results will bear some relevance to other applications. In fact, as it will be mentioned in a subsequent section of this report, there is reason to believe that the results of the two experiments being described bear at least as much relevance to medical diagnosis as to the military context in which they were observed.

Edwards and Dodson have both recently illustrated how the Bayesian paradigm might be useful in military command and control systems where threat diagnoses or threat evaluations are performed (refs. 10, 11, 12). With respect to Bayes' theorem, Edwards retains the original statement while Dodson has presented a modified version which will be described in detail in section IV of this report. Edwards, however, has been considerably more explicit about the role of humans in threat-diagnosis systems. The usefulness of either approach hinges rather heavily on the ability of humans to estimate the conditional probabilities implied by the Bayesian paradigm, particularly in exceedingly complex situations. The present report describes two experiments, performed in a simulated threat-diagnosis situation, in which humans were required to estimate the conditional probabilities P(H|D) and P(D|H). The purpose of the experiments was to investigate human performance in estimating these conditional probabilities in a complex situation and to explore the implications and relevance of the results for the design of multiman-machine information-processing systems in which threat diagnosis or threat

evaluations are performed. More important perhaps, and as intimated at the outset of the report, one would hope that current research, stemming from a reconsideration of Bayes' theorem as a formal statement, will lead at least to some methodological innovations useful for psychology and will otherwise aid in the study of the logical structure of the task of revising one's opinions in the light of experience.

II. THE ROLE OF HUMANS IN THREAT-DIAGNOSIS SYSTEMS

Within the context of military affairs individuals are required to provide hypotheses or make diagnoses which will account for the occurrence of certain critical events in some hostile or potentially hostile environment. An intelligent judgment about the form and magnitude of the threat implied by these events is required before effective counteraction can be recommended. This "threat-diagnosis" activity is to be expected in any military agency in which the significance of incoming intelligence data is assessed. The command and control facilities of the North American Air Defense Command (NORAD) provide a specific example of the locus of such activity. In an age of ever-increasing sophistication of weapon systems, when available response times to hostile action are measured in minutes and seconds, the consequences of incorrect diagnoses of environmental events are frightening to contemplate. The demands placed upon the humans who must provide these diagnoses are indeed formidable. Specifically, there may be a wide range of potential explanatory hypotheses; the volume of intelligence data (often of a diverse character) may be enormous; these data will always be fallible to an unknown degree and often contradictory; and the cause and effect relationships between data and hypotheses may be exceedingly abstruse or unknown. It has become increasingly apparent that humans who provide threat diagnoses in these complex circumstances have need of assistance.

Edwards has recently described how Bayes' theorem might be put to good use in the design of threat-diagnosis systems (refs. 11, 12). Specifically, he advocates a system design in which information is processed probabilistically and in which certain portions of the threat-evaluation task are automated. The output of such a system would be posterior probability estimates or estimates of the probability that various alternative hypotheses of interest to the system account for the occurrence of whatever intelligence data the system possesses. From Bayes' theorem these posterior probability estimates follow upon specification of $P(H)$ and $P(D|H)$. Using the argument described in the preceding section of this report, Edwards contends that the initial or prior probabilities [$P(H)$] which the system entertains at the beginning of some period can be arbitrarily chosen (as long as they are not too close to 0 or 1) and will soon be overcome by the mass of incoming data. He maintains, therefore, that the only information needed for calculation of posterior probabilities (which, of course, would be done on high-speed computers) are $P(D|H)$ values, since the $P(H|D)$ values calculated on cycle n-1 become the prior probabilities used for calculation on cycle n. $P(D|H)$ would, of course, be estimated by experts, presumably with considerable past experience with the data in question. If the arguments about Bayes' theorem have merit and if humans can make reasonable estimates of $P(D|H)$ or some other quantity from which $P(D|H)$ can be recovered, then one faces the intriguing possibilities of automated diagnosis or hypothesis selection since the assignment of posterior probabilities across the hypothesis set is merely an arithmetic task, one superbly and almost instantaneously performed by a computer.

III. CONDITIONAL PROBABILITY ESTIMATION BY HUMANS: A REVIEW OF RECENT RESEARCH

Edwards' proposal has succeeded in generating a growing amount of research on human commerce with the probabilistic information implied by Bayes' theorem. For the most part, this research has involved comparisons, under several different circumstances, of human and various automated estimations of posterior probabilities. At present, the proposition that posterior probability estimation ought to be automated in diagnostic systems rests upon empirical evidence which although suggestive, is not compelling. It is possible to group the various studies into two classes according to the experimental disposition of the environmental contingency rules or P(D|H).

A. P(H|D) Estimation by Humans, P(D|H) Given

Two experiments using fairly complex stimulus environments and one fairly simple experiment are representative of this general condition. In an unpublished experiment Hays, Phillips, and Edwards (described by Edwards and Phillips, ref. 13) displayed prior probabilities and P(D|H) values directly to their subjects. They found that the subjects were suboptimal estimators of P(H|D) upon comparing these estimates with calculations of P(H|D) using Bayes' theorem. In the first experiment performed using the Ohio State University multiman-machine system simulator (ref. 14), the P(H|D) estimates provided by the primary decision maker on an eight-man threat-evaluation team were compared with P(H|D) estimates calculated on the basis of Dodson's modification of Bayes' theorem (see section IV). (This modification hereafter will be termed the "MBT.") The computer-implemented MBT solutions were significantly superior to the human's estimates over the course of the experiment. In a very simple experimental setting described by Edwards (ref. 15) subjects were asked to estimate the posterior probability that sequences of red and blue poker chips were being drawn from either of two bookbags (a 70% red-30% blue bag or a 70% blue-30% red bag). These estimates, compared with those calculated from Bayes' theorem, were uniformly conservative.

With respect to the first two experiments mentioned above, there is very little assurance that the subjects in either experiment were in any position, by virtue of training or experience, to use the displayed P(D|H) values to best advantage. Hays, Edwards, and Phillips gave no instructions about how to use these P(D|H) values except "the obvious qualitative statements"; e.g., if a datum is highly likely under hypothesis A and unlikely under hypothesis B, and the datum occurs, it favors hypothesis A (ref. 13). In the OSU experiment great pains were taken to present the P(D|H) values to the subject in a form which was thought to be most meaningful to him. In fact, verbal abstractions of the required 20 x 103 P(D|H) matrix were provided in the hopes of increasing its meaning. In both experiments, however, one has very little idea of how well the subjects were able to utilize this P(D|H) information. If they disregarded this information because it was unfamiliar or meaningless, one might argue that the superiority of the mathematical solutions would be obvious under these conditions and that very little has been shown in these two experimental comparisons between human and automated P(H|D) estimations.

B. P(H|D) Estimated by Humans, P(D|H) Unknown

In one condition of the recent Kaplan and Newman study (ref. 16), naive subjects were asked to estimate P(H|D) directly from data. This condition was called the "non-PIP" condition (PIP meaning Probabilistic Information Processing). In the non-PIP condition P(D|H) was neither explicitly estimated by the subjects nor displayed to them. Human P(H|D) estimates in the non-PIP condition were compared with computer-implemented Bayesian solutions of P(H|D) based upon P(D|H) estimations made by the subjects in the second condition of the experiment which was called the "PIP" condition. Although the highest probability was always assigned to the correct hypothesis in both PIP and non-PIP conditions, the PIP condition resulted in significantly higher P(H|D) values.

One might criticize the manner in which the sequential nature of the task in the non-PIP condition was interpreted to the subjects. Presumably, by a careful analysis of human P(H|D) estimates, one would hope to be able to show the extent to which humans are adept at revising opinions on the basis of experience. The sequential and cumulative aspects of this process are apparent. Yet in the experiment being cited, the only instructions given to the subjects with regard to the sequential acquisition of experience in the non-PIP condition was the meager statement: "If you wish, you may turn back the pages and see what responses you have given before" (ref. 16, Appendix A, p. 6). No mention is made of observing the data in any cumulative or sequential fashion which an intelligent judgment of P(H|D) would require.

In the second experiment performed by Southard, Schum, and Briggs (ref. 17) the posterior probability estimation performance of the primary decision maker on an eight-man threat-evaluation team was evaluated as this individual was given an increased amount of control over a Bayesian hypothesis-selection aid. Although this individual's performance was independent of the aid configuration, his accuracy in estimating P(H|D) was very nearly identical to the accuracy of MBT solutions of P(H|D) calculated on the basis of the same data used by the human estimator. In addition, with an increasing number of trials (there were 240 in all), the human's P(H|D) estimates showed slight superiority over those calculated by means of the MBT. One very interesting result was the human's shift from conservative P(H|D) estimates to more definite commitment-type (very high) estimates as the experiment progressed (ref. 17, see figure 6).

With one exception the results of the experiments cited above indicate that human estimations of P(H|D) are inferior to solutions of P(H|D) based upon Bayes' theorem. Without exception, all of these studies are introductory in character and the results, therefore, should be interpreted with some caution. Edwards' general conclusion is that humans fail to extract all the consistency or certainty that exists in probabilistic information (ref. 15). This conclusion is indicated by the conservative estimates that the humans typically produced (compared with Bayesian estimates) in all experiments except the one in which conservatism diminished with experience. However, one may well question whether or not the design and conduct of the various experiments cited above actually allowed or encouraged the humans to perform at a higher level. With more adequate instructions and better procedures

gaining experience in tasks involving probabilistic information. the humans could perhaps have been induced to perform even more creditably than they actually did. There is further danger in unrestrained generalization on the basis of these early studies. In all studies, with the possible exception of the second study by Southard, Schum, and Briggs (ref. 17), naive subjects performed briefly in unfamiliar circumstances. Presumably, however, persons in real-life situations performing crucial hypothesis-selection tasks are exceedingly competent people with considerable experience in dealing with data relevant to the situation.

IV. A SIMULATED THREAT-DIAGNOSIS SYSTEM

The two experiments being reported were performed using the multiman-machine system simulator at the Ohio State University Laboratory of Aviation Psychology. The major hardware components of this simulation facility include an IBM 7094-1401 computer system, digital display interface equipment, and phone and closed-circuit television facilities for system operator intercommunication. The computer facilities provided the means for generating a complex real-time stimulus environment or data base simulating the movements of the surface and air forces of a hypothetical adversary called "Aggressor." In both experiments a subject-operator team attempted to evaluate or diagnose the threat posed by deployments of Aggressor's forces on the basis of intelligence information obtained on simulated reconnaissance overflights of Aggressor's territory. Since both experiments were performed using the same simulation facilities, there were many features common to both experiments. The purpose of this section of the report is to describe the stimulus environment or data base, the tasks for members of the threat-evaluation team, the types of human and automated performance being evaluated and compared, and the measures of human and automated performance. All of these features are common to both experiments and will be described in enough detail to render the two specific experiments and their results intelligible to the reader.

A. Stimulus Environment Characteristics

The stimulus environment or data base refers to the characteristics of the scenario presented to the threat-evaluation team. This scenario, as mentioned above, represented the maneuvers or deployments of the surface and air forces of the hypothetical aggressor whose activities were under surveillance. The precise features of the stimulus environment and the procedures for generating events to be portrayed in this environment have been described in considerable detail in a previous report (ref. 14). Briefly, however, Aggressor territory was defined as a square area 1024 miles on each side. Within this territory there occurred orderly buildups or deployments of Aggressor's forces which were called developmental groupings. The computer facilities allowed the experimenter to depict the buildup of these developmental groupings in a time-dependent fashion. In the experiments being reported there could be as many as 25 individual developmental groupings in various stages of buildup in Aggressor territory at any one time. Figure 1, a highly simplified abstraction of the basic scenario, should help to make this buildup process clear. The large triangles in the figure represent certain fixed installations such as forts, supply depots, and airfields. The clusters of small circles represent units of Aggressor's surface and air forces. The term "developmental grouping" refers to the clusters of these units which, in figure 1, are shown in various stages of buildup. Each developmental grouping was given a unique number (e.g., No. 101) which allowed mutual identification of the grouping by the experimental subjects on the threat-evaluation team and the experimenter. Observe in figure 1 that there are developmental groupings in various stages of buildup. A developmental grouping reached the terminal stages of buildup when all of its associated units had reached their final locations along one of the four borders, e.g., No. 101 and No. 103 in figure 1 have terminated. When all

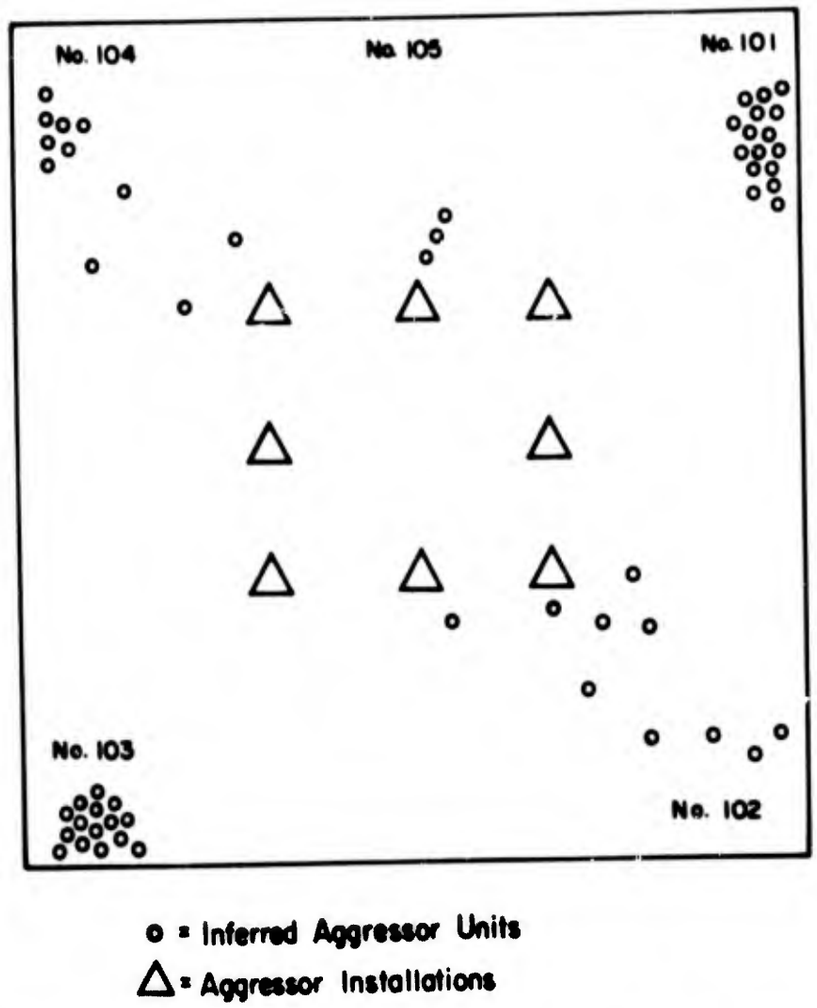


Figure 1. Aggressor Territory (Highly Simplified).

the units in a developmental grouping had reached their terminal locations and had remained there for a set period of time, they disappeared from the environment. No. 102, No. 104, and No. 105 illustrate developmental groupings in various intermediate stages of buildup as their units are still moving from the centrally located fixed installations out to terminal peripheral locations. System "load" was defined as the number of developmental groupings terminating in an experimental session. In the two experiments being reported six developmental groupings terminated each session. Certain members of the threat-evaluation team determined the existence and characteristics of these developmental groupings by a computer interrogation procedure described below as the system's task is discussed.

B. A Modification of Bayes' Theorem

Before describing the precise characteristics and significance of these developmental groupings, it will be wise to proceed with a discussion of Dodson's modification of Bayes' theorem at this point since Dodson's formulation has influenced the design of the stimulus environment to a very great extent. Dodson's modification of Bayes' theorem (MBT) provides for

generalization to two types of situations not accounted for in the original statement (ref. 10). First, Bayes' theorem assumes only two possible states of an event or datum, occurrence or nonoccurrence. The MBT, however, is applicable to situations in which an event or datum may exist in any number of states. Also, Bayes' theorem assumes complete knowledge about which of the two states of a datum (occurrence or nonoccurrence) has been observed. Dodson's formulation, on the other hand, allows for observational uncertainty about which state within a datum class has in fact occurred. This observational uncertainty can be indicated by assigning probabilistic estimates that the datum is in each of the possible states.

Equation 2 is Dodson's MBT with notational modifications introduced by the authors in an effort to preserve a single notational system throughout the report.

$$P(H_i|D) = \sum_{k=1}^{\mu} P(D_k) \left[\frac{P(H_i)\ P(D_k|H_i)}{\sum_{i=1}^{n} P(H_i)\ P(D_k|H_i)} \right] \qquad \text{(Eq. 2)}$$

This equation applies to the evaluation of the probabilistic responses given to the μ states of one event or data set. $P(H_i|D)$ is the a posteriori probability that hypothesis i is true given the probabilistic estimates that the various μ states in the data class have been observed. In the bracketed expression $P(H_i)$ is the a priori probability of hypothesis i and $P(D_k|H_i)$ is the conditional probability that the data class in question will be observed in state k if hypothesis i is true. The denominator is a normalizing constant which assures that the a posteriori probabilities sum to 1.00 across the n hypotheses. $P(D_k)$ is the probability that the k^{th} state of the data class is the state being observed. Note that $\sum_{k=1}^{\mu} P(D_k) = 1.00$. For each data class there are μ states or conditions where μ varies according to the data class being considered. There may be observations in many different data classes to be evaluated by means of the MBT. Therefore, the $P(H_i|D)$ calculated by means of the observations and conditional probabilities for one data class become the a priori probabilities $P(H_i)$ used in the calculation of $P(H_i|D)$ for the next data class and so on until all data observations have been evaluated. The assumption of data class independence is necessary since these MBT calculations are sequential combinations of probabilistic estimates across the different data classes. The results of these calculations may be misleading depending upon the form and the extent of data class dependencies. In the present experiments the independence assumption is justified since the experimenter's procedure for selecting the "true" levels in each data class for every developmental grouping assured data class independence. In order to generate the characteristics of each developmental grouping, the experimenter selected one level from each of the 25 data classes at random according to the previously specified "true" $P(D_{jk}|H_i)$ distributions in each data class.

A description of the stimulus environment features which match the requirements of a Bayesian diagnostic paradigm can now be presented. First, in the two experiments being reported Aggressor was allowed eight mutually exclusive and exhaustive response alternatives. These response alternatives

were in fact "hypotheses" from the point of view of the threat-evaluation team and were used to account for the occurrence of the intelligence data obtained with respect to the developmental groupings. The various response alternatives can also be considered as possible strategies to be employed by Aggressor in initiating some hostile action across the borders of his homeland area. These hypotheses or response alternatives were simply labeled alphabetically A through H. This arbitrary labeling scheme was initiated because of a methodological necessity discussed in detail in a previous report (ref. 17). The procedure seems to have had little or no effect on the subjects' performance. With experience in dealing with environmental events the subjects learned rather quickly what hypothesis A "meant," B "meant," and so on. The second environmental feature concerns the multiple event or data classes. The attributes of each developmental grouping could be described in terms of 25 different types of information or data. These 25 attribute data classes are listed in Appendix I of this report. In general, they referred to such developmental grouping features as (a) infantry-armored constituency, (b) artillery, missile, rocket, and air support, (c) logistics support, and (d) spatial and temporal arrangement of forces (order of battle). As Appendix I also illustrates, each of the 25 attribute data classes had between two and eight possible states or conditions. This is one particular feature of the stimulus environment which matches Dodson's paradigm. For example, data class X (Tactical Air Support Squadrons) has four possible states, i.e., in any developmental grouping there could be either 0, 1, 2, or 3 tactical air support squadrons. Only one state of each of the 25 data classes was "correct" for each developmental grouping.

In summary, the stimulus environment consisted of the homeland area of a hypothetical adversary called Aggressor. The Aggressor activities under surveillance consisted of buildups of surface and air forces in this homeland area. These buildups were termed "developmental groupings" and could be described in terms of 25 attributes or dimensions. There could be as many as 25 groupings in various stages of buildup at any one time although only six terminated in each experimental session.

C. Tasks Performed by Members of the Threat-Evaluation Team

The purpose of this section is to describe how members of the threat-evaluation team provided answers to the following questions: (a) what events are taking place in Aggressor's territory, and (b) what is the significance of these events, or more specifically, what is Aggressor's intention with respect to each developmental grouping? Figure 2 illustrates the flow of information in the simulated system and describes the two basic task levels at which the answers to the above questions were sought. The first task level concerns the intelligence staff officers (ISOs). These individuals attempted to provide answers to the first question indicated above. It was the task of this portion of the system to locate the various Aggressor developmental groupings and to describe their characteristics in terms of the 25 attributes listed in Appendix I. For each developmental grouping under surveillance the ISOs produced probabilistic estimates that the various levels or states in each attribute data class were in fact being observed with respect to that grouping. The ISOs received information about events in Aggressor territory only after they had initiated a two-stage information retrieval sequence. The first phase was initiated by the "Chief of Staff" who specified an area

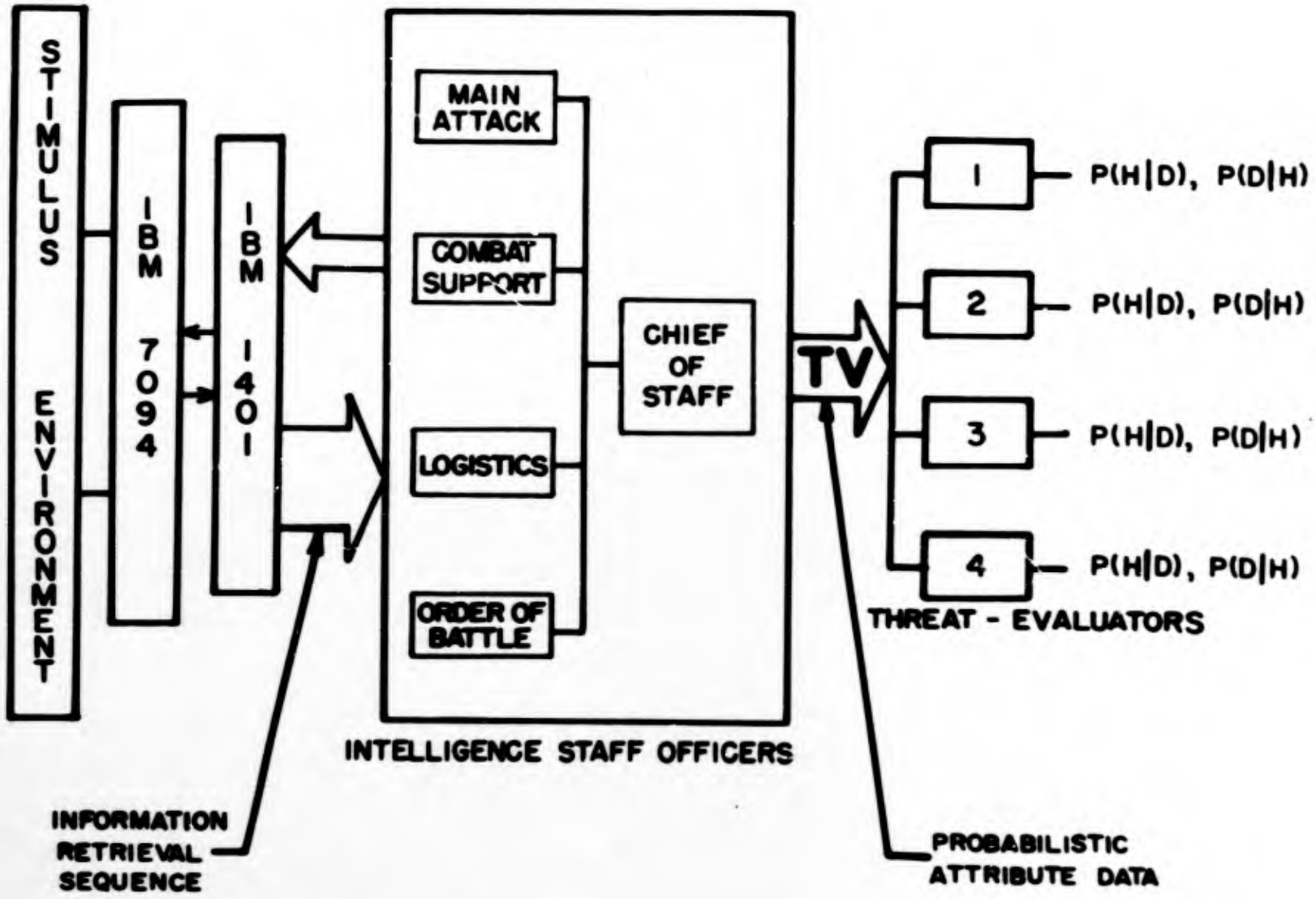


Figure 2. Simulated Threat-Diagnosis System.

of Aggressor territory over which simulated reconnaissance missions were to be flown. The computer facilities simulated these overflights and also simulated the activities of the many individuals who interpreted and catalogued the photographic, radar, and infra-red sensor data thus obtained on the overflights. After an "overflight" has been completed and the resulting data "processed" and "catalogued," the ISO team could proceed with the second phase of the retrieval sequence. In this second phase the ISOs could make direct interrogations of computer storage by means of their digital display consoles and ask specific questions about events in the territory which had been put under surveillance. These questions were directly related to the area of responsibility which the ISO had assumed. Observe in figure 2 that there are four ISOs who, in effect, served as "content experts." Each of these individuals was responsible for developing a subset of the attribute data classes indicated in Appendix I. For example, the "Main Attack" ISO was responsible for developing information about attribute data classes I, II, and III in Appendix I for each developmental grouping. The information which the ISOs received on their digital display consoles was in the form of verbal and numerical descriptions of the type, number, activity, and location of collections of mobile weapons, vehicles, and aircraft (collectively termed "elements"). Using tabled reference sources which related numbers and types of elements to Aggressor units, they attempted to infer the existence of the various Aggressor units in a developmental grouping and the spatial and temporal arrangements of these inferred units. From the information obtained using the procedure indicated above, the ISOs could provide estimates of the probability that the various states or levels of each of the 25 attribute data classes had been observed with respect to a particular developmental grouping. Table 1 should help to make clear the exact form of the ISO responses. Table 1 illustrates the form of the probabilistic responses made by the ISO responsible for data classes X and XI on the basis of information retrieved with respect to developmental grouping No. 101. He is confident, for example, at .70 that there are two tactical air squadrons in this developmental grouping but he allows for possible underestimation or overestimation. Such responses were provided by the ISOs for each of the 25 attribute data classes for every developmental grouping under surveillance. Since the events in the stimulus environment were time-dependent, information became obsolete

TABLE 1

AN ILLUSTRATION OF AN INTELLIGENCE STAFF OFFICER RESPONSE

	State or Level			
	0	1	2	3
X. Tactical Air Support Squadrons		.10	.70	.20
XI. Aerial Reconnaissance Squadrons		.90	.10	

and the two-phase retrieval sequence was initiated many times during an experimental session. This, of course, caused the ISOs to revise their data state estimates quite frequently. The final estimates which they made during the terminal stages of buildup of a developmental grouping were collected and relayed to the next task level in the system (the threat evaluators, TEs, shown in figure 2). Thus, the output of the ISO level was a set of probabilistic estimates of the state or condition of each of the 25 attribute data classes for each terminating developmental grouping.

The second task level shown in figure 2 consisted of those individuals whose task was to evaluate the threat posed by each developmental grouping and thus provide answers to the second of the two questions posed above. Each of the TEs produced estimates of $P(H|D)$ and $P(D|H)$ on the basis of the probabilistic attribute data relayed to them from the ISO level. Since the $P(H|D)$ and $P(D|H)$ estimates represented the most important behavior in the present study, the use of several subjects to provide these estimates was an experimental necessity. Equipment limitations precluded the use of more than four subjects in the TE role. It must be emphasized that each TE produced his own estimates of these conditional probabilities. In fact, TE performance was carefully monitored to assure the experimenter of independent estimates. The attribute data inputs to each of the TEs were identical.

As previously mentioned, six Aggressor developmental groupings terminated each day. When it was apparent that a grouping had terminated its buildup, the ISO team relayed its final attribute data estimates for that grouping to the TE level by means of closed-circuit television. On the basis of these data each TE estimated the probability that each of the eight Aggressor response alternatives or strategies could have accounted for the occurrence of these attribute data [$P(H|D)$]. More appropriately this term should be symbolized as $P(H|D_I, D_{II}, \ldots, D_{XXV})$ since there were, in fact, 25 data state estimates presented simultaneously. For convenience however, the term $P(H|D)$ will be used with the understanding that D stands for the entire "package" of data state estimates for a developmental grouping. Each TE therefore provided six $P(H|D)$ estimates during each experimental session. The TEs were required to normalize their $P(H|D)$ estimates for each grouping, i.e., their estimates were required to have unit sum across the eight hypotheses. At the beginning of each session (with the exception of the first session) the TEs were informed about the six "true" hypotheses applicable to the data seen in the previous session.

On the basis of these verified data-hypotheses relationships provided at the start of each session and applicable to the six groupings which had terminated on the previous day and on the basis of similar accumulated verified data-hypotheses relationships established in earlier sessions, each TE estimated $P(D|H)$ for every state of each of the 25 data classes given each of the 8 hypotheses. More appropriately, these estimates should be symbolized as $P(D_{jk}|H_i)$ since a TE was actually estimating the probability that the jth level of data class k would be observed if hypothesis i were true. This involved generation of an 8 x 103 $P(D_{jk}|H_i)$ matrix since there were 8 hypotheses and 103 possible data states across all data classes. This $P(D_{jk}|H_i)$ matrix was generated by each TE once during each experimental session. These $P(D_{jk}|H_i)$ estimates served two purposes: (a) they were used as inputs to the MBT in order to provide one class of automated estimates of $P(H|D)$ for

comparison with the TEs' estimates of $P(H|D)$, and (b) they provided each TE with a method for summarizing and accumulating the past experience about events in the stimulus environment necessary in order for the TE to make his own estimates of $P(H|D)$.

D. Performance Comparisons

The primary purpose of both experiments being reported was to compare three types of $P(H|D)$ estimates: (a) unaided human estimates of $P(H|D)$, (b) MBT solutions of $P(H|D)$ on the basis of human estimates of $P(D|H)$, and (c) MBT solutions of $P(H|D)$ calculated on the basis of Dodson's "self-adapting" version of the MBT (ref. 10). Although the means for acquiring the first two estimates were discussed in the preceding section, an additional fact about these two estimates must be made apparent. At the beginning of the first experiment and at the beginning of each of three conditions in the second experiment the four TEs were told to assume that the prior probabilities of each of the hypotheses were equal (i.e., .125 since there were eight hypotheses). With respect to the MBT solutions of $P(H|D)$ using the TE estimates of $P(D|H)$, the prior probability term $[P(H_i)]$ in equation 2 was set at .125 for each calculation. There were two reasons why the prior probabilities were treated in this manner. First, the effects of varying prior probabilities upon subsequent $P(H|D)$ estimates is an experimental issue in its own right. At this juncture in the research series the subjects were each told to assume equal prior probabilities in order to reduce the complexity of an already difficult experimental situation. Second, the events in the stimulus environment were designed so that in each experimental condition each hypothesis was "true" an equal number of times. Therefore, setting $P(H_i) = .125$ in each MBT solution of $P(H|D)$ using the TE estimates of $P(D|H)$ was judged to be minimally damaging to the accuracy of these solutions.

The third solution or estimate of $P(H|D)$ mentioned above was based upon Dodson's suggestion about how the MBT could be made to adapt itself to changing environmental events. This adaptation feature has as its basis parameters which regulate the rate at which information obsolesces and an expression for feedback about the true state of affairs existing in the environment at the time each observation was made. These parameters and the expression describing feedback are applied by Dodson to the $P(H_i)$ and $P(D|H_i)$ terms in equation 2. The parameters and the feedback expression describe how these terms are to be updated on every trial or observation cycle in a sequence. In the experiments being reported the a priori probability was set at .125 in all calculations provided from this self-adapting MBT as well as from the other MBT solution described above. At this juncture in the experimental series the experimenter's concern was limited to the possibility of control over the adaptation process strictly with respect to the contingency relationships between each level of every data class and each hypothesis $[P(D_{jk}|H_i)]$. Equation 3, with notational modification, is Dodson's expression for $P(D_{jk}|H_i)$ illustrating the adaptive or "learning" features allowed by the parameters.

$$P(D_{jk}|H_i)_v = \frac{P(D_{jk}|H_i)_{v-1} + K_v \left[P(D_{jk})_v \, P(F_i)_v \, w_{iv}\right]}{1 + K_v \left[P(F_i)_v \, w_{iv}\right]} \qquad \text{(Eq. 3)}$$

where

$P(D_{jk}|H_i)$ = the conditional probability of the k^{th} state or condition of data class j given the occurrence of hypothesis H_i.

v = a particular observational trial or cycle number. (v-1) refers to the preceding observational trial or cycle.

$P(D_{jk})_v$ = the probability that the k^{th} level of data class j has been observed in cycle v. (The values were provided by the ISOs upon observations made of the stimulus environment.)

$P(F_i)_v$ = the probability that H_i is to be associated with the input pattern of $P(D_{jk})$ in cycle v. This term is essentially the feedback from the environment as to what actually happened in association with the attribute data for a particular grouping. The term applies to the strategy (H_i) Aggressor actually used in cycle v. $P(F_i)_v$ assumes only two values, 0.0 or 1.0. If $P(F_i)_v = 0.0$, then H_i was not Aggressor's strategy in cycle v; if $P(F_i)_v = 1.0$, then H_i was Aggressor's strategy in cycle v.

K_v = the parameter which regulates the extent to which $P(D_{jk})_v$ and $P(F_i)_v$ are allowed to modify all condition probabilities. K_v can assume any value in the range $0 \leq K_v \leq \infty$. When $K_v = 0$, no adjustment of the preceding conditional probability (on the v-1th cycle) in made, i.e., equation 2 reduced to:

$$P(D_{jk}|H_i)_v = P(D_{jk}|H_i)_{v-1}$$

As K_v approaches infinity, $P(D_{jk}|H_i)_v$ approaches $P(D_{jk})_v$.[1] This means that $P(D_{jk}|H_i)$ on v^{th} cycle is entirely determined by the most recent estimation of $P(D_{jk})$. Fairly

---

[1] Let $P(D_{jk}|H_i)_{v-1}$, $[P(D_{jk})_v\ P(F_i)_v\ w_{iv}]$, and $[P(F_i)_v\ w_{iv}]$ be constants in any cycle v; call them C_1, C_2, and C_3, respectively. Then:

$$\lim_{K_v \to \infty} \frac{C_1 + K_v C_2}{1 + K_v C_3} = \lim_{K_v \to \infty} \frac{\frac{(C_1 + K_v C_2)}{K_v}}{\frac{(1 + K_v C_3)}{K_v}} = \lim_{K_v \to \infty} \frac{\frac{C_1}{K_v} + C_2}{\frac{1}{K_v} + C_3} = \frac{C_2}{C_3}$$

$$= \frac{P(D_{jk})_v\ P(F_i)_v\ w_{iv}}{P(F_i)_v\ w_{iv}} = P(D_{jk})_v$$

large changes of $P(D_{jk}|H_i)_v$ in the direction of most recent estimations of $P(D_{jk})$ can be made with fairly small values of K_v. In the present experiments K_v was set at .5. This value introduced moderate rather than drastic changes toward the most recent estimates of $P(D_{jk})$.

w_{iv} = a parameter which regulated the extent to which the input sets of $P(D_{jk})_v$ will be associated with a specific H_i. In terms of the present experiment w_{iv} represents the extent to which the conditional probabilities associated with any Aggressor response alternative (or strategy) determined by previous data are modified by current data. In effect, w_{iv} is a vernier weight which allows one to control differentially the adjustments of conditional probabilities for each of the hypotheses. K_v, on the other hand, can be considered a more gross weight affecting all conditional probabilities across all of the hypotheses. In the present experiments w_{iv} was defined more precisely as:

$$w_{iv} = [P(H_i|D)_v - P(F_i)_v]^2 \qquad \text{(Eq. 4)}$$

where

$P(H_i|D)_v$ = the self-adapting MBT calculation in cycle v.

$P(F_i)_v$ = environmental feedback as to the correct hypothesis in cycle v. Recall that $P(F_i)_v$ assumed only two values, 1.0 if H_i was true and 0.0 if H_i was not true.

This definition of w_{iv} says, in effect, that the closer the calculated $P(H_i|D)$ was to being correct the smaller will be the change induced in the values of $P(D_{jk}|H_i)$ on the following cycle. Thus, large changes will not be induced when $P(D_{jk}|H_i)$ values are relatively accurate. Equation 4 shows why values of w_{iv} were limited to the range $0 \leq w_{iv} \leq 1.0$. Note, by observing equation 3, that a change in $P(D_{jk}|H_i)$ values could occur only in the hypothesis category correct in that cycle, i.e., where $P(F_i)_v = 1.0$. When $P(F_i)_v = 0.0$, the right hand side of equation 3 reduced to $P(D_{jk}|H_i)_{v-1}$.

Equations 3 and 4 thus specify how the $P(D_k|H_i)$ term in equation 2 is to be automatically adjusted on the basis of experience. Since $P(H_i)$ was constant throughout the experiments, automatic solutions of $P(H|D)$ for each developmental grouping could be obtained from the self-adapting MBT upon entering into the computer the $P(D_{jk})$ estimates (probabilistic attribute data state estimates) provided by the ISOs for each developmental grouping.

E. Performance Measures

In both experiments two types of measures were taken with respect to the human and automated estimates of $P(H|D)$. The first measure, called "verified certainty," was simply the value of $P(H|D)$ in the correct hypothesis category. Table 2 should help to make this measure (and the one which follows) clear.

TABLE 2

AN ILLUSTRATION OF P(H|D) RESPONSES

		Hypotheses							
		A	B	C	D	E	F	G	H
No. 101	Human	.10	.70	.10	.05	.05			
	MBT	.08	.60	.18	.04	.04	.02	.02	.02
No. 102	Human	.50	.40	.05	.05				
	MBT	.20	.30	.20	.10	.10	.05	.03	.02

Assume in both examples (developmental grouping Nos. 101 and 102) that B is the true hypothesis. The verified certainty scores for the human are .70 and .40, for the MBT .60 and .30. It is also of interest to observe the precise number of occasions on which either the human or automated first choice or highest P(H|D) estimates were correct. For this reason a second method of scoring was introduced. These scores, called "dichotomous scores," indicate the number of occasions on which the highest or first-choice P(H|D) estimate was placed in the true hypothesis category. In table 2 observe that for No. 101 both human and MBT first-choice estimates are correct since hypothesis B was known to have been true. In No. 102, however, the MBT placed its highest P(H|D) value in the correct category while the human did not. This latter example illustrates how different interpretations of the relative performance of human and MBT can arise depending upon the scoring procedure one uses. Although the human's verified certainty score for No. 102 was higher than the MBT score, the MBT placed its highest estimate in the correct category and the human did not.

Verified certainty scores were also used to evaluate the performance of the ISOs in estimating the probability of the various states within a data class. Verified certainty, in this instance, was simply the value of $P(D_{jk})$ placed in the true data class state for the particular developmental grouping.

For the P(D|H) estimates produced by the TEs a somewhat different scoring procedure was used. The P(D|H) score was called an "agreement score" (α_j) and was defined as follows:

$$\alpha_j = 1 - \frac{\sum_{k=1}^{\mu} |d_k|}{2} \qquad 0 \leq \alpha_j \leq 1.0 \qquad \text{(Eq. 5)}$$

where

μ = the number of states or conditions in data class j.

k = the k^{th} state or condition in data class j.

$d_k = [P(D_{jk}|H_i)_{estimates} - P(D_{jk}|H_i)_{true}]$

i = the i^{th} hypothesis category

There was one α_j score for each data class under each of the eight hypothesis categories. Perfect agreement between the estimated conditional probabilities assigned across a data class given some hypothesis and the true conditional probabilities for this class and hypothesis yielded a score of $\alpha_j = 1.0$. Complete lack of agreement yielded $\alpha_j = 0.0$.

V. EXPERIMENT I: HUMAN ESTIMATIONS OF POSTERIOR PROBABILITIES OVER AN EXTENDED SERIES OF TRIALS

A. Objectives

The general character of the first experiment was exceedingly simple. Two questions were asked: (a) how does increased experience affect human performance in the task of estimating $P(H|D)$, and (b) to what degree will human $P(H|D)$ estimates match those provided by both the MBT using human-estimated $P(D|H)$ and the self-adapting MBT? In the writers' opinion, not all of the previous studies reviewed in section III of this report provided sufficient instructions, time, or procedures to allow the subjects to make more correct and more confident estimates of $P(H|D)$. There was an indication in the second experiment performed by Southard, Schum, and Briggs (ref. 17) that conservatism in estimating $P(H|D)$ diminishes with experience. This result, however, needed confirmation since data were collected from only one subject. From one point of view, an experiment devoted solely to the effects of experience upon human performance may seem rather trivial and improvement in performance with experience is the sort of thing one would naturally expect. Unfortunately, most studies so far offer little notion about what to expect from persons who are experienced in dealing with events in some environment of concern. The allegation is simply that humans are conservative or suboptimal estimators of posterior probabilities and that they do not extract maximum certainty from probabilistic information. The purpose of the present experiment was to observe the extent to which this statement will have to be qualified when humans are given the opportunity to become familiar with environmental events and proficient in dealing with probabilistic statements describing these events.

B. Experimental Procedures

The present experiment consisted of 30 consecutive 4-hour sessions. In each session six Aggressor developmental groupings terminated. On the basis of the probabilistic attribute data provided by the ISOs for each grouping, each TE produced his estimates of $P(H|D)$. The attribute data were presented to each TE individually and simultaneously by means of closed-circuit television. In addition to these televised data, each TE was also able to follow the buildup of each grouping in Aggressor territory. The ISOs maintained an edge-lighted plexiglas display board upon which all Aggressor activity under surveillance was continuously posted. A television camera was focused on this board and a continuous TV picture of it was available to each TE. Since there were six developmental groupings terminating in every session, the TEs each produced 180 $P(H|D)$ estimates throughout the experiment. In each session the TEs also generated their estimates for the required 8×103 $P(D_{jk}|H_i)$ matrix. TE performance was carefully monitored to assure independence of effort.

For each of the 180 developmental groupings terminating during the course of the experiment the experimenter obtained:

(1) Four unaided human estimates of $P(H|D)$.

(2) Four MBT solutions of P(H|D) calculated individually on the basis of each TE's P(D|H) estimates.

(3) One self-adapting MBT solution of P(H|D).

Various types of knowledge of results were given to the ISOs and TEs. Complete feedback with respect to the correctness of the environmental data produced by the ISOs was judged to be unrealistic. For this reason, verification of the true state of only five of the 25 classes of attribute data (for each of the six terminating developmental groupings) was provided for the ISO team on the session following the termination of the groupings. These five classes were chosen at random. Each TE was also given this attribute data feedback. Each TE was also provided each session with the six correct hypotheses explaining the occurrence of the data seen in the previous session. One assumption, therefore, was that previous courses of action actually taken by Aggressor could always be recognized.

The two MBT solutions were also made available to the TEs. At the start of each session each TE was able to compare his unaided P(H|D) estimates made in the previous session with the MBT P(H|D) solutions based upon his P(D|H) estimates from the previous session and with the self-adapting MBT solutions. The TEs did not see each others' estimates. Finally, feedback about the correct P(D|H) estimates was not provided for the TEs because it seems highly unrealistic to assume that the true environmental contingency rules could ever be known by the threat-evaluation team.

C. Subjects, Training, and Instructions

All of the subjects in the experiment were either upperclass undergraduates or students in the graduate or various professional schools at Ohio State University. They were volunteers and were paid at rates determined by length of service and amount of responsibility assumed in the threat-evaluation system. All of the subjects had served in at least one of two previous experiments and all had received the extensive 11h-hour training program discussed in detail in a previous report (ref. 14). The ISOs in the present experiment were already well trained in their respective tasks. The TEs, however, were given further training consisting of lectures and practice problems involving P(H|D) and P(D|H) estimation. The practice problems were similar in content and difficulty to those actually encountered in the experiment. This training program consisted of five sessions each of 4-hour duration.

The TEs were instructed to produce their most accurate estimates of P(H|D) and P(D|H). Speed of response was not emphasized in this experiment.

D. Results

1. Human and automated posterior probability estimation as represented by verified certainty scores: As mentioned previously with respect to posterior probability estimates, verified certainty scores simply indicated the value of the estimate placed in the correct hypothesis category. Figure 3, therefore, illustrates the change, as environmental experience increases, in the size of the certainty estimates placed in correct hypothesis categories

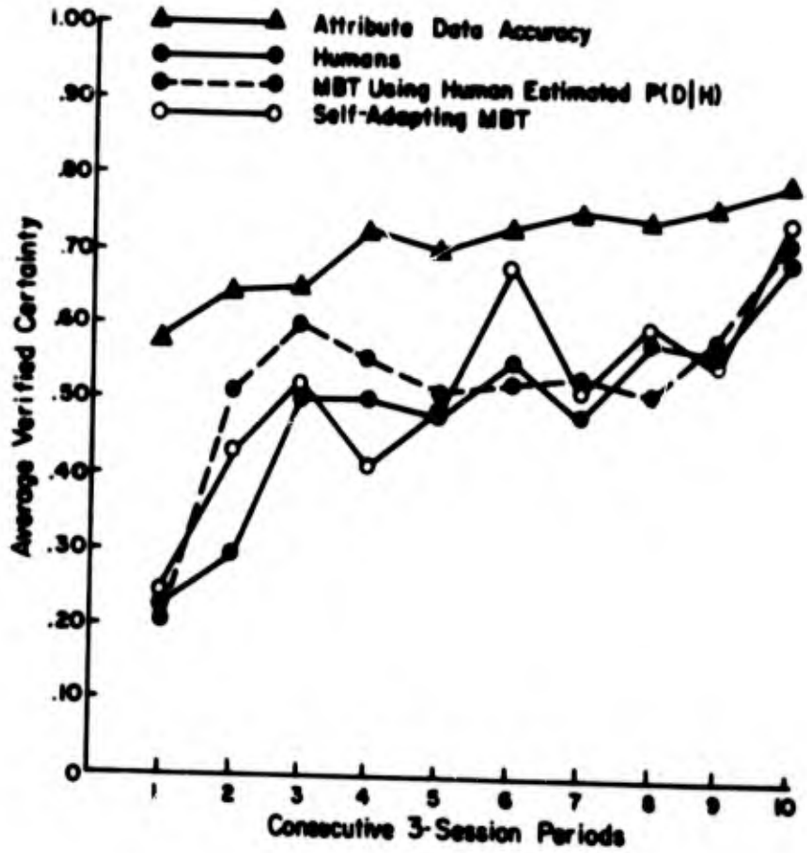


Figure 3. Human and Automated Posterior Probability Estimates over 30 Experimental Sessions.

by the four TEs (taken as a group), the four MBT solutions on the basis of TE P(D|H) estimates (taken as a group), and the self-adapting MBT. On the abscissa are three-session periods. The ordinate values refer to the average verified certainty scores obtained with reference to the 18 developmental groupings terminating in each three-session period. The solid circles connected by the solid line indicate unaided human performance. Each of these data points represents the average verified certainty scores of the four TEs (taken as a group) for the 18 developmental groupings terminating in that three-day period, i.e., each data point represents the overall average of human verified certainty scores in 72 P(H|D) estimates. The solid circles connected by the broken line represent the average verified certainty score made by the MBT solutions using TE estimates of P(D|H) for the 18 developmental groupings in that three-session period. Since there were four separate solutions for each of the 18 developmental groupings, each of these data points represents the average verified certainty score of 72 P(H|D) solutions. The open circles represent the performance of the self-adapting MBT solutions. Since only one such solution was provided for each developmental grouping, these data points represent the average verified certainty score of 18 P(H|D) estimates.

As figure 3 indicates, higher P(H|D) estimates were placed in the correct hypothesis categories by both MBT solutions early in the session sequence.

The MBT using TE estimates of P(D|H) yielded higher verified certainty scores than the humans until about session 15. The self-adapting MBT yielded higher verified certainty scores than the humans until about session 9. Thereafter, in both cases, the human and automated estimates are quite similar except for the one irregularity in the sixth three-session sequence. Figure 3, although a representation of group data with respect to the TEs and the MBT solutions using their P(D|H) estimates, is also quite representative of individual performance. Graphs (not shown) for each of the TEs showed the same early superiority up to session 15 for the solutions incorporating the human-estimated P(D|H), and up to session 9 for the self-adapting MBT. After session 15, in all cases the performance of the humans and these MBT solutions were quite similar.

Also illustrated in figure 3 is the fact that the accuracy of the attribute data upon which the various P(H|D) estimates were based was not constant throughout the experiment. The accuracy of these data, indicated by the solid triangles in figure 3, increased slightly as the experiment progressed. Each data point represents the verified certainty scores averaged over all 25 data classes and over all of the 18 developmental groupings terminating in the three-session period. Recall that verified certainty scores, with respect to the attribute data classes, refer to the value of the data state estimate placed by an ISO in the correct state of the data class for a particular developmental grouping. The measure indicated in figure 3 is admittedly a gross indication of accuracy since averaging was performed across data classes having several different possible states. The purpose of its inclusion in the figure is merely to point out that a portion of the increase in P(H|D) estimation accuracy may be due to the fact that the data upon which these estimates were based became slightly more accurate as the experiment progressed. That the entire P(H|D) estimation increase is not due to the data accuracy increase alone can be seen by comparing the slopes of the attribute data and P(H|D) curves between points at various locations on the graph. Moreover, the same attribute data were common to all three types of P(H|D) estimates and the primary interest was in a comparison of the relative accuracy of these estimates.

Table 3 lists the means and standard deviations of these verified certainty scores over the entire 30 sessions.

A discussion of the distributions of the verified certainty scores at this juncture will facilitate subsequent discussion of the statistical analysis of these scores. Figure 4 illustrates the distributions of verified certainty scores for the four unaided human estimates of P(H|D) and for the four MBT solutions based upon the human estimates of P(D|H). On the abscissa are verified certainty score class intervals. The ordinate refers to the frequency of verified certainty scores occurring in these class intervals. The data points represent scores pooled from all four subjects and from all four MBT solutions based upon the subjects' P(D|H) estimates.

It is quite apparent from figure 4 that the greater number of estimates placed in the correct hypothesis categories was either very high (.905 to 1.000) or very low (.000 to .104) for both the humans as a group and the MBT solutions incorporating the human estimates of P(D|H). There were, in other words, large numbers of definite commitment-type responses made by the humans

TABLE 3

MEANS AND STANDARD DEVIATIONS OF VERIFIED CERTAINTY SCORES

Subject	Unaided Human		MBT Using Human Estimated P(D|H)		Self-Adapting MBT	
	Mean	SD	Mean	SD	Mean	SD
1	.420	.391	.507	.435	.520	.388
2	.596	.433	.571	.429		
3	.495	.429	.529	.407		
4	.447	.415	.482	.438		
Overall Subjects	.490	.422	.523	.428		

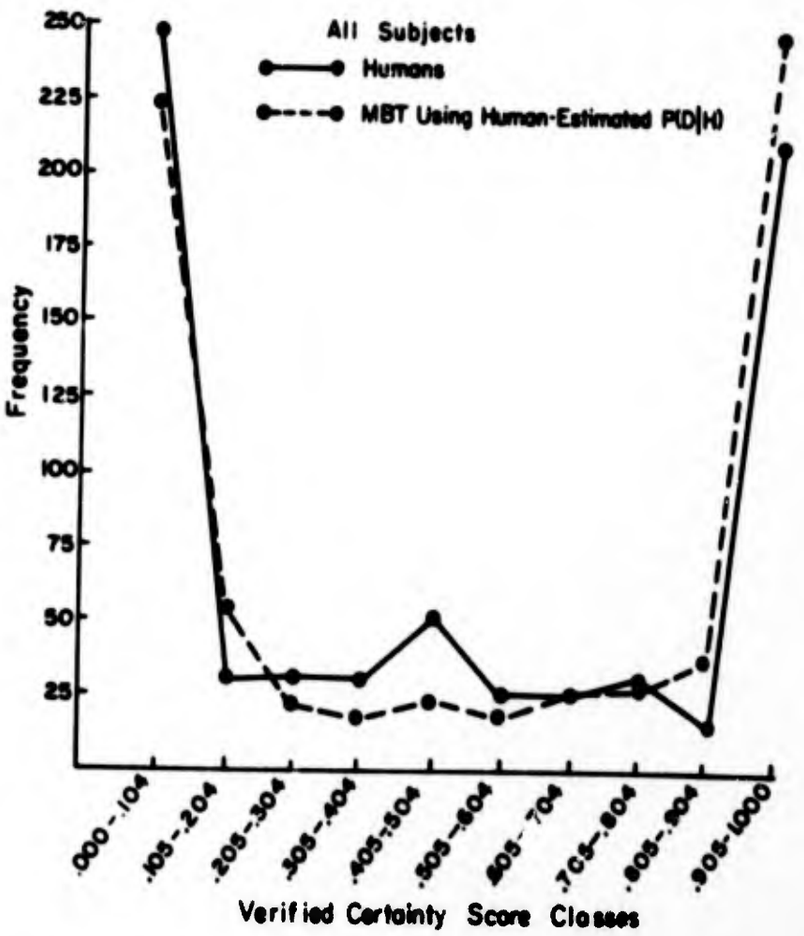


Figure 4. Distributions of Verified Certainty Scores, Experiment I.

and the MBT. If the humans' P(H|D) estimates were generally conservative, one would have expected higher frequencies of scores in the middle class intervals. Notice, however, that the distribution of human scores does show a considerably greater frequency in the class interval .405 and .504. The score distributions for the individual subjects shown in Appendix II indicate that this greater human score frequency in class interval .405 to .504 was due, almost entirely, to one subject (subject No. 1). Also observe in Appendix II the extremely similar score distributions for Subjects 2 and 3 and the MBT solutions based upon their P(D|H) estimates. Figure 5 illustrates the distribution of verified certainty scores for the self-adapting MBT. The lower frequencies in each class interval are due to the fact that only one solution was obtained for each developmental grouping.

Since figures 4 and 5 show only the frequency of various sizes of estimates placed in the correct hypothesis category, distributions of all P(H|D) estimates are needed in order to provide a more complete response profile for the humans and the MBT. Accordingly, figure 6 illustrates, in terms of ten P(H|D) estimate size class intervals, the frequency of all P(H|D) estimates in these class intervals for the four humans and the four MBT solutions

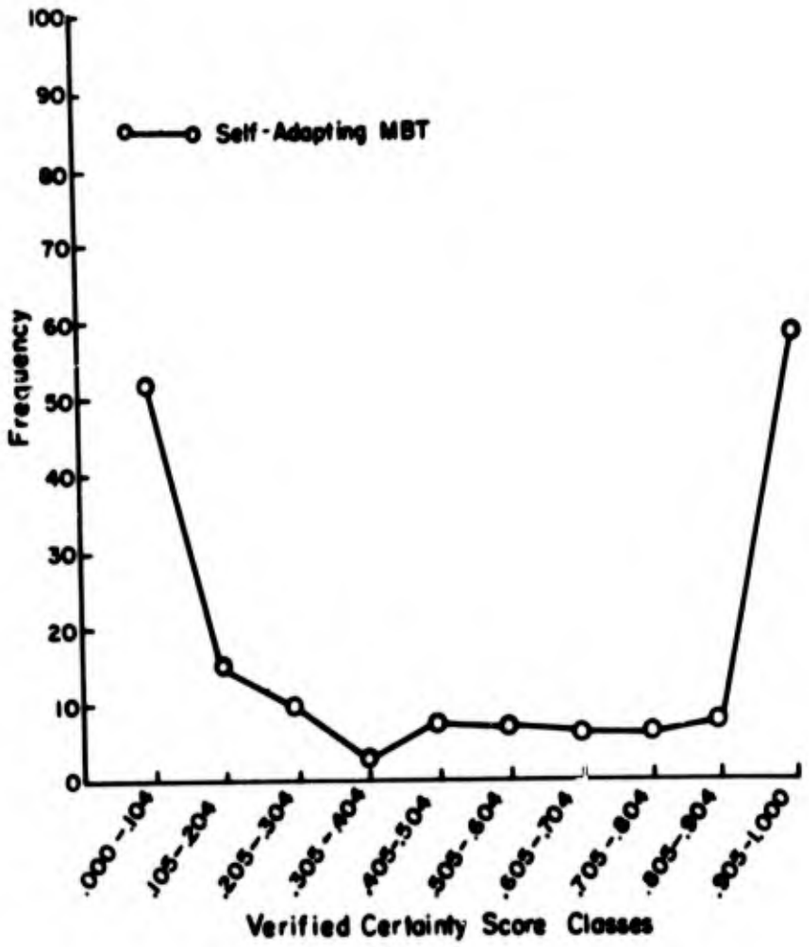


Figure 5. Distribution of Verified Certainty Scores for the Self-Adapting MBT.

based upon the human estimates of P(D|H). The data points in either of the two distributions are not independent of each other since the distributions deal with estimates which must have unit sum across the eight hypothesis categories. For example, for every estimate of 1.0 recorded there were seven estimates of 0.0 recorded. Both distributions show large numbers of extremely high and, concomitantly, extremely low P(H|D) estimates. The greater frequency of human estimates in the .405-.504 class interval is again due essentially to the performance of one subject. Appendix III, showing distributions of all P(H|D) estimates for each subject, illustrates the extreme preference for estimates in this interval by subject 1. It is apparent that subject 1 made .50-.50 estimates using only two hypothesis categories on a large number of occasions. Subject 4 contributed to the greater human frequency of estimates in the .205-.504 range. Apparent from his distribution in Appendix III is his use of .30-.30-.40 estimates in only three hypothesis categories. These preceding statements are warranted because of the fact that the humans typically expressed their estimates only to the nearest .10 or .05. Subjects 2 and 3 again show response patterns remarkably similar to those of the MBT solutions incorporating their P(D|H) estimates. In the distributions shown

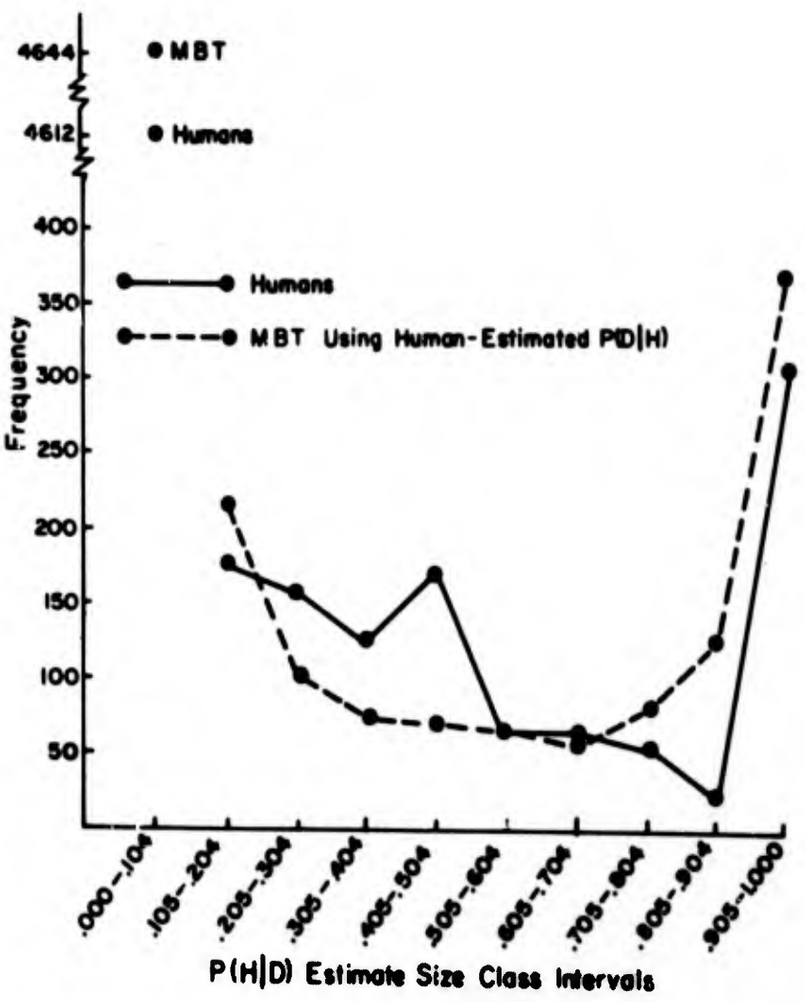


Figure 6. Distributions of All P(H|D) Estimates.

in figure 6 and Appendix III there were in every case a very large number of estimates in the class interval .000-.104. This was to be expected, of course, since there were fairly large numbers of estimates in the class interval .905-1.000. In order to illustrate the precise frequency of estimates in the interval .000-.104 and still portray the major portion of the distributions without recourse to a scale transformation, the ordinates were broken in two places and the two data points were placed adjacent to the correct frequencies as they appear on the ordinate. The data points in the first two class intervals were not connected lest the reader be presented with a distorted view of the correct distributions.

Figure 7 shows the distribution across the ten class intervals of all the P(H|D) estimates produced by the self-adapting MBT. The form of this distribution is quite similar to that of the MBT solutions using the human-estimated P(D|H) values.

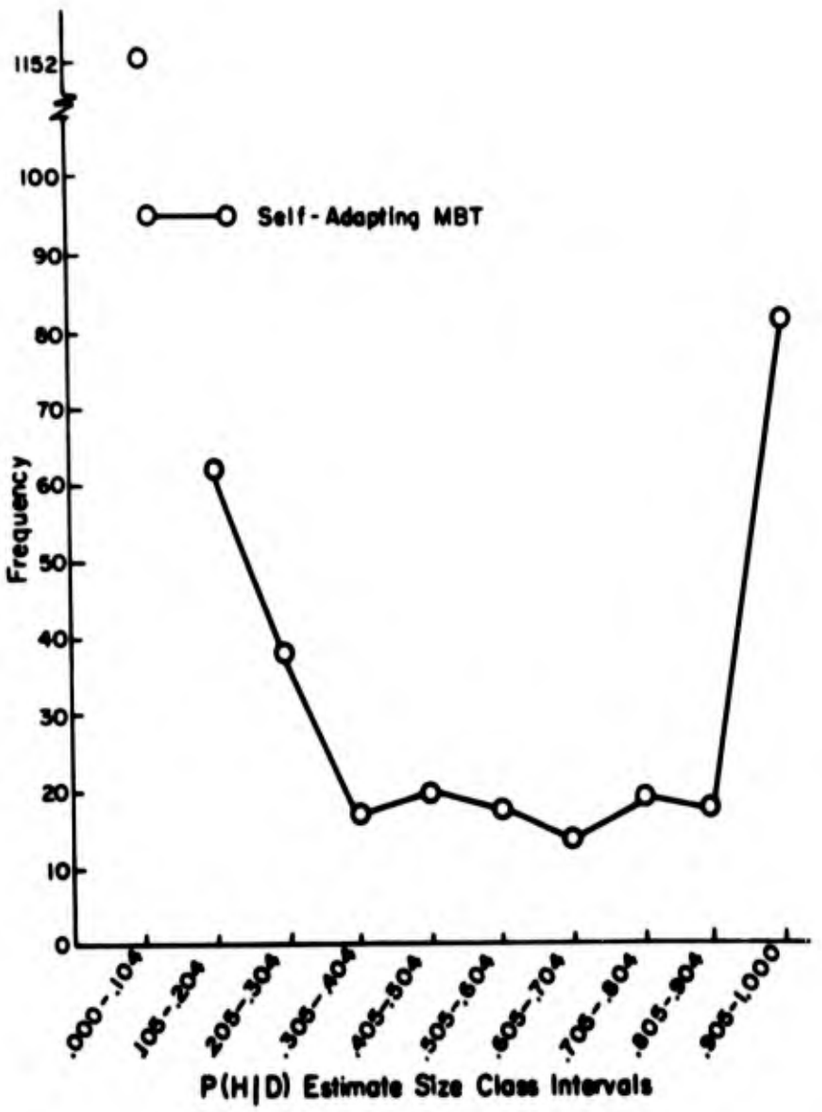


Figure 7. Distribution of All P(H|D) Estimates by the Self-Adjusting MBT.

Recalling the dramatic verified certainty score distribution change observed in an earlier experiment (ref. 17), it is also of some interest to observe how the distributions of verified certainty scores for the humans and the MBT changed as environmental experience increased. Figure 8 illustrates the form of the distributions of verified certainty scores for the four human subjects during the first, middle, and last ten-session periods during the experiment. The graph clearly illustrates the increase in the number of large scores as the experiment progressed. With the exception of subject 1, to whom one can attribute the preponderance of scores in the

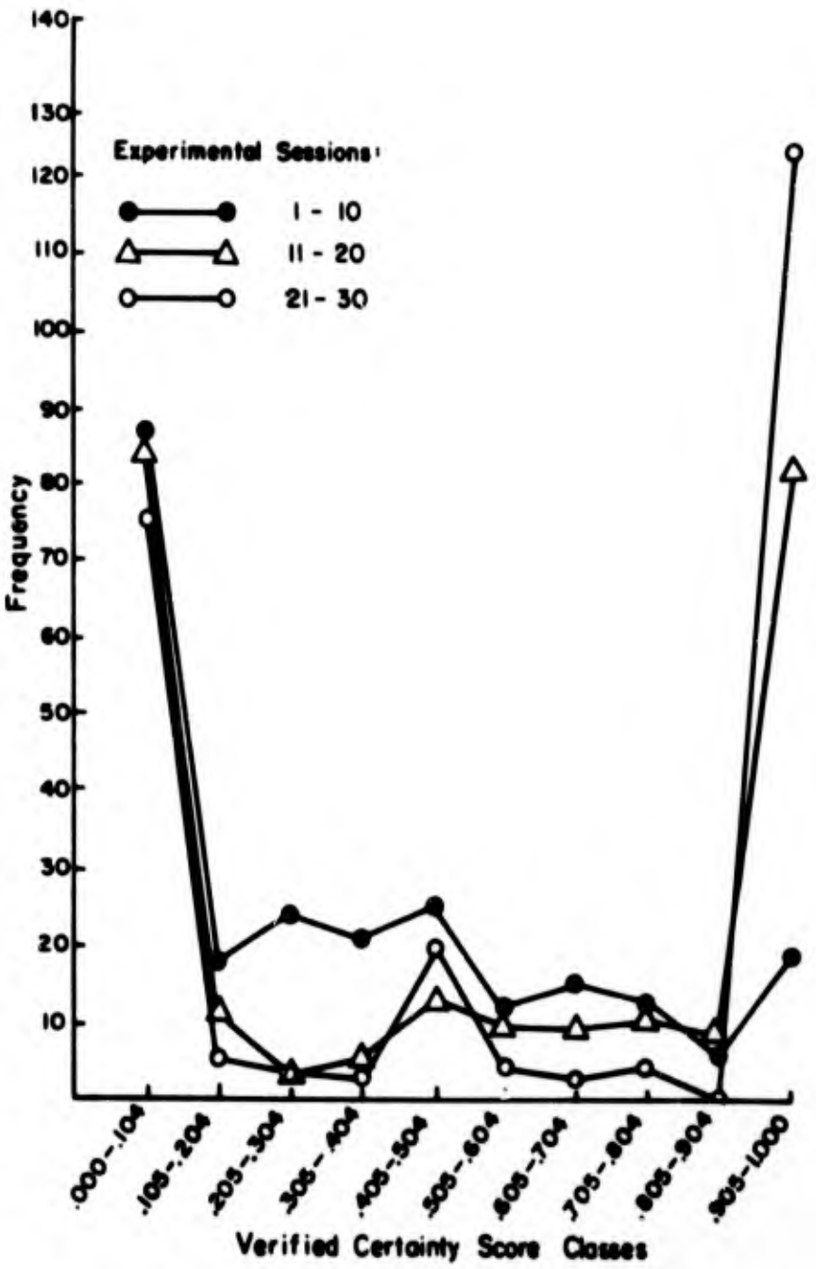


Figure 8. Distributions of Verified Certainty Scores for Humans in the First, Middle, and Last Thirds of Experiment I.

.405-.504 class interval, the subjects' score distributions changed during the experiment and by session 21 resembled very closely the form of the MBT distributions shown in figure 9. Notice in figure 9 that the MBT using the human estimates of P(D|H) produced very few early scores in the middle score class intervals. There was, however, a fairly substantial increase in very large scores for the MBT as the experiment progressed.

What the verified certainty score distributions do not indicate, of course, is the degree of relationship between the human estimates and the two types of MBT estimates as the scores for each developmental grouping

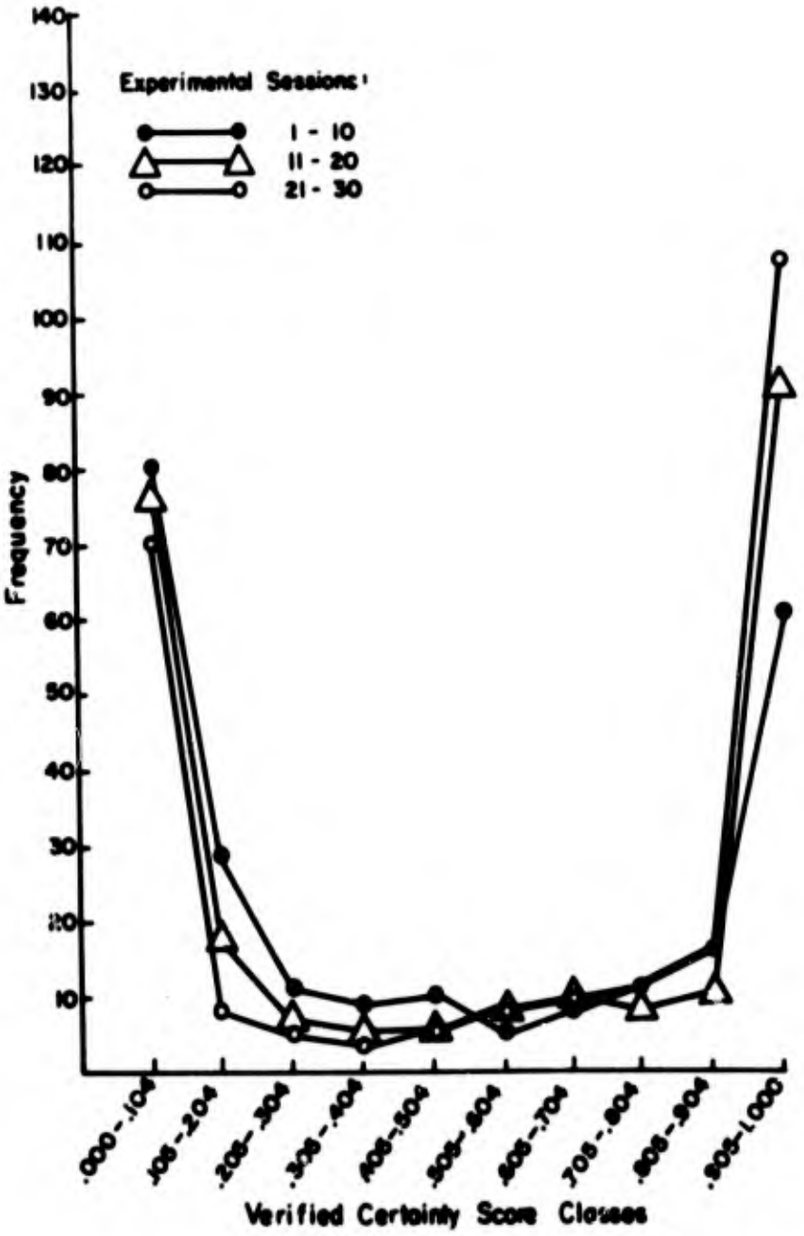


Figure 9. Distribution of Verified Certainty Scores for the MBT in the First, Middle, and Last Thirds of Experiment I.

are examined. Rank-order correlations were, therefore, calculated between each subject's verified certainty scores and the scores for the MBT using his P(D|H) estimates, between each subject's scores and those for the self-adapting MBT, and between the scores for the two types of MBT solutions. The Pearson product-moment correlation was not employed primarily because of the decidedly bimodal character of each score distribution. A fairly large number of tied scores occurred for each subject because of the tendency on the part of the subjects to round off their estimates to the nearest .05. The correction for tied ranks described by Siegel (ref. 18) was incorporated in all calculations of Rho. Table 4 lists the rank-order correlations between the human and MBT verified certainty scores.

Most of the Rho values in table 4 are of moderate size and all are highly significant. This latter finding is perhaps not too surprising in view of the number of degrees of freedom in the calculations. As a rough measure of relationship in the present instance, the Rho values presented are probably satisfactory as long as one does not try to interpret them in the exact way that a product-moment correlation would be interpreted. The reason for using the two-tailed probability values is discussed below.

Now that the characteristics of the verified certainty score distributions have been presented, the statistical analysis of these scores can be discussed. The primary factor affecting the choice of a statistical hypothesis-selection method was the form of the score distributions shown in figures 4 and 5. Clearly, a distribution-free method was called for.

TABLE 4

RANK-ORDER CORRELATIONS BETWEEN HUMAN AND MBT VERIFIED CERTAINTY SCORES
(df = 177)

Rank-Order Correlation between Verified Certainty Scores for:	Subject No.	Rho	t	P (Two-Tailed)
I. Each subject and the MBT using his P(D|H) estimates	1	.58	9.549	<.001
	2	.53	8.373	<.001
	3	.55	8.703	<.001
	4	.55	8.709	<.001
II. Each subject and the self-adapting MBT	1	.38	5.448	<.001
	2	.56	8.943	<.001
	3	.55	8.801	<.001
	4	.47	6.989	<.001
III. MBT calculated from each subject's P(D|H) estimates and the self-adapting MBT	1	.57	9.277	<.001
	2	.70	13.140	<.001
	3	.59	9.609	<.001
	4	.59	9.709	<.001

Another consideration involved the degree to which the score samples for each of the three groups were related. All three types of P(H|D) estimates were based upon the same attribute data. In addition, the human estimates and one form of MBT estimate were presumably based upon the same set of P(D|H) values. For these reasons the three verified certainty score samples were considered as being related. In view of these considerations, and because there were only three groups to be compared, the Wilcoxon matched-pairs test was performed to test the significance of the differences between each of the three groups. Before presenting the results of the significance tests, the specific experimental hypothesis being examined warrants further discussion. It has been previously mentioned that although there were data in existence which tend to show superiority of automated estimates of P(H|D) over unaided human estimates, one might argue that the humans in most of these studies were at some disadvantage because of lack of experience and insufficient instructions. From one point of view, then, under better conditions the results might indicate no differences or even differences in favor of human subjects. At this juncture in the research, therefore, the experimenters preferred to allow for the possibility of performance differences in either direction. This was particularly the case with respect to the self-adapting version of the MBT since the parameters involved in its calculation were more or less arbitrarily chosen. In view of these considerations, the tests of significance performed on the results of this experiment were two-tailed. The results of the Wilcoxon tests are as follows:

a. The overall difference between the verified certainty scores for the 180 P(H|D) estimates produced by the four humans and those of the four MBT solutions using the human estimates of P(D|H) did not meet the conventional requirements for statistical significance. The average human verified certainty score for each of the 180 developmental groupings was compared with the average MBT solution. The hypothesis of no difference was not rejected. N (the number of greater-than-zero differences) = 173, Z = -1.2324, p (two-tailed)< .2196.

b. The overall difference between the average human verified certainty score for each of the 180 groupings and the self-adapting MBT solution also was not statistically significant. N = 172, Z = -1.619, p (two-tailed) < .1052.

c. The overall difference between the average MBT solution using the human estimates of P(D|H) and the self-adapting MBT solution for every developmental grouping was not significant. N = 175, Z = .0715, p (two-tailed) < .9442.

2. Dichotomous score analysis: The other method of measuring human and MBT P(H|D) estimation performance involved an account of the number of occasions on which the humans or the MBT placed their highest or first-choice P(H|D) estimates in the correct hypothesis categories. Their highest estimations were, therefore, treated in a dichotomous fashion as being either correct or incorrect. Table 5 illustrates the number of occasions on which the highest P(H|D) estimates were placed in the correct hypothesis category by the humans and by both types of MBT solutions. For each subject and for each MBT solution there were 179 possible correct hypotheses. (One developmental

TABLE 5

DICHOTOMOUS SCORES IN EXPERIMENT I

Subject	Unaided Human	MBT Using Human's P(D|H)	Self-Adapting MBT
1	62	95	
2	111	105	
3	91	96	
4	71	91	
Total	335	387	96

grouping result was discarded. This also accounts for the fact that there were 177 instead of 178 degrees of freedom in calculating the significance of the Rho values.)

The lower total recorded in table 5 for the self-adapting MBT is due to the fact that only one such solution was calculated for each developmental grouping. When the scores for the MBT using human estimates of P(D|H) are compared with the associated human scores one finds MBT superiority in all but one case. The overall difference between the humans as a group and the MBT solutions using their P(D|H) estimates was found to be statistically significant in favor of the MBT [Wilcoxon test, N = 28, Z = -2.573, p (two-tailed) < .01]. To make this comparison the total number of dichotomous scores made by the four humans in each experimental session was compared with the similar total for the four MBT solutions. Comparing the self-adapting MBT total against each human total, one also finds MBT superiority in the same three cases. The differences between each subject's total and the self-adapting MBT total were not analyzed statistically.

3. P(D|H) estimation accuracy: It should already be apparent from the performance of the MBT using the subjects' estimates of P(D|H) that these conditional probabilities were estimated quite reasonably. However, it is extremely difficult, if not impossible, to present a single summarized description of the overall accuracy of these estimates. The agreement scores (α_j) defined in an earlier section are specific to a particular data class. Pooling or averaging these α_j scores over data classes with differing states or conditions does not seem justifiable because of the different probabilistic structure underlying these data classes. Moreover, the agreement scores are also specific to a certain hypothesis category. Pooling each α_j score over the eight hypothesis categories seems more defensible than the previous pooling procedure, but a description of the accuracy of estimation specific to certain hypotheses is lost in the process. The only other alternative is to present the average α_j score for each data class under every hypothesis (a matrix of 200 values). For the purposes of the present report, however, a less precise but considerably more convenient representation will be provided.

In Appendix IV are shown the α_j score distributions for each of the threat evaluators in four selected data classes, each having 8, 6, 4, or 2 alternative states or conditions. (The data class numbers I, III, IX, and XIX refer to those listed in Appendix I.) The α_j scores in each distribution have been pooled across the eight hypothesis categories. The mean α_j score for each subject in each data class is also given. The distribution for data class IX (two alternatives) is the best distribution, in terms of location, of the 25 possible distributions. That for data class XIX is the worst of the 25. All others lie somewhere between these extremes. The abscissa in each graph shows α_j score class intervals. Recall that $\alpha_j = 1.00$ means perfect agreement between the estimated and true $P(D_{jk}|H_i)$ values while $\alpha_j = 0.0$ means perfect lack of agreement between these values. The ordinate shows the frequency of α_j scores in these class intervals.

The most critical feature underlying an interpretation of these scores is the fact that the $P(D_{jk}|H_i)$ values were estimated by the subjects upon observation of repetitive events. This fact will be taken into consideration when the experimental results are interpreted in the next section.

E. Discussion and Interpretation of Results

There seem to be at least four major considerations which one ought to keep in mind as the results of this experiment and the one which follows are discussed and interpreted. These considerations define, more or less, the types of situations in which the results of these experiments stand the greatest chance of application. First, the subjects gained experience in an environment whose critical events were essentially repetitive in nature. To facilitate $P(D|H)$ estimation the subjects kept rather careful accounts of the number of times a certain data state occurred in the presence of some hypothesis. The stimulus environment, therefore, can generally be described as "frequentistic" in nature since relative frequencies could be used to indicate the probability of critical events. There seems to be nothing unrealistic about such a situation. Indeed, in the field of medicine, for example, careful records are maintained with respect to the occurrence of certain symptoms given various known causes or states of nature. The second consideration is that the design of the stimulus environment and the subjects' tasks allowed for situations in which one is not always precisely sure about what datum one is observing. Dodson's MBT allows for this observational uncertainty. Third, in both experiments subjects were used who were already quite experienced in dealing with the critical environmental events. The performance of these individuals will surely be different from that of a collection of naive subjects who perform only briefly in unfamiliar circumstances. Finally, the estimators of $P(H|D)$ were always subsequently informed about the true hypothesis and consequently about the quality of their estimates.

The first major result of the experiment was that with experience the human's performance eventually matched that of the MBT solutions. The verified certainty scores for the MBT using human estimates of $P(D|H)$ were superior in every instance until about session 15. Presumably, this indicates that a greater amount of consistency (or predictability) existed in the $P(D|H)$ values than the subjects actually perceived while making their $P(H|D)$ estimates. With increased experience the subjects gained a greater appreciation for the consistency which their own $P(D|H)$ estimates said was

in the environment and which the MBT "recognized" quite early in the sequence of trials. There is another explanation for the performance increase. This explanation was indicated in a questionnaire administered to the four TEs after every four sessions during the experiment. It seems that, out of necessity, the humans, when initially confronted with the mass of attribute data, acted to reduce the complexity of the situation and in so doing lost valid or useful predictory information. With experience, however, the better subjects took greater notice of a larger number of data classes in which there existed consistency or predictability. This improved consistency recognition was reflected in improved scores. The increase in the number of large estimates with experience is clearly shown in figure 8. There is another potential explanation for this finding. There were no explicit costs or payoffs in these experiments and, therefore, the subjects were never penalized for incorrect diagnoses. High scores in themselves were certainly rewarding to the subjects. One wonders whether or not the extreme estimates would still be obtained in a direct cost-payoff situation. Under the conditions of the experiment, therefore, the optimal strategy for the humans was probably not Bayesian. Beyond a certain threshold of certainty the subject might just as well have estimated 1.0 since there was nothing to lose. Most assuredly it was an easier task to write 1.0 in one hypothesis category than to give vernier estimates across all eight categories even though the estimate may have been much higher than the consistency of the data justified.

In terms of verified certainty scores the MBT solutions using human estimates of $P(D|H)$ were not superior to the unaided human estimates, while in terms of the dichotomous scores the MBT solutions were significantly superior. The latter explanation in the preceding paragraph should help to account for this discrepancy. In comparing the relative performance of MBT and humans the average verified certainty scores by themselves may be misleading because of the apparent increased tendency of the subjects to use higher certainty estimates than the data justify. If it is true that the subjects tended to maximize the size of their estimates (which the MBT does not do), then both types of measures are needed in order to provide an accurate evaluation of the relative performance of human and MBT.

A more thorough analysis of individual performance proved to be highly informative. The best performance in terms of both measures were those of subjects 2 and 3. It happens that their response profiles (Appendixes II and III) closely match those of the MBT using their $P(D|H)$ estimates. In addition, the questionnaire revealed that a high percentage of the attribute data influenced their $P(H|D)$ estimates. Subjects 1 and 4 adopted a considerably less sensitive approach. Already mentioned was their preference for estimates in only two or three categories. Table 3 also confirms their lack of sensitivity. Observe that these two subjects have the lowest mean scores and also the lowest score standard deviations. The questionnaire revealed the startling fact that even after the 28th session both subjects were completely discarding information in 8 or more of the 25 data classes. Subject 2, on the other hand, utilized the information from as many as 23 data classes in his estimates.

The questionnaire also helped to establish the fact that none of the humans processed the data from each class in an iterative fashion.

Apparently the P(H|D) estimation task for the humans involved taking notice of those data which were found to be the best discriminators among the various hypotheses, eliminating certain of the hypotheses on the basis of these data, and then loading with certainty estimates the remaining hypotheses.

VI. EXPERIMENT II: HUMAN AND AUTOMATED ESTIMATION OF POSTERIOR PROBABILITIES UNDER SEVERAL LEVELS OF INPUT INFORMATION FIDELITY

A. Objectives

Individuals faced with the task of evaluating events in a hostile environment will seldom, if ever, have access to information about these events which is utterly precise or reliable. Indeed, these tasks involve the use of information which is less than perfect for any one of a number of reasons. At least the following causes of decreased input information fidelity are worthy of note:

1. Information may be gathered by sensors with varying degrees of inherent resolving power. On any given occasion, this resolving power may be degraded for some reason or another.

2. Information is lost or altered in any transmission or recoding process. Verbal descriptions or pictorial representations of events are always somewhat less than the event itself.

3. Relevant information may be available only after a painstaking search through irrelevant or extraneous information. Important information may be lost in the process of filtering out relevant signals from "noise."

4. Information may become available which has been intentionally disseminated in order to obscure certain critical events and thus to deceive an observer or evaluator of these events. Thus, authentic information may be confused with that which is spurious.

Each of these factors, alone or in combination with others, produces a common result: discriminability among environmental events is reduced.

In evaluating the efficacy of the Bayesian paradigm for information-processing systems, it is surely of interest to see how the various probabilistic estimates made by humans or automated devices fare under various levels of input fidelity. Edwards, in his description of a probabilistic information-processing system suggests that "a probabilistic system can afford to accept and use with profit information so seriously fallible or degraded that it would be excluded or ignored in deterministic systems" (ref. 18, p. 1). The probabilistic system which Edwards has in mind is one in which posterior probabilities estimates are calculated on the basis of Bayes' theorem using as inputs the $P(D|H)$ estimations produced by humans. One implication from Edwards' statement seems to be that if Bayes' theorem is optimal in its extraction of consistency from probabilistic data, then as data become more fallible and consistencies among data become more obscure or difficult for humans to perceive, the superiority of the Bayes-theorem-based solutions over unaided human estimates of $P(H|D)$ ought to become more apparent. The justification for this implication is as follows. As the fidelity of data describing environmental events decreases, uncertainty

about what has been observed in connection with the hypothesized states of nature increases. Contingent data-hypothesis relationships [P(D|H)] become more obscure or abstruse. The assertion then is that since Bayes' theorem is a mathematically optimal method for combining these conditional probabilities, solutions of P(H|D) based upon Bayes' theorem will be superior to estimates of P(H|D) produced by humans if the humans are not able to incorporate all the consistency contained in the contingencies or P(D|H) values as they produce P(H|D) estimates. In other words, there may simply be more consistency or predictability in the P(D|H) values (even though based upon fallible data) than the humans will be able to recognize and reflect in their estimates of P(H|D).

The major purpose of this experiment is to compare, under several levels of input fidelity, human estimates of P(H|D) and MBT solutions of P(H|D) based upon the human estimates of P(D|H). The variable in the experiment speaks directly to a performance comparison involving human estimates of two types of conditional probabilities [P(H|D) and P(D|H)]. For this reason the performance of the self-adapting MBT will not be statistically analyzed in comparison with the other two types of estimates. Another reason for not including the self-adapting MBT in the analysis is that only one such solution will be available for each developmental grouping, while there will be four each of the other two types of estimates. Individual comparison of self-adapting MBT with each subject's estimates are of limited interest in this experiment. The performance of the self-adapting MBT will, however, be presented descriptively.

B. Input Fidelity as a Variable

The input information which was utilized by the team of ISOs in developing the 25 types of attribute data consisted of verbal descriptions of the type, number, and activity of Aggressor vehicles, mobile weapons, and aircraft. These descriptions were based upon photo, radar, and infra-red sensor records obtained on simulated reconnaissance overflights of the territory of the hypothetical adversary. The overflights and the interpretation of the obtained sensor records were simulated, of course, by the computer facilities. Of the four factors mentioned above which act to reduce discriminability among environmental events, the first, sensor resolution, can be manipulated most systematically in the present stimulus environment configuration. In order to simulate the fact that there might exist several graded quality levels of sensor data, different levels of verbal description were available for each event to be observed by the ISOs. There were three quality levels of photo information and two each of radar and infra-red. To illustrate how discriminability among events in the stimulus environment is contingent upon level of sensor description, consider the following example. The verbal description of a 2-1/2-ton cargo truck given to an ISO using a level 1 photo (highest level) is "2-1/2-ton cargo truck," an unequivocal description. Under a level 2 photo, however, the description would read "medium size wheeled vehicle." In this case, the operator could not distinguish a 2-1/2-ton cargo truck from a 5-ton truck, an amphibious armored carrier, or a 140-mm rocket launcher since they are also described as "medium size wheeled vehicles" under a level 2 photo. Using level 3 photo (poorest), the verbal description of a 2-1/2-ton cargo truck would be "self-propelled vehicle." Discrimination is now very poor indeed since there are many types of mobile

weapons and vehicles also described as "self-propelled vehicle" under a level 3 photo. It is also true that discriminations among critical events, which the 25 attribute data classes reflect, are based upon frequency counts of these vehicles, weapons, and aircraft. One can systematically induce uncertainty into these counts as the quality level of sensor records is reduced. For example, a certain unit in Aggressor's surface forces contains 131 wheeled vehicles as seen under photo level 1 (highest quality). This vehicle count happens to be unique and the unit can always be identified. Before the ISO receives his vehicle total, however, the experimenter can, by means of a programmed algorithm, add or subtract a random number from the original total. This random number is chosen within a certain range which indicates the resolution of the sensor. Suppose, for example, that the degraded sensor was capable of providing an accuracy of ±10% on vehicle totals. If the original total were 131, the experimenter could, on each occasion, add algebraically a random integer between -13 and 13 to the original total, making the possible range of vehicle totals 118 through 144. This will cause the ISO to confuse this unit with other units whose range of totals falls within the range of totals possible for the unit in question. This procedure for degrading totals can be supplied independently of the procedure for degrading verbal descriptions.

In all previous experiments the ISOs have always been allowed unrestricted access to top-quality sensor records. (This was also the case in Experiment I reported above.) In the present experiment, where fidelity of input is the variable of interest, the experimenter manipulated both the verbal and numerical descriptions provided by sensor records, thus simulating the effects of degraded photo, radar, and infra-red images. Following are the chosen levels of the fidelity variable and the method by which these levels were induced:

1. Level I. Highest fidelity: No change was induced here over previous experiments. ISOs had unrestricted access to top-level photo, radar, and infra-red sensor records.

2. Level II. Intermediate fidelity: ISOs had access only to the poorest quality photo, radar, and infra-red records, but only the verbal descriptions showed degradation (i.e., vehicle, weapon, and aircraft totals did not suffer degradation). This condition induced considerable confusion in the estimates provided by the ISOs in 13 of the 25 attribute data classes.

3. Level III. Poor fidelity: ISOs had access only to poorest quality photo, radar, and infra-red sensor records. In this case, however, the experimenter degraded both vehicle description and vehicle, weapon, and aircraft totals. This condition induced considerable confusion in the estimates provided by the ISO in 24 of the 25 attribute data classes. Sensor resolution was set at ±10%.

The particular configurations in levels II and III were chosen because of their even effects upon the 25 attribute data classes, i.e., approximately half were affected by level II and all but one by level III. Following is the effect which the experimenter hoped to induce by this manipulation. In the description of the first experiment, it was mentioned that the ISOs

produce probabilistic estimates of the level or state of each attribute data class for every Aggressor developmental grouping. On the basis of top-quality sensor records they receive unequivocal verbal descriptions and exact weapon and vehicle totals. In a large number of cases unique identifications of Aggressor surface and air units were possible with this top-quality input information and the ISO could estimate the state or level of a certain data class with some confidence. Under the reduced fidelity conditions (for those data classes which are affected by the degradation level) lack of discriminability should force him to be considerably less certain about the level or state of the data class in question. The effects of increased attribute data uncertainty (caused by decreased sensory record fidelity) upon the P(H|D) estimates produced by the humans and the MBT were, therefore, the primary concern of this experiment.

C. Experimental Design

There were actually two variables in the experiment: sensor record fidelity (three levels) and mode of P(H|D) estimation (three levels of which only two were to be analyzed statistically). The plan of the experiment, in terms of these two variables, is shown in table 6. The experiment consisted of 30 4-hour sessions. In each session, as in the previous experiment, six developmental groupings terminated. The first 10 sessions were performed under fidelity level I, the next 10 under fidelity level II, and the last 10 under fidelity level III. Breaking up the sessions per level in an attempt to balance out possible residual effects was not possible because of the nature of the task. For each mode of estimation each 10-session period was a completely new learning experience. At the beginning of each 10-day session the humans and the MBT operated upon an entirely different set of environmental contingencies and Aggressor response alternatives. Although the 25 types of attribute data remained the same, the relationships between these data and the eight hypotheses (i.e., the true P(D|H) values) were changed in each 10-session period. The numbers in the cells of table 6 refer to the number of individual P(H|D) estimations made under each mode of estimation. Since there were four humans, six developmental groupings terminating per session, and 10 sessions per fidelity condition, there were

TABLE 6

PLAN OF EXPERIMENT II

Mode of P(H|D) Estimation	Fidelity Condition		
	I (Top) Sessions 1-10	II (Intermediate) Sessions 11-20	III (Low) Sessions 21-30
Human	240	240	240
MBT using human-estimated P(D|H)	240	240	240
Self-adapting MBT	60	60	60

240 estimates produced in each of the first two modes for each fidelity level. Only one self-adapting MBT solution was calculated for each grouping, making the total under this mode only 60 per fidelity condition.

D. Subjects

The subjects serving as ISOs were the same in this experiment as in the previous one. With one exception, the TEs were the same as those who served in Experiment I.

E. Results

The major results of the experiment in terms of verified certainty scores are illustrated in figure 10. Along the abscissa are the three levels of sensor record fidelity. The ordinate refers to the average verified

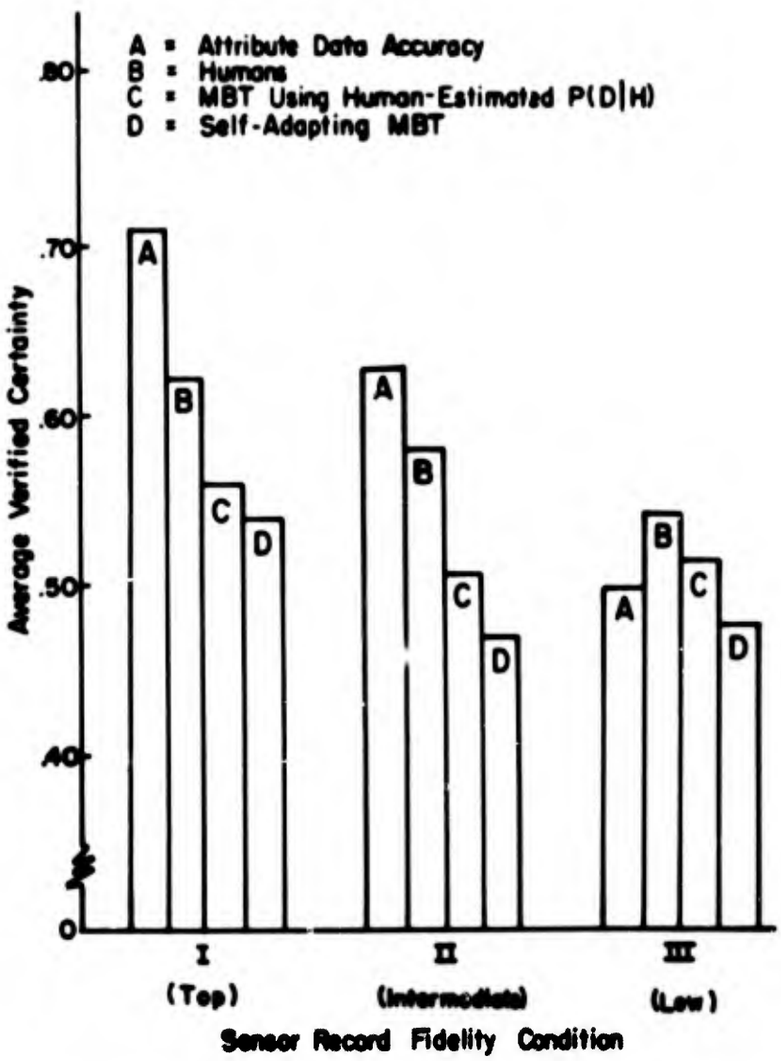


Figure 10. Posterior Probability Estimation Accuracy under Several Levels of Input Data Fidelity.

certainty scores for the 60 developmental groupings terminating in the 10-session period for each fidelity level. The data points for the humans and the MBT incorporating the human estimates of P(D|H) are averages over 240 P(H|D) estimates. Each self-adapting MBT data point is an average over 60 P(H|D) estimates. The attribute data points are gross estimates of the average accuracy in all data classes for all developmental groupings terminating in each of the three 10-session periods. The accuracy of the attribute data estimates (bar A) decreased rather sharply as the simulated sensor records were degraded. Reduction in accuracy of these data caused a corresponding decrease in P(H|D) estimation accuracy by humans and MBT. This P(H|D) accuracy decrease was slightly more drastic for the humans (bar B) than for either the MBT using the human estimates of P(D|H) (bar C) or the self-adapting MBT (bar D). Surprisingly, the humans, as a group, placed higher average estimates in correct hypothesis categories than either MBT solution.

The distributions of these verified certainty scores (Appendix V) were of exactly the same form as those obtained in the first experiment. In every case there was a distinctly bimodal distribution with extremely high frequencies at either end of the distribution. Again, as in the first experiment, the use of a distribution-free method of analysis was imposed. As previously mentioned, there were two variables of interest: fidelity condition and mode of P(H|D) estimation. The major experimental hypothesis referred, however, to an interaction between these variables. It was specifically stated in the preceding section that the MBT using the human P(D|H) estimates ought to be increasingly superior as fidelity decreased. A distribution-free method of analysis which permits one to examine interactions is the Ranks Test for Matched Data described by Bradley (ref. 19). The mean verified certainty scores shown in table 7 for the humans and the MBT using the human estimates of P(D|H) were compared using the ranks test.

TABLE 7

VERIFIED CERTAINTY SCORE MEANS AND STANDARD DEVIATIONS IN EXPERIMENT II

Mode of P(H|D) Estimation	Subject	Fidelity Condition					
		I		II		III	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Human	1	.538	.437	.572	.456	.586	.428
	2	.671	.397	.597	.442	.564	.452
	3	.638	.445	.563	.456	.479	.440
	4	.635	.443	.601	.457	.528	.457
MBT using human-estimated P(D|H)	1	.556	.396	.490	.390	.506	.373
	2	.529	.430	.530	.435	.553	.381
	3	.577	.415	.537	.441	.513	.405
	4	.567	.410	.466	.401	.498	.382

The first result of the ranks test was that the main effect of fidelity condition induced a statistically significant decrease in the overall verified certainty scores for both humans and MBT ($S = 26$, $p = .04$). The interaction between fidelity condition and mode of estimation was not significant ($S = 14$, $p = .27$). Analysis of the main effect due to mode of estimation (i.e., human versus MBT) presented a problem. In table 7, if one pools the results across all three fidelity conditions and then ranks these results, one obtains a 2×4 matrix of ranks. Now observe that all the values for the four humans will be higher than those corresponding for the MBT and, therefore, all human scores will receive a rank of 2 and all MBT scores will receive a rank of 1. The value of S, the sum of the differences between the ranks, is maximum in this case and equals 8. There are $2^n = 2^4 = 16$ ways the differences can be summed and there are only two ways of getting an S value as large as 8. The smallest p value, therefore, is $2/16 = .125$. Even though the maximum difference was observed, the small size of the matrix of ranks precluded obtaining a statistically significant result. There is another way to analyze the overall differences between the human and MBT scores which makes no additional assumptions. Resorting to an exact probabilities test, one can compare each human's score with his corresponding MBT score in each of the three fidelity conditions. These comparisons are shown in table 8.

Summing the differences between each of these two scores, one obtains $\Sigma d = .650$. Now, there are $(2)^{12}$ or 4096 possible ways in which the signs of the 12 given differences can arrange themselves. Given the differences which exist and rearranging the signs in all possible ways, one must calculate the number of occasions on which an absolute value of greater than or equal to .650 could occur. It turns out that there are 12 ways of getting $|\Sigma d| > .650$ and only one way of getting $\Sigma d = .650$. Therefore, the probability of a difference greater than or equal to $|\Sigma d| = .650$ by chance, given the observed differences in table 8, is $13/4096 = .0032$. The main effect of mode of estimation in terms of verified certainty scores was statistically significant ($p = .0032$) in favor of the humans.

TABLE 8

DIFFERENCES BETWEEN HUMAN AND MBT AVERAGE VERIFIED CERTAINTY SCORES IN EXPERIMENT II

Subject	1			2			3			4		
Fidelity Condition	I	II	III	I	II	III	I	II	III	I	II	III
Human Score	.538	.572	.586	.671	.597	.564	.638	.563	.479	.635	.601	.528
MBT Score	.556	.490	.506	.529	.530	.553	.577	.537	.513	.567	.466	.498
Difference d	-.018	.082	.080	.142	.067	.011	.061	.026	-.034	.068	.135	.030

TABLE 9

RHO VALUES: EXPERIMENT II

Rank-Order Correlation between Verified Certainty Scores for:	Subject	Fidelity Condition I	II	III
I. Humans and MBT using human P(D|H)	1	.66	.45	.50
	2	.71	.51	.30
	3	.50	.67	.51
	4	.67	.52	.32
II. Humans and self-adapting MBT	1	.52	.59	.53
	2	.64	.49	.39
	3	.60	.56	.40
	4	.63	.64	.46
III. Both MBT solutions	1	.82	.75	.70
	2	.69	.74	.60
	3	.78	.81	.72
	4	.77	.69	.78

TABLE 10

DICHOTOMOUS SCORES: EXPERIMENT II

	Subject	Fidelity Condition I	II	III	Total
I. Humans	1	30	32	29	90
	2	43	36	37	116
	3	39	34	27	100
	4	47	34	28	99
	Total	149	136	121	406
II. MBT using human P(D|H)	1	34	30	36	100
	2	34	32	37	103
	3	35	33	31	99
	4	36	30	32	98
	Total	139	125	136	400
III. Self-adapting MBT		34	35	35	104

Appendix VI shows human and MBT performance in terms of verified certainty by session in each experimental condition. The similarity of the three "learning" curves in each condition is considerable. The rank-order correlations between these three types of estimates in each fidelity condition are given in table 9.

All Rho values in table 9 are significant beyond $p < .02$ and the significance test for all values was calculated with 58 degrees of freedom.

Table 10 shows the dichotomous scores for the humans and MBT in each fidelity condition. The scores shown for each individual subject and MBT solution in each of the three conditions refer to the number of occasions (out of a possible 60) on which the first-choice or highest P(H|D) estimate was placed in the correct category. The first two types of data shown in table 10 were analyzed in the same manner as the verified certainty data. The main effect due to fidelity condition and the interaction between fidelity condition and mode of estimation [human versus MBT using human P(D|H)] were tested using the Ranks Test for Matched Data. Neither the main effect ($p = .273$) nor the interaction ($p = .653$) was statistically significant.

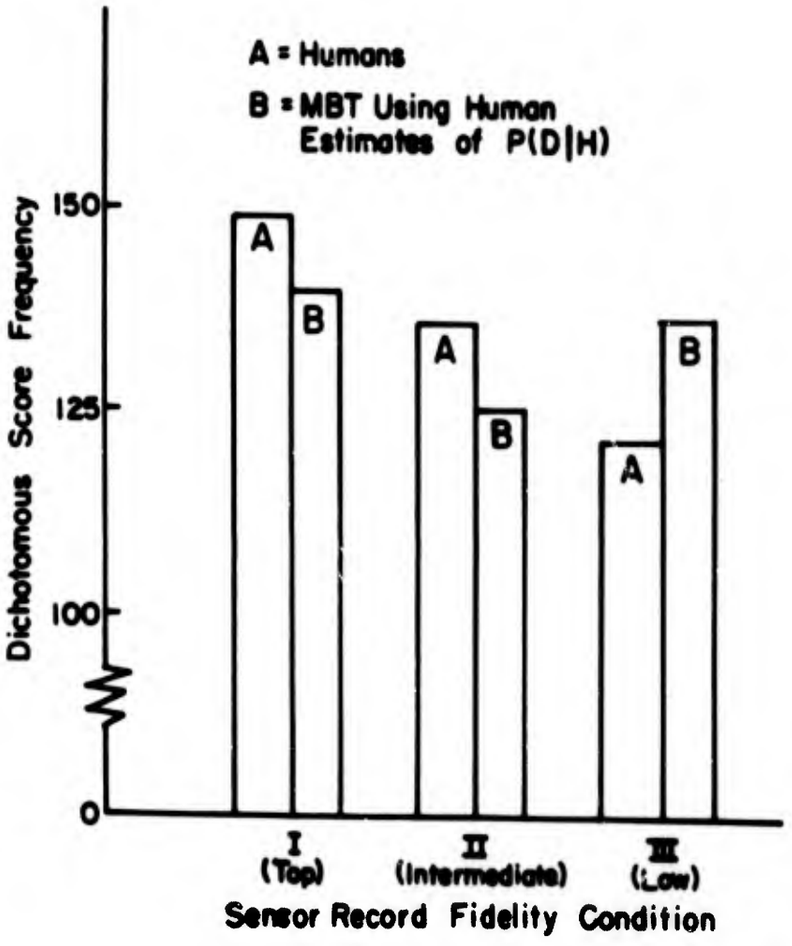


Figure 11. Dichotomous Scores for Humans and MBT under Each Fidelity Level.

The main effect due to mode of estimation was tested by the exact probability method described for the verified certainty scores. The hypothesis of no difference could not be rejected ($p = .378$). What is fairly interesting, of course, is that the largest difference between the overall human scores and the MBT scores occurred at the lowest fidelity level and favored the MBT. This result is shown graphically in figure 11.

F. Discussion and Interpretation of Results

In terms of verified certainty scores, the statistical significance of the main effect due to the fidelity variable merely shows that the experimental procedure of degrading the simulated sensor records was sensitive enough to cause an overall decrement in the certainty with which $P(H|D)$ estimates were placed in the correct hypothesis categories. This result, by itself, is neither surprising nor especially important, but it does indicate that experimental control was achieved in a very complex situation. The statistical significance of the main effect due to mode of $P(H|D)$ estimation is very important but difficult to interpret since the observed result seems to imply that human performance was greater than "optimal." A careful interpretation of this result demands at least the following: (a) a more thorough explanation of what is meant by optimal performance, (b) a more careful specification of the situation in which the humans performed, and (c) an examination of the performance scoring procedure.

Bayes' theorem is a statement which is consistent with the definition of a conditional probability. $P(H|D)$, or the probability of some hypothesis after the impact of some datum, is a conditional probability and can be defined in terms of set theory notation as:

$$P(H|D) = \frac{P(H \cap D)}{P(D)} \qquad \text{(Eq. 6)}$$

Now, since $P(H \cap D)$ is equivalent to $P(D \cap H)$ and $P(D \cap H)$ is equal to $P(H)P(D|H)$, $P(H|D)$ can be expressed in terms of conditional probabilities:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \qquad \text{(Eq. 7)}$$

Equation 7 is Bayes' theorem and calculations of $P(H|D)$ on its basis are optimal in the sense that they arise because of formally consistent combinations of probabilities. Presumably, Dodson's modification of Bayes' theorem (equation 2) is similarly consistent in its extension to the situations mentioned in section IV of this report. Now, if humans were optimal in the sense that Bayes' theorem is optimal, they would aggregate probabilities in this formally consistent manner and then reflect this consistency in their $P(H|D)$ estimations. Previous research has indicated that humans are not particularly good at aggregating probabilistic data and that their estimates of $P(H|D)$ are conservative (with respect to Bayes' theorem calculations) because there is more certainty in the data than they can recover (ref. 15). Presumably, maximum certainty could be recovered only by the formally consistent probability aggregation accomplished in Bayes' theorem. The results of the present experiment lead one also to consider situations in which

human estimates of P(H|D) may actually be higher than the certainty in the data justifies.

Appendix V clearly shows the larger number of extreme estimates provided by the humans in each experimental condition. To be examined now are the reasons why these extreme estimates were made. It was mentioned in connection with the results of the preceding experiment that the humans had nothing to lose by trying to maximize the size of their scores. In addition, the tasks of estimating both conditional probabilities P(H|D) and P(D|H) on the basis of large amounts of data were most demanding and arduous. Surely it was an easier task, once a preliminary discrimination among hypotheses had been made, to provide a single high estimate in one category than to adjust meticulously estimates in eight categories.

The verified certainty scoring procedure may be misleading in the light of what has been said about optimal performance. The point is that higher verified certainty scores may not necessarily mean superior performance. Surely, it is just as undesirable to be excessively certain as it is to be overly conservative in one's estimates.

It was specifically hypothesized that the superiority of the MBT solutions of P(H|D) calculated on the basis of human-estimated P(D|H) would be increasingly superior to the human estimates of P(H|D) as fidelity decreased. In terms of the verified certainty scores, this interaction was not statistically significant. One can observe, however, that the smallest difference between the humans and the MBT occurred at the lowest fidelity level. Another lower level may have brought forth MBT superiority. Indeed, this suggestion is substantiated below as the results of the dichotomous score analysis are discussed.

Dichotomous scores, as previously mentioned, show only whether or not the first-choice or highest P(H|D) estimate was placed in the correct hypothesis category. Failure to obtain any statistically significant main effects due to mode of P(H|D) estimation rather confirms the interpretation of the verified certainty results. It also indicates how well, in fact, the humans did perform when compared to the MBT. Although the interaction effect was not significant, the largest difference between human and MBT performance occurred at the lowest fidelity level and was in favor of the MBT. This possibly indicates that with a lower level of fidelity the main experimental hypothesis might have been confirmed. Indeed, an experiment is now in progress to test this notion.

VII. SUMMARY AND CONCLUSIONS

Two experiments have been reported in which human posterior probabilities estimates have been compared with those calculated on the basis of a modification of Bayes' theorem. Human performance was observed in a complex simulated threat-evaluation setting. The major features of the two experiments qualifying their results and affecting generalization of the results are as follows: (a) the critical events under surveillance by the subjects in both experiments were essentially repetitive in nature, and probabilistic statements about these events were essentially frequentistic; (b) the stimulus environment and experimental procedures were designed such that observational uncertainty was induced in the subjects' tasks; (c) individuals, already experienced in dealing with events in the stimulus environment, were used as subjects in both experiments; and (d) those subjects providing posterior probabilities estimates were always informed about the quality of their estimates.

The major results of the experiments were as follows:

1. Human performance at estimating posterior probabilities on the basis of large amounts of data was decidedly superior to what one might have concluded on the basis of previous studies.

2. The role of experience in dealing with environmental events seems to be especially important. With experience the subjects, either by incorporating more data in their judgments or by perceiving more certainty in the data, were increasingly able to place larger posterior probabilities estimates in the correct hypothesis categories.

3. Although human estimates of posterior probabilities may be conservative on occasion because of suboptimal extraction of certainty from probabilistic data, the present experiments lead one to recognize that there are situations in which humans may provide higher estimates than the certainty in the data justifies. Several reasons for these excessive estimates, specific to the present experimental situation, were discussed in the preceding section.

4. Although a decrease in input data fidelity appears to play a significant role in depressing the size of posterior probability estimates, the hypothesis of increased superiority of the MBT under lower levels of fidelity was not substantiated by the data. Further levels of data degradation below those actually used in the experiment may have brought forth the predicted superiority of the MBT solutions over the human estimates of posterior probabilities. The results of Experiment II seem to justify further research on this issue.

5. It is apparent that more than one type of scoring procedure is needed if one expects to get an accurate account of performance in these tasks. Without the use of the dichotomous

scores in the first experiment one might have tended to devaluate the overall performance of the MBT calculations on the basis of the human estimates of P(D|H). In the second experiment, the verified certainty scores by themselves may have been similarly misleading.

There are two classes of conclusions which may be drawn from these experiments. The first class relates to the implications of the present research for the design of threat-evaluation or other diagnostic systems and the other class refers to the implications of the present research for the general study of human commerce with probabilistic information. With respect to the design of diagnostic systems, the present research tends to confirm the notion that automated Bayesian hypothesis selection on the basis of expert human estimation of the conditional probabilities P(D|H) may well prove useful in systems with a diagnostic mission. Although the present research hopefully gives a better idea about what to expect in real-life systems from experienced people, the subjects in the present experiment performed under no time stress, nor was there a meaningful cost-payoff arrangement. The frequentistic nature of the probabilistic data being processed by the simulated system suggests that the present results may be important for the medical diagnosis situation in which a frequency notion of probability is often indicated. Many military diagnosis systems must deal with environmental processes which are not specifiable in terms of long-run frequencies. For this reason, Bayesian automated hypothesis selection as a diagnostic aid needs to be evaluated in nonfrequentistic environments. Indeed, the Engineering Psychology Laboratory at the University of Michigan will soon have the facility for evaluating human performance in these situations.

Finally, one may ask what the comparison of human and MBT posterior probability estimates suggests as far as the study of human commerce with probabilistic information is concerned. Although there were some striking similarities between the MBT and human performance, it seems clear that analogous processes were not involved. The response profiles for the better subjects were almost identical to those of the MBT solutions calculated on the basis of their P(D|H) estimates. In addition, the better subjects used a very large amount of the input data. The performance of all subjects was moderately though significantly correlated with that for both types of MBT calculation. However, calculation of the MBT solutions involved an iterative and consistent aggregation of all that available data. The human P(H|D) estimates, on the other hand, seem to have been based upon the following process. The input data set was first "filtered" or "shrunken" in an effort to reduce the complexity of the situation. The retained data were then used to make preliminary and final discriminations among the available hypotheses. The final step in the process involved assigning numbers to indicate the subject's certainty that the discrimination had been an accurate one.

APPENDIX I

ATTRIBUTE DATA CLASSES

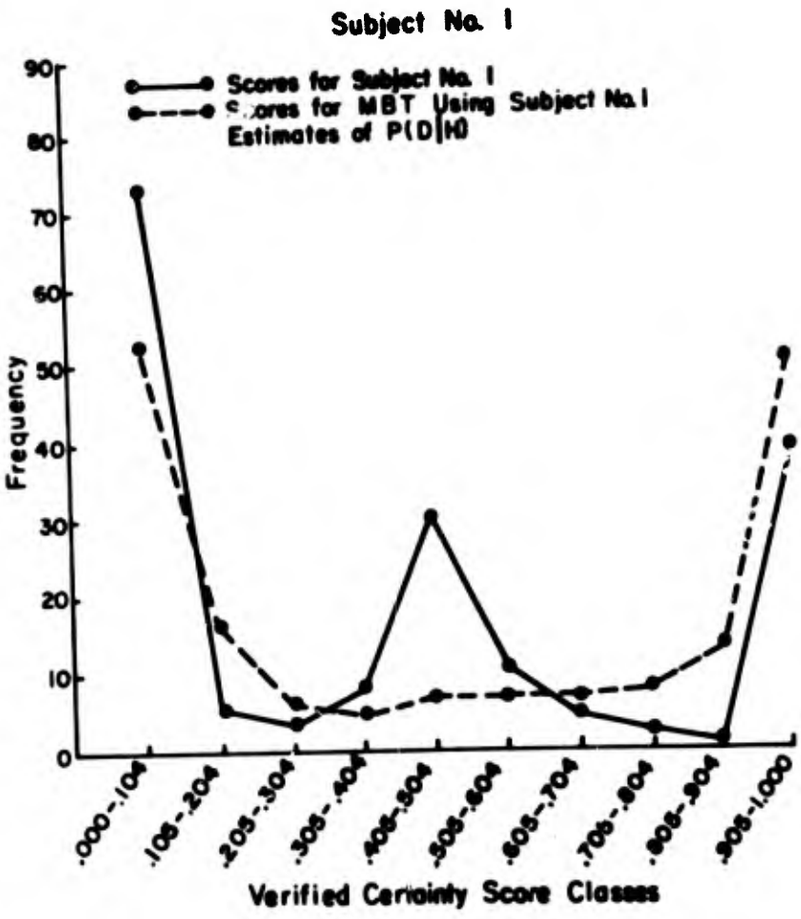
Data Class	Number of Possible States	Description
I. Mechanized Rifle Battalions	8	The states or levels of data classes I through XII all refer to numbers of battalions or squadrons of the type indicated by the various data class labels indicated in column 1. The first level in every data class refers to zero battalions or squadrons.
II. Medium Tank Battalions	7	
III. Heavy Tank Battalions	6	
IV. Artillery Battalions (range up to 10,000 meters)	3	
V. Artillery Battalions (range up to 20,000 meters)	3	
VI. Artillery Battalions (range up to 30,000 meters)	3	
VII. Rocket Battalions	3	
VIII. Intermediate Range Ballistic Missile Battalions	3	
IX. Ground Reconnaissance Battalions	2	
X. Tactical Air Support Squadrons	4	
XI. Aerial Reconnaissance Squadrons	4	
XII. Surface to Air Missile Battalions	4	

Data Class	Number of Possible States	Description
XIII. Units of Fire for Infantry and Armored Units (Main Attack Forces)	4	This data class refers to the amount of ammunition being carried by road or rail convoys which provide logistics support for infantry and armored units.
XIV. Units of Fire for Artillery Missile, and Rocket Units (Combat Support Forces)	4	Refers to the amount of ammunition being carried by supply convoys for these three classes of units.
XV. Dispersal Distance between Supply Units	4	Distance in miles between terminal positions of supply convoys.
XVI. Supply Timing for Main Attack Units	3	Refers to temporal order of appearance of supply units and units being supplied.
XVII. Supply Timing for Combat Support Units	3	Same as XVI.
XVIII. Terminal Activity Zone	5	Refers to the distance in miles from the border of the most forward units in a developmental grouping.
XIX. Terminal Activity Development Pattern	4	Refers to the configuration or placement of units laterally along the border of contention after these units have reached their terminal positions.

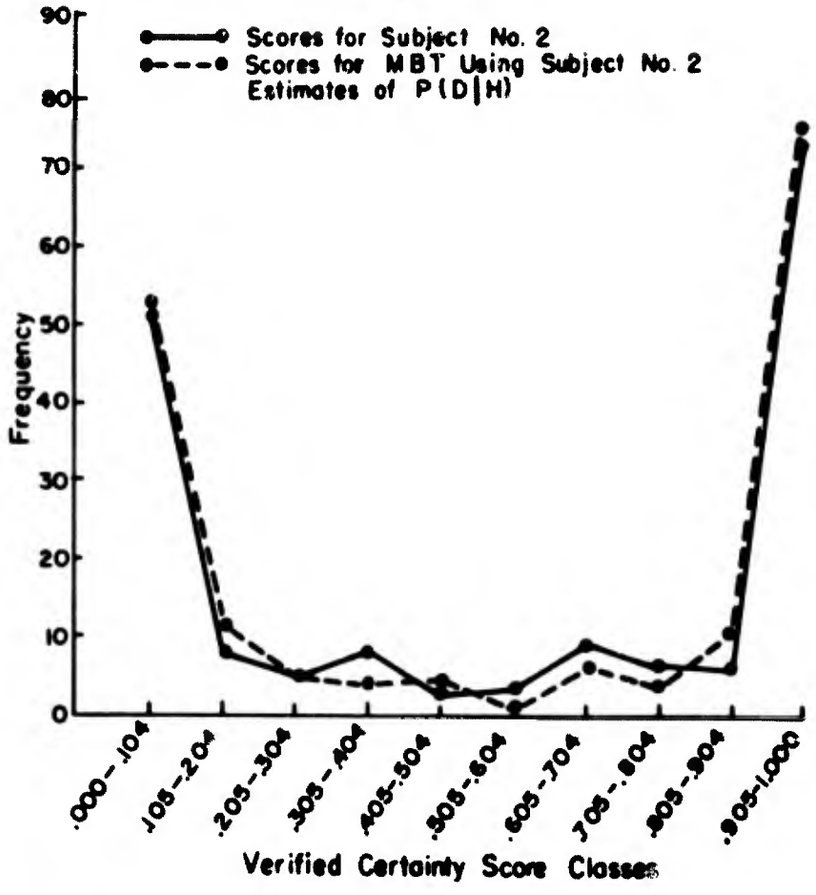
Data Class	Number of Possible States	Description
XX. Attack Position Lateral Dispersion	5	Dispersal distance in miles along a border of contention of an entire developmental grouping.
XXI. Attack Position Depth	4	Distance in miles involved in the placement of forces perpendicular to a border, i.e., the distance between the most forward unit in a grouping and rearmost unit.
XXII. Attack Buildup Timing	3	Refers to the temporal order of appearance at terminal positions along a border of contention of main attack units and combat support units.
XXIII. Transportation Methods	3	Refers to the combination of road, rail, and air facilities used to transport Aggressor units in any developmental grouping.
XXIV. Ground Transportation Speed Class	5	Road and rail convoy speed during the buildup of a developmental grouping.
XXV. Developmental Period	6	Length of time in days from beginning to termination of a buildup of a developmental grouping.

APPENDIX II
DISTRIBUTIONS OF VERIFIED CERTAINTY SCORES IN EXPERIMENT I

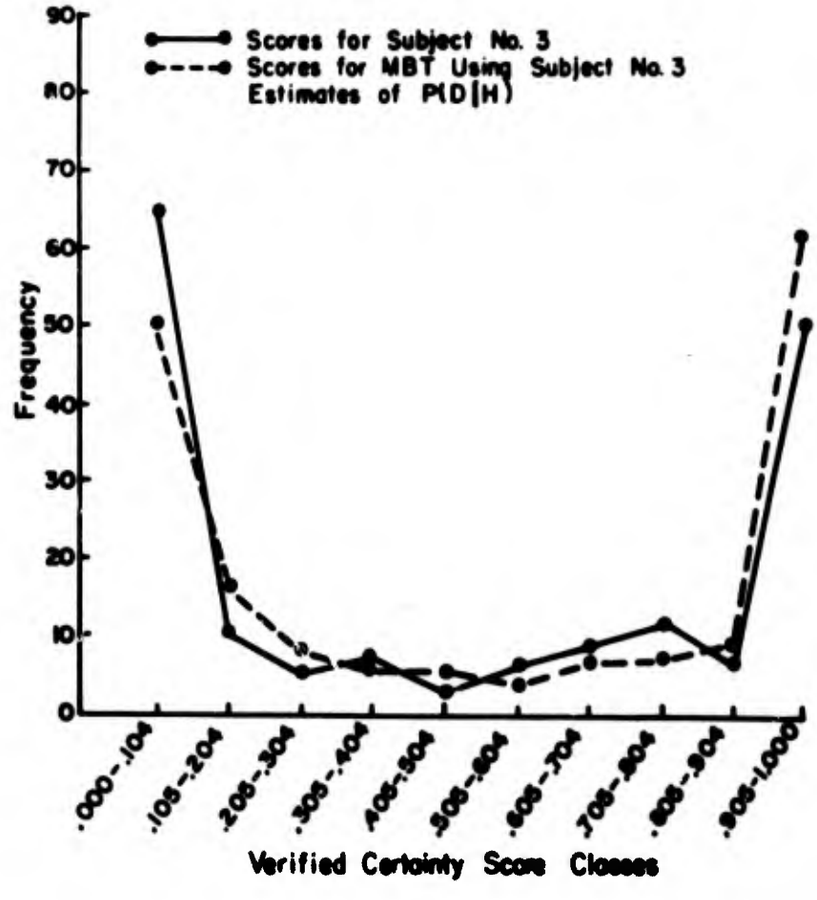
The following four graphs illustrate the distributions of verified certainty scores in Experiment I for each of the four threat evaluators and the MBT solutions incorporating their P(D|H) estimates.

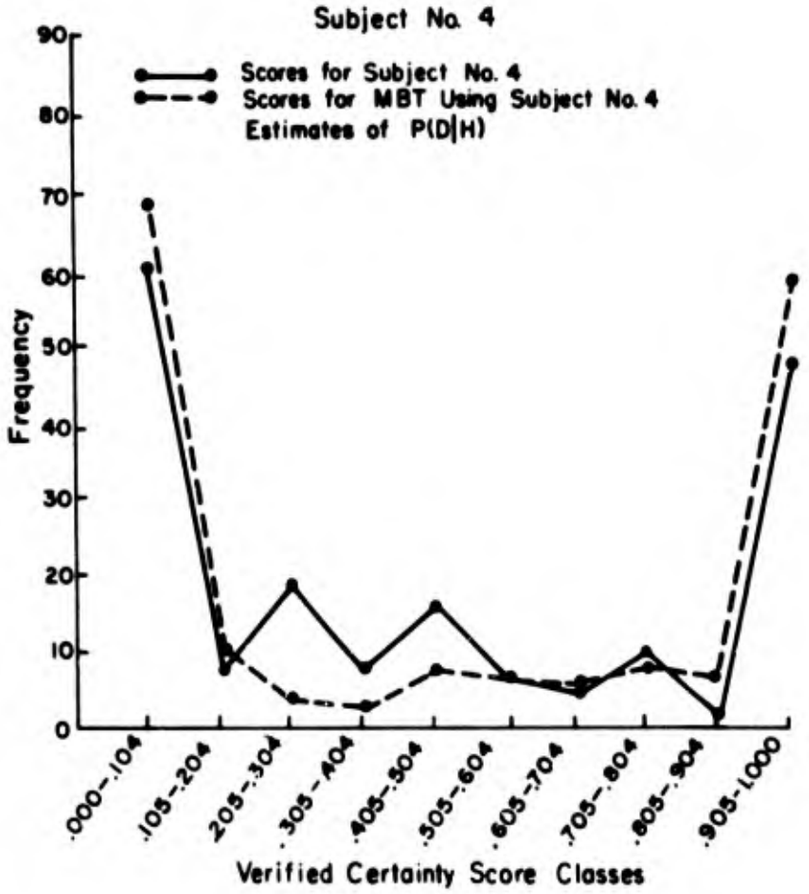


Subject No. 2



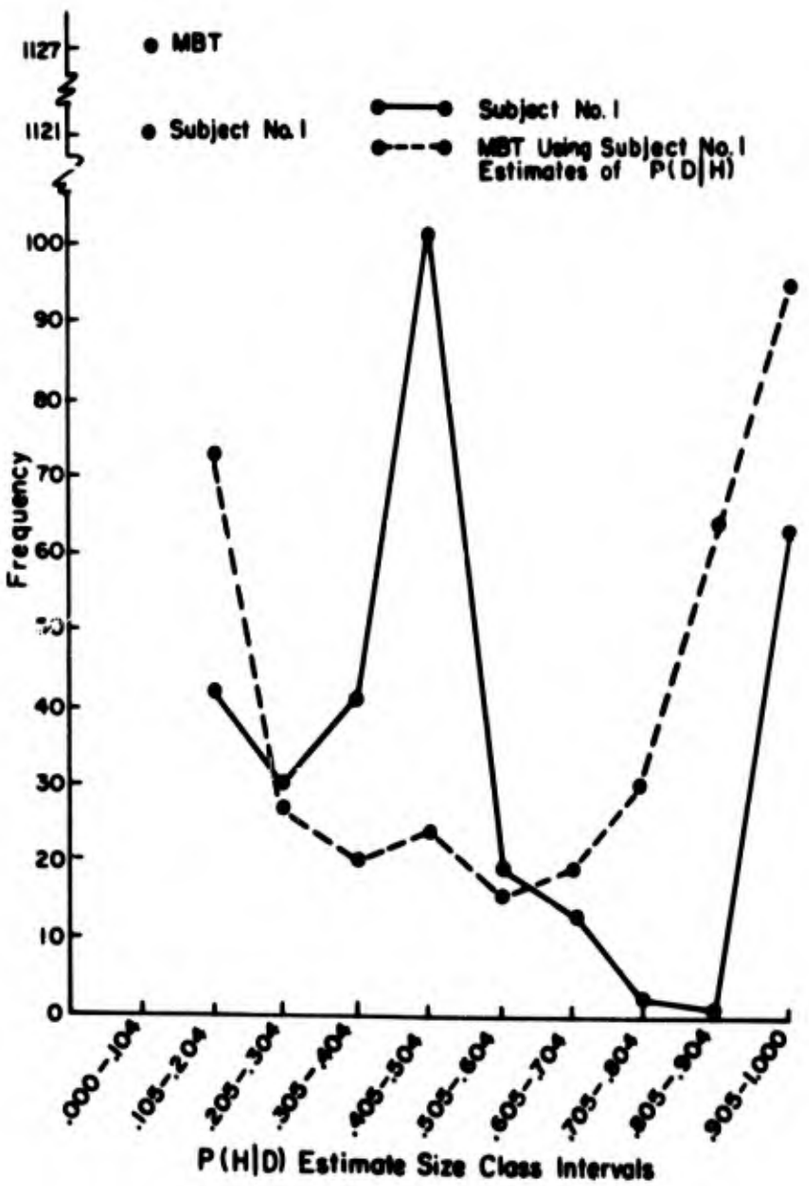
Subject No. 3

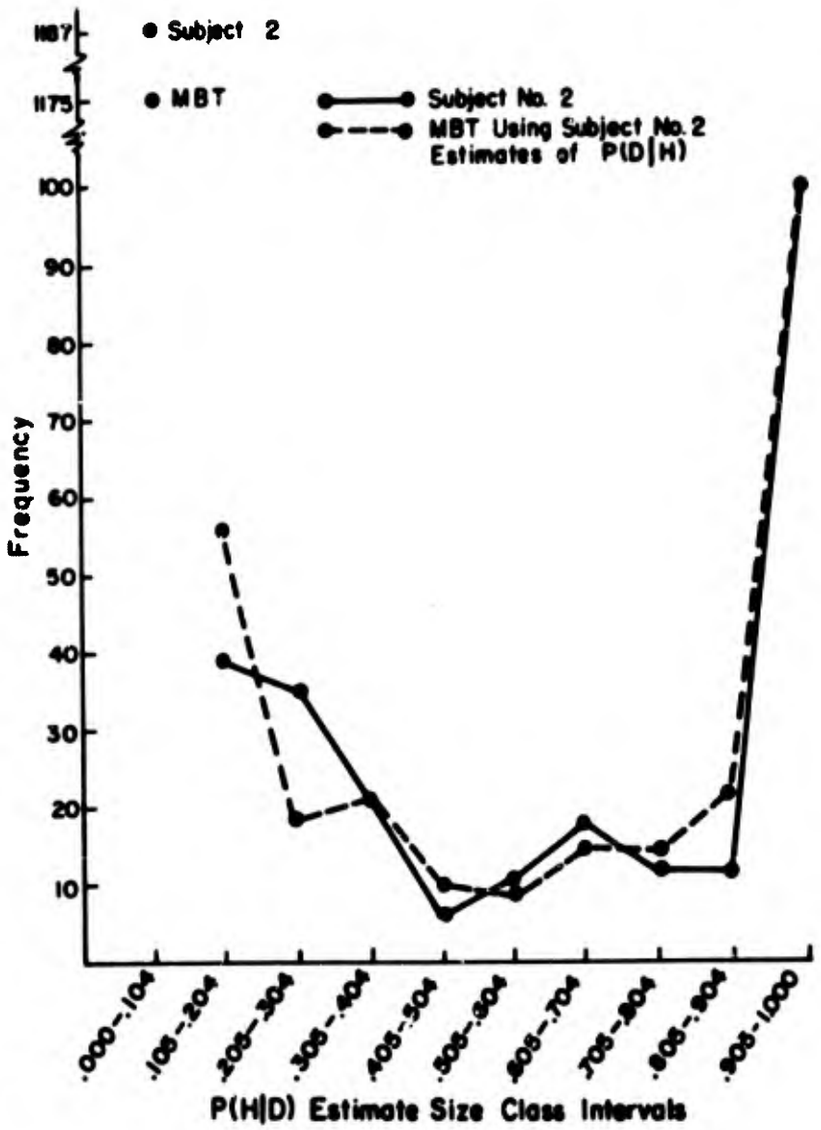


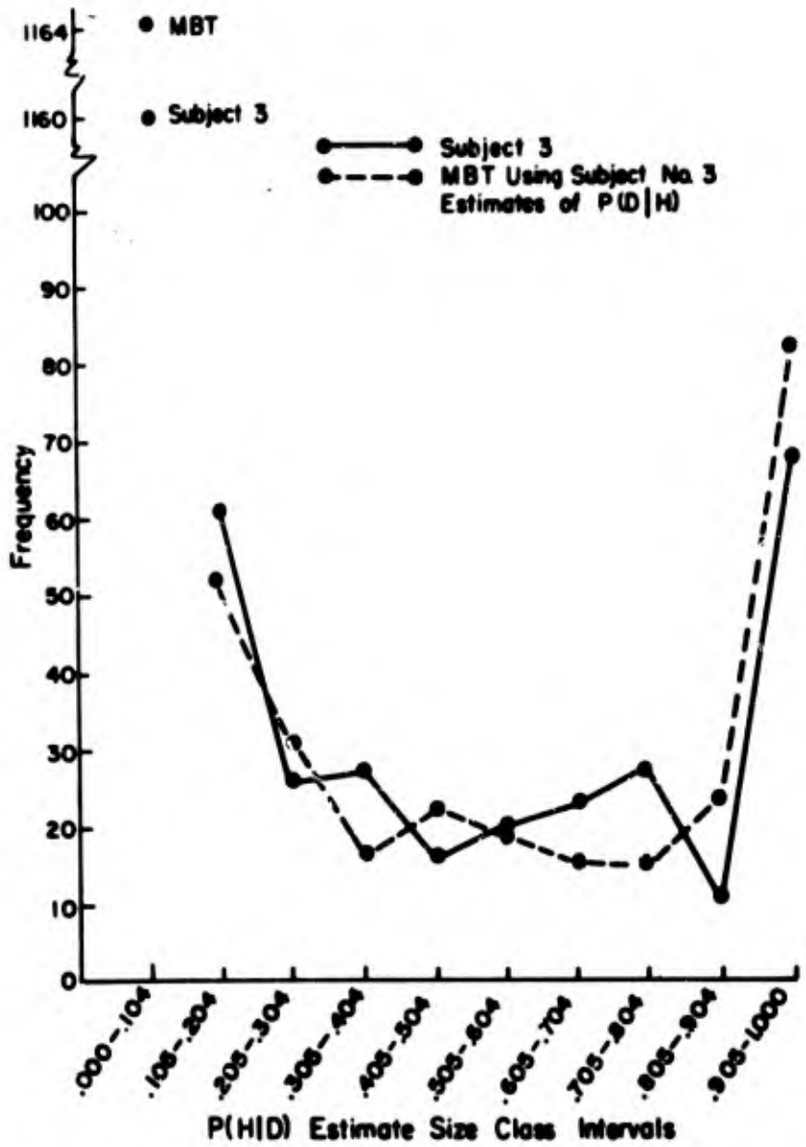

Subject No. 4
Scores for Subject No. 4
Scores for MBT Using Subject No. 4
Estimates of P(D|H)
Frequency
90
80
70
60
50
40
30
20
10
0
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Verified Certainty Score Classes

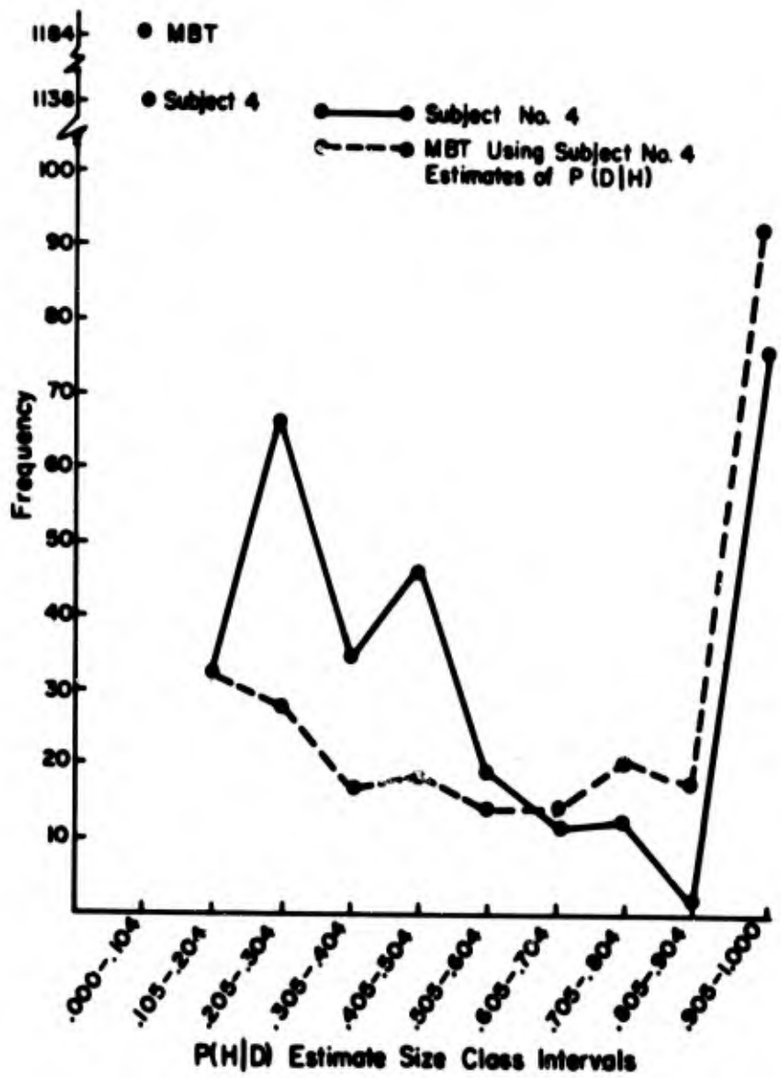
APPENDIX III
DISTRIBUTIONS OF P(H|D) ESTIMATES MADE IN EXPERIMENT I

The following four graphs illustrate the distributions of all P(H|D) estimates made in Experiment I by each threat evaluator and the MBT solution calculated on the basis of his P(D|H) estimates.



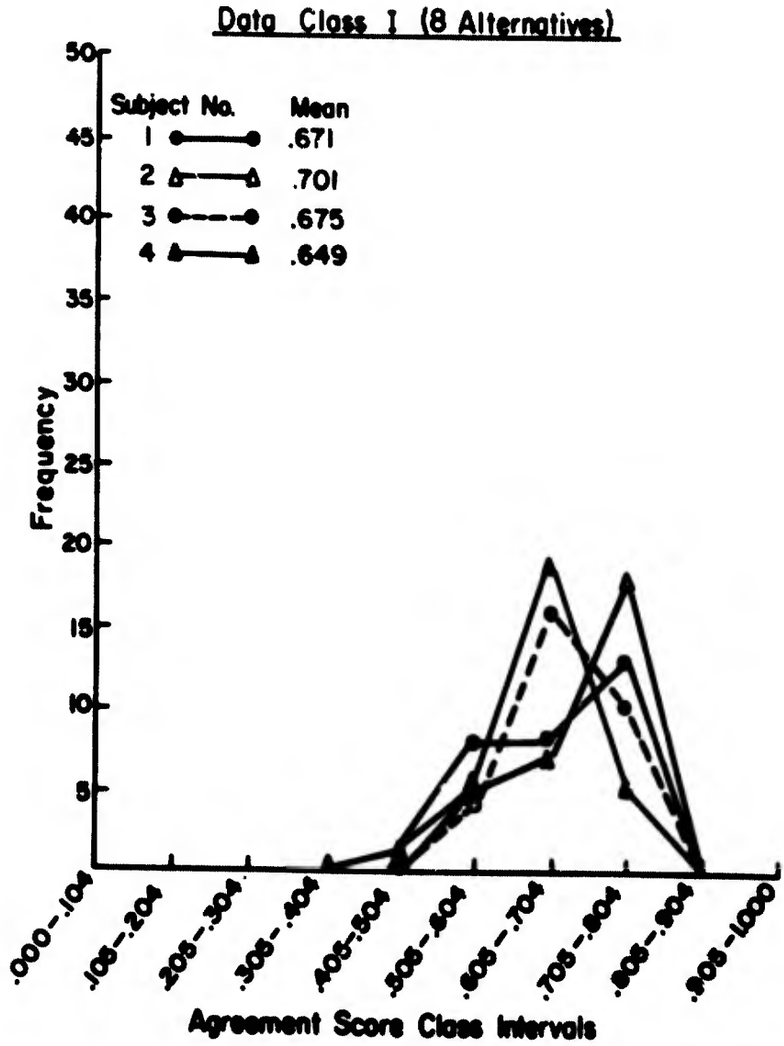

1187
1175
● Subject 2
● MBT
Subject No. 2
MBT Using Subject No. 2 Estimates of P(D|H)
100
90
80
70
60
50
40
30
20
10
Frequency
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
P(H|D) Estimate Size Class Intervals

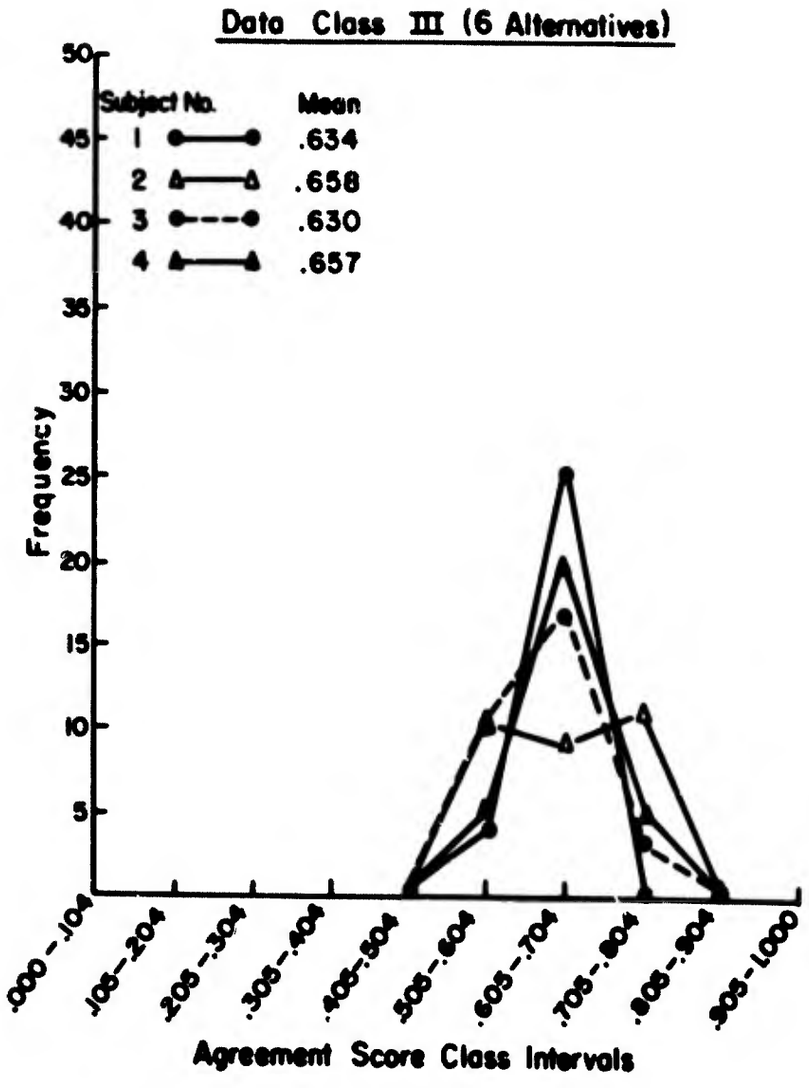

1164
1160
100
90
80
70
60
50
40
30
20
10
0
Frequency
● MBT
● Subject 3
Subject 3
MBT Using Subject No. 3 Estimates of P(D|H)
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
P(H|D) Estimate Size Class Intervals

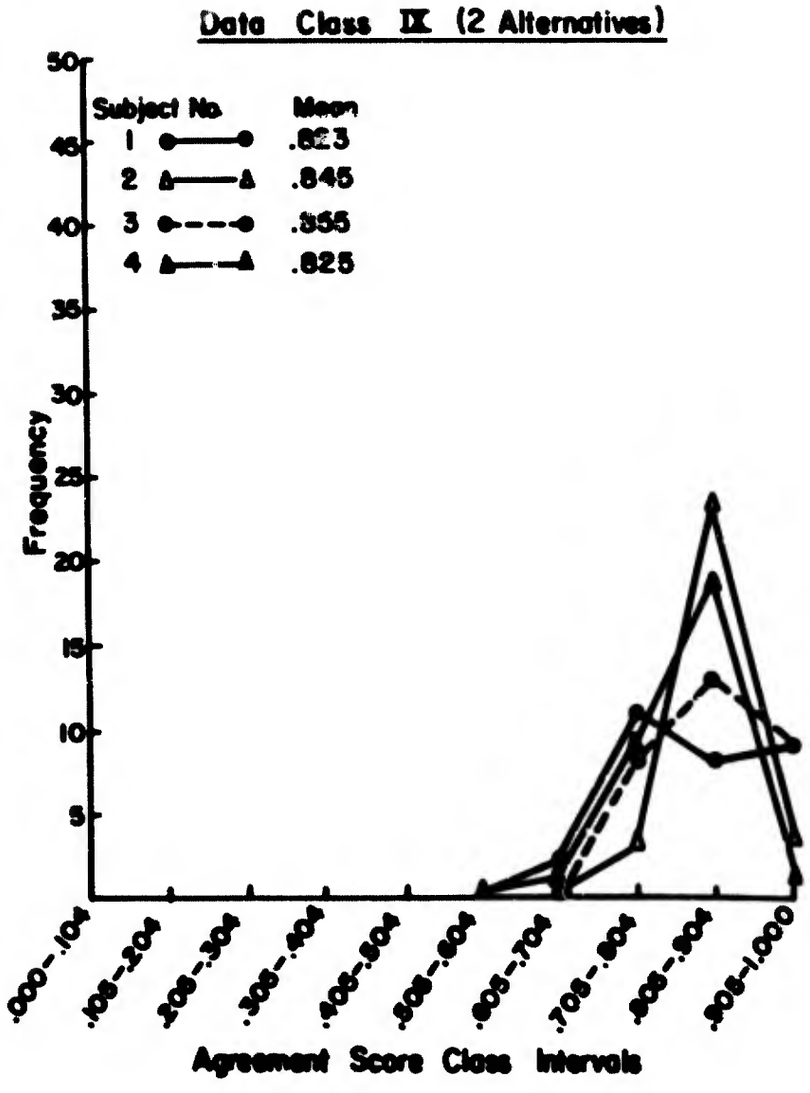

1184
1138
100
90
80
70
60
50
40
30
20
10
Frequency
● MBT
● Subject 4
Subject No. 4
MBT Using Subject No. 4 Estimates of P (D|H)
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
P(H|D) Estimate Size Class Intervals

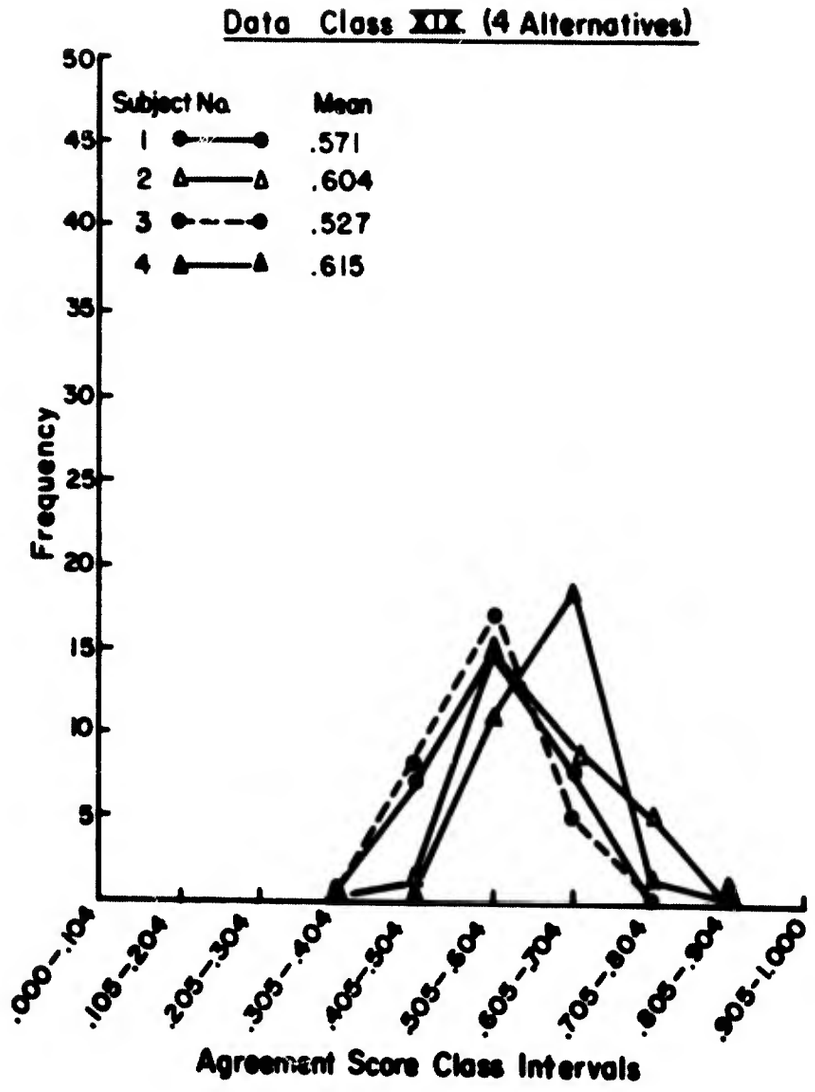
APPENDIX IV
DISTRIBUTIONS OF AGREEMENT SCORES IN EXPERIMENT I

The following four graphs illustrate representative distributions of agreement scores (a_j) in Experiment I for each of the four threat evaluators' P(D|H) estimates.



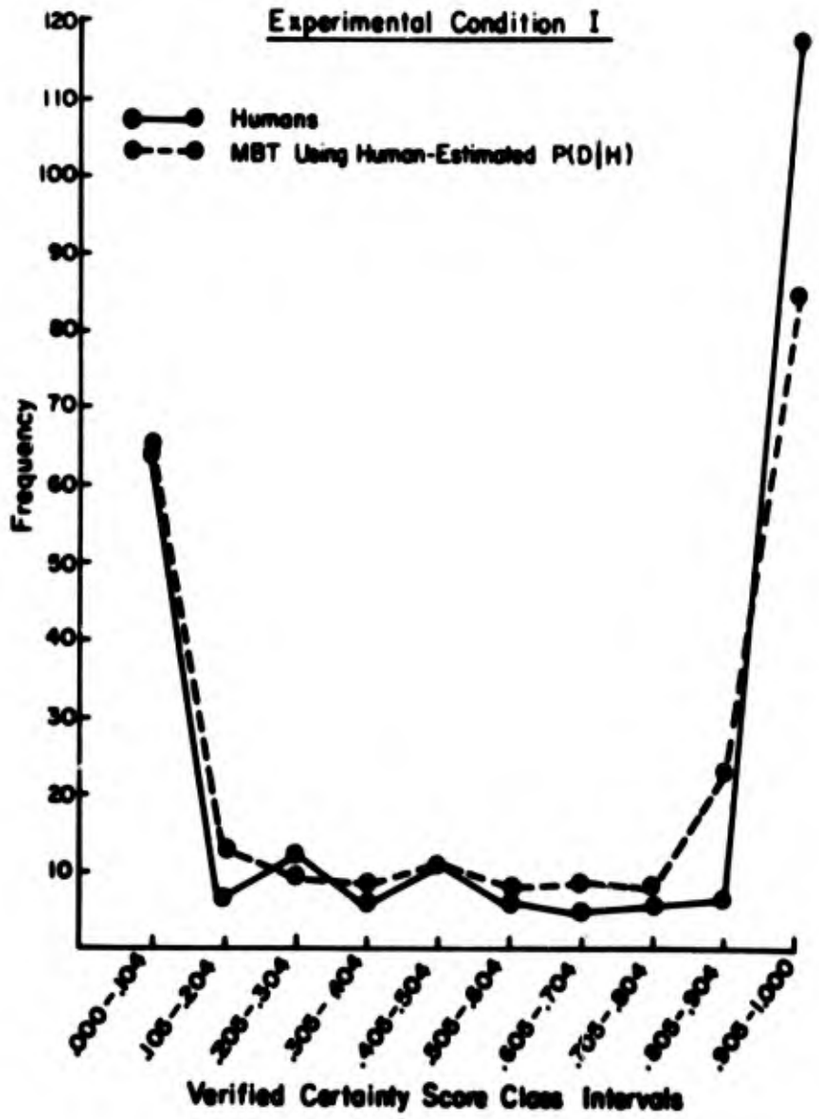

Data Class III (6 Alternatives)
Subject No.
Mean
1 .634
2 .658
3 .630
4 .657
Frequency
50
45
40
35
30
25
20
15
10
5
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Agreement Score Class Intervals

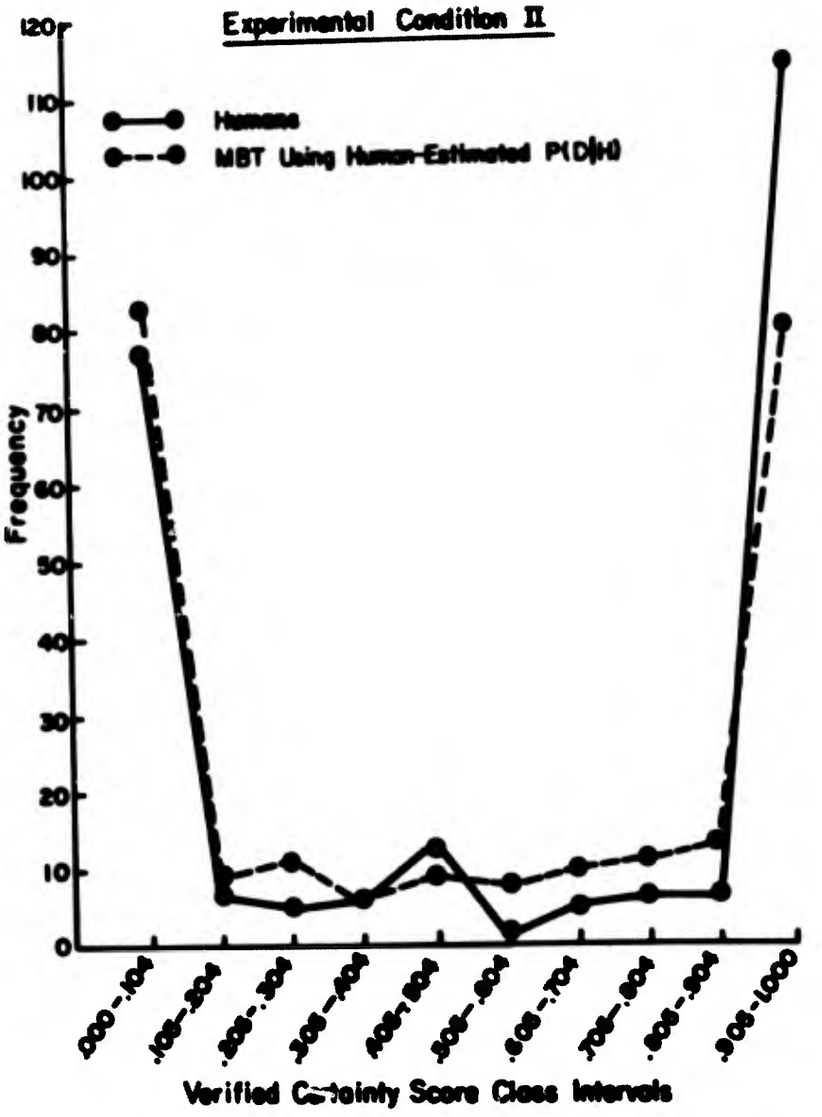

Data Class IX (2 Alternatives)
Subject No.
Mean
1 .823
2 .845
3 .855
4 .825
Frequency
50
45
40
35
30
25
20
15
10
5
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Agreement Score Class Intervals

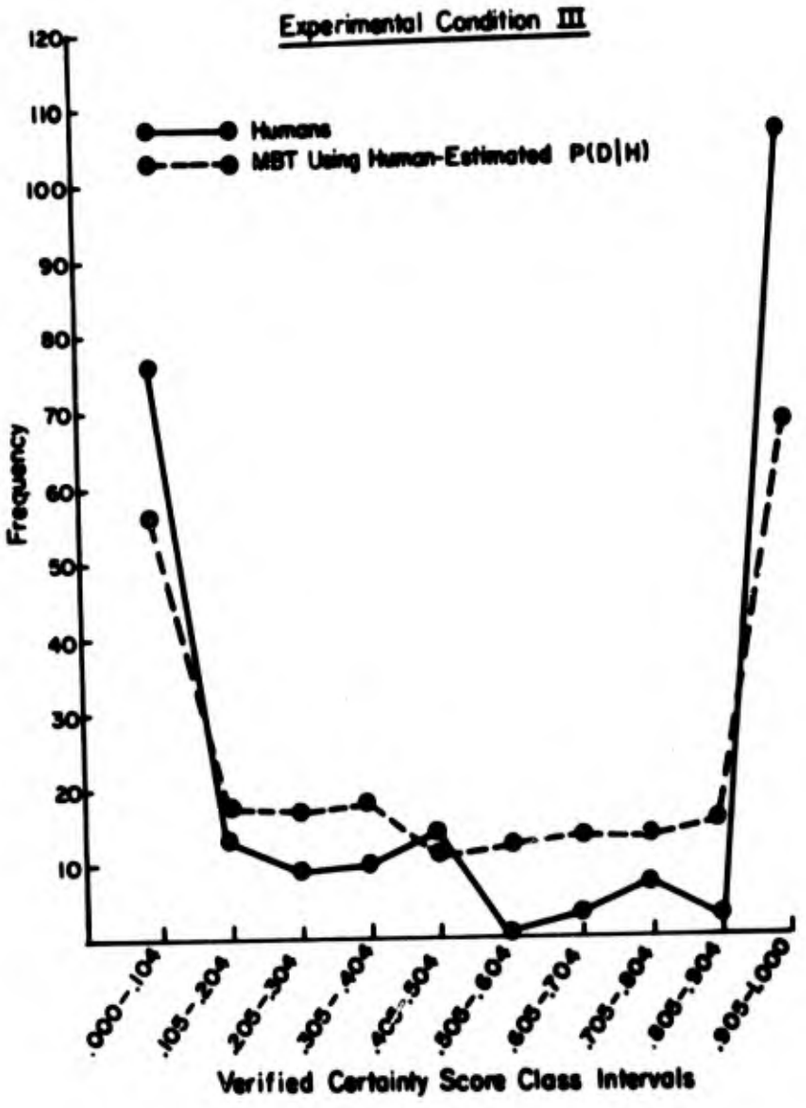

Data Class XIX (4 Alternatives)
Subject No.
Mean
1 .571
2 .604
3 .527
4 .615
Frequency
50
45
40
35
30
25
20
15
10
5
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Agreement Score Class Intervals

APPENDIX V
DISTRIBUTIONS OF VERIFIED CERTAINTY SCORES IN EXPERIMENT II

The following three graphs illustrate the distributions of verified certainty scores in each condition of Experiment II for the humans (as a group) and the MBT solutions calculated on the basis of their P(D|H) estimates.

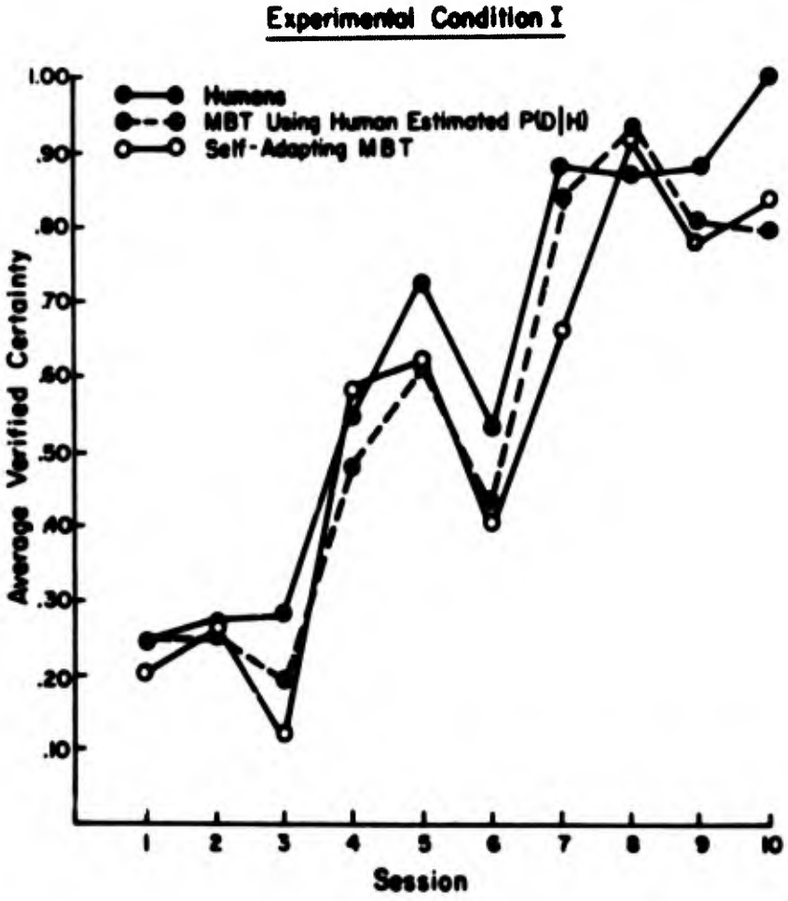



Experimental Condition II
Humans
MBT Using Human-Estimated P(D|H)
Frequency
120
110
100
90
80
70
60
50
40
30
20
10
0
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Verified Certainty Score Class Intervals

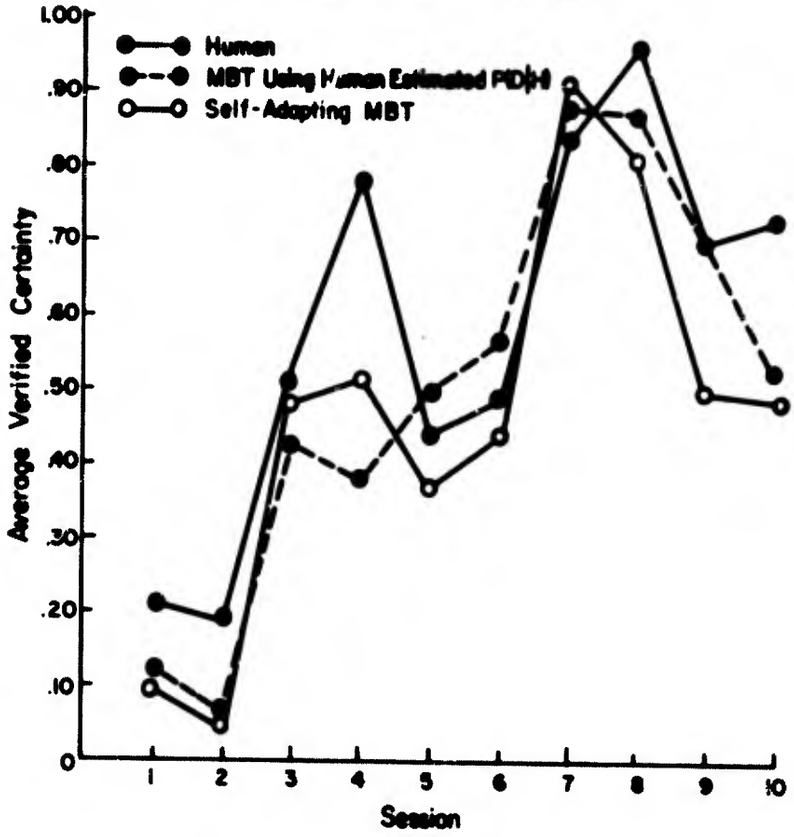

Experimental Condition III
Humans
MBT Using Human-Estimated P(D|H)
Frequency
120
110
100
90
80
70
60
50
40
30
20
10
.000-.104
.105-.204
.205-.304
.305-.404
.405-.504
.505-.604
.605-.704
.705-.804
.805-.904
.905-1.000
Verified Certainty Score Class Intervals

APPENDIX VI
VERIFIED CERTAINTY SCORES IN EXPERIMENT II

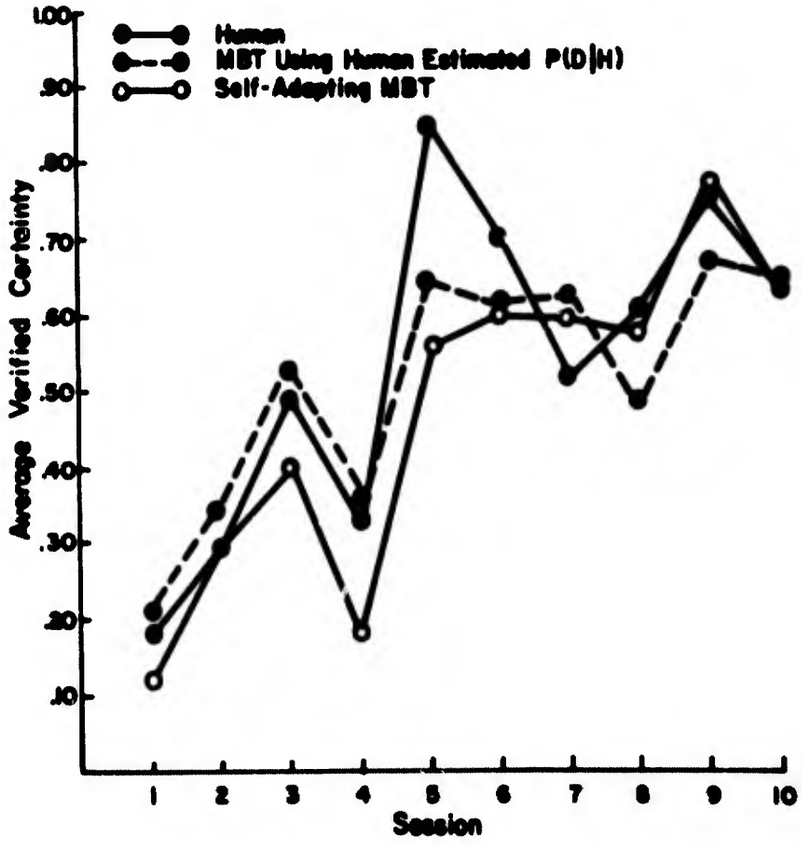
The following three graphs illustrate human and MBT performance in terms of verified certainty scores for each of the three fidelity conditions in Experiment II.



Experimental Condition II



Experimental Condition III



REFERENCES

1. Chernoff, H., and L. E. Moses, Elementary Decision Theory, John Wiley and Sons, Inc., New York, N. Y., 1959.

2. Shuford, E. H., Some Bayesian Learning Processes, Division of Mathematical Psychology, Institute for Research, State College, Penna., Report No. 2.

3. Bayes, T., "An Essay towards Solving a Problem in the Doctrine of Chances," Philosophical Transactions of the Royal Society, Vol 53, pp 370-418, 1763.

4. Good, L. J., "Kinds of Probability," Science, Vol 129, pp 443-446, 1959.

5. Fisher, R. A., The Design of Experiments, 7th Edition, Hafner Publishing Company, New York, N. Y., 1960.

6. Uspensky, J. V., Introduction to Mathematical Probability, McGraw-Hill Book Company, Inc., New York, N. Y., 1937.

7. Savage, L. J., The Foundations of Statistics, John Wiley and Sons, Inc., New York, N. Y., 1954.

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

9. Savage, L. J., "Bayesian Statistics," Decision and Information Processes, The Macmillan Company, New York, N. Y., pp 161-194, 1962.

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

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

12. Edwards, W., Probabilistic Information Processing in Command and Control Systems, ESD Technical Documentary Report 62-345, Electronic Systems Division, L. G. Hanscom Field, Bedford, Mass., March 1963.

13. Edwards, W., and L. Phillips, Man as a Transducer for Probabilities in Bayesian Command and Control Systems, presented at the annual meeting of the American Association for the Advancement of Science, Philadelphia, Penna., December 1962.

14. Southard, J., D. A. Schum, and G. E. Briggs, An Application of Bayes' Theorem as a Hypothesis-Selection Aid in a Complex Information-Processing System, AMRL Technical Documentary Report 64-51, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio, August 1964.

15. Edwards, W., Probabilistic Information Processing by Men, Machines, and Man-Machine Systems, SDC Technical Memorandum 1418/000/01, System Development Corporation, Santa Monica, Calif., August 1963.

16. Kaplan, R. J., and J. R. Newman, A Study in Probabilistic Information Processing, SDC Technical Memorandum 1150, System Development Corporation, Santa Monica, Calif., April 1963.

17. Southard, J. F., D. A. Schum, and G. E. Briggs, Subject Control Over a Bayesian Hypothesis-Selection Aid in a Complex Information-Processing System, AMRL Technical Report 64-95, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio, September 1964.

18. Siegel, S., Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Company, Inc., New York, N. Y., 1956.

19. Bradley, J. V., Distribution-Free Statistical Tests, WADD Technical Report 60-661, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio, August 1960.