AMRL-TR-64-95

SUBJECT CONTROL OVER A BAYESIAN HYPOTHESIS-SELECTION AID IN A COMPLEX INFORMATION-PROCESSING SYSTEM

JACK F. SOUTHARD
DAVID A. SCHUM
GEORGE E. BRIGGS

OHIO STATE UNIVERSITY

SEPTEMBER 1964

BEHAVIORAL SCIENCES LABORATORY
AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
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.

Qualified requesters may obtain copies from the Defense Documentation Center (DDC), Cameron Station, Alexandria, Virginia 22314. Orders will be expedited if placed through the librarian or other person designated to request documents from DDC (formerly ASTIA).

Do not return this copy. Retain or destroy.

Stock quantities available at Office of Technical Services, Department of Commerce, Washington, D. C. 20230.

Change of Address

Organizations receiving reports via the 6570th 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.

800 - November 1964 - 448-11-311

BLANK PAGE

SUBJECT CONTROL OVER A BAYESIAN HYPOTHESIS-SELECTION AID IN A COMPLEX INFORMATION-PROCESSING SYSTEM

JACK F. SOUTHARD
DAVID A. SCHUM
GEORGE E. BRIGGS

FOREWORD

The research described in this report 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 June 1963 to 15 April 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, Human Engineering Division, Behavioral Sciences Laboratory.

The authors are pleased to acknowledge the invaluable assistance rendered by several members of the laboratory staff. Messrs. Edwin R. Lassettre and Lonnie D. Whitehead assumed a major share of the responsibility for designing and preparing the large amount of computer programing required in this research project. These two individuals must be accorded a primary share of the credit for the enormous programing effort which was necessary to generate the complex stimulus environment described in the report. Miss Carolyn Black assisted in the collection and analysis of data. Miss Janis Frye and Miss Barbara Lindig provided invaluable editorial assistance.

This technical report has been reviewed and is approved.

WALTER F. GRETHER, PhD
Technical Director
Behavioral Sciences Laboratory

ABSTRACT

This report describes the second experiment in a series devoted to estimating the effectiveness of automated hypothesis selection in man-machine systems in which threat evaluations or threat diagnoses are being performed. In the experiment an eight-man team produced evaluations of various threats posed by a hypothetical aggressor. These evaluations were made on the basis of intelligence information gathered on simulated reconnaissance overflights of the homeland area of the aggressor. IBM 1401 and 7090 computer facilities provided the means for generating the complex stimulus environment or data base. The primary output from this threat evaluation team was a series of a posteriori probabilities estimations produced by the team's commanding officer (CO). These estimations represented the CO's judgments as to the most likely of the four response alternatives available to aggressor in deploying his forces along a border of contention. In three of the four experimental conditions the CO was provided with a hypothesis-selection aid based upon a modification of Bayes' theorem (MBT). In these three conditions the CO was permitted to exert an increasing amount of control over the MBT-aid mechanism. He exerted control either by adjustment of certain parameters in the MBT model or by direct insertion of conditional probabilities into the model. The purpose of the experiment was to observe whether increasing control over the MBT-aid mechanism would increase the user's acceptance of the aid and improve his threat-diagnosis performance. The CO's threat-evaluation performance did improve during the course of the experiment but independently of the MBT-aid configuration. Throughout the experiment, solutions of a posteriori probabilities based upon the MBT were calculated by the experimenter for comparison with the human estimates. These two sets of estimates were strikingly similar. The CO's estimates, although quite conservative in early trials, became noticeably less conservative as the experiment progressed. The overall difference between the accuracy of the CO and MBT estimations was negligible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

BLANK PAGE

I. INTRODUCTION

Inferences about the present or impending posture of hostile military forces, predictions of the strategy to be used in the deployment of these forces, and discriminations relative to the existence of certain classes of weapons or vehicles are examples of the types of judgments required of individuals who evaluate intelligence information. Essentially, each of these judgments involves selection of one or several hypotheses which best account for the occurrence of fragments of intelligence data often contradictory and always fallible to an unknown degree. In an age of ever-increasing weapons system sophistication, 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. In certain military information-processing systems the diagnostic requirements are indeed formidable and seem to have entirely outstripped human capabilities. Specifically, the range of possible hypotheses which account for environmental events may be large; the set of input data to be evaluated may be large and of a diverse character; the environmental rules or contingencies which relate these data to the alternative hypotheses may be exceedingly abstruse or unknown; and relatively instantaneous diagnoses may be required.

The possibility of automating certain aspects of the threat-diagnosis function in complex information-processing systems (in which the data to be evaluated are probabilistic in nature) has been suggested by Edwards (refs. 2, 3, 4) and Dodson (ref. 1). Computer-implemented solutions of a posteriori probabilities (estimates of the probability of the various alternative hypotheses in the light of new data) based upon Bayes' theorem can readily be provided either as aids to a human threat-evaluator who makes the final hypothesis selection or as final hypothesis-selection mechanisms themselves. The formal justification for these procedures rests upon the notion that Bayes' theorem provides an optimal method for revising opinions (or selecting hypotheses) on the basis of experience because it allows maximum certainty or consistency to be extracted from the probabilistic data at hand. The empirical justification rests upon a limited number of studies which tend to show that humans are unable to extract all of the certainty existing in probabilistic data and that, in general, their estimates of posterior probabilities are quite conservative (refs. 3, 5).

The multiman-machine systems simulation facility at The Ohio State University Laboratory of Aviation Psychology has been adapted to provide the vehicle necessary to evaluate certain aspects of the Bayesian paradigm for information-processing systems. Computer facilities (IBM 7090 and 1401) permit generation of a complex, real-time stimulus environment simulating the movements of the surface and air forces of a hypothetical aggressor in his homeland area. Up to 20 basic strategies or response alternatives are available to the aggressor in maneuvering and deploying his forces. Twenty-five classes of data (see Appendix I) are used to describe the attributes or characteristics of these deployments which individually are called "developmental groupings." There exists a set of contingency rules relating the data classes to the response alternative set. These features of the stimulus environment are described in detail in the first report in

the current series (ref. 6). A team of eight system operators attempts to evaluate each developmental grouping as it unfolds in the simulated hostile environment. Certain members of this team who are called "intelligence staff officers" (ISOs) attempt to extract from the environment, by means of simulated reconnaissance overflights, information which will enable them to make inferences about the state or condition existing in each of the 25 data classes with respect to each of the developmental groupings under surveillance. These data classes refer, in general, to such developmental grouping features as (a) infantry-armoured constituency, (b) artillery, missile, rocket, and air support, (c) logistic support, and (d) spatial and temporal arrangement of forces (order of battle). Each of these 25 data classes has between two and eight possible states or conditions, only one of which applies in a particular developmental grouping. The ISOs estimate the probability that a data class (j) is in state or level (k) for each developmental grouping. On the basis of these probabilistic judgments regarding each of the 25 data classes for every developmental grouping, the commanding officer (CO) of the evaluation team estimates the probability that each of the aggressor response alternatives (hypotheses) could account for the data observed in connection with each of the groupings. The CO, therefore, performs the hypothesis-selection tasks of primary concern in the present research. His responses are posterior probabilities estimates (in this case, estimates of the probability of the several aggressor response alternatives in the light of the attribute data). A detailed explanation of the individual tasks performed by each team member is also provided in the first research report.

The experiment described in the present report is the second of a series devoted to estimating the effectiveness of a Bayesian paradigm for information-processing systems where hypothesis-selection or diagnostic functions are performed. In the experiment the user of the Modified Bayes Theorem (MBT) hypothesis-selection aid (the CO) was given an increasing amount of control over the aid mechanism itself. The purpose of the experiment was to observe the effects of this increased control capability upon the CO's a posteriori probability estimates. The particular features which allow this control increase are described in the following section of the report. The rationale for treating increased MBT-aid control as a variable is as follows: Inspite of compelling evidence for the usefulness of the MBT as a hypothesis-selection aid, the individual or individuals for whom this aid is intended may be quite reluctant to rely upon the probabilistic data which the aid provides. Such reluctance is commonly found when an attempt is made to introduce new equipment or techniques to operational personnel. The experimental issue which was raised in this connection concerns the method by which acceptance and use of a hypothesis-selection aid could be increased in a multiman-machine system context. One possibility is that acceptance and use of the aid will increase as greater control over the aid is given to its user. With little or no control over the hypothesis-selection aid, the user is likely to view it merely as a "black box" whose sensitivities are not apparent. The likelihood of acceptance of the aid under such a condition would perhaps be fairly low. In order to control the MBT, the user would certainly be required to understand both the principles involved in its use and the effects of the changes which he could introduce at several points in the model itself. Given sufficient understanding of the MBT so that he might make a rational comparison of the capabilities

and limitations of the MBT with his own, the user may be more likely to accept it as an aid.

The concept of the MBT as a hypothesis-selection aid has merit if we suppose that human beings will be required to provide the final diagnoses or estimates of a posteriori probabilities. Another possibility, suggested by Edwards (refs. 2, 3), is that estimates of a posteriori probabilities be provided by a computer-implemented Bayesian solution on the basis of contingent datum-hypothesis relationships [P(D|H)] estimated by humans. If these P(D|H) values can be judged with reasonable accuracy by humans, the diagnoses or a posteriori estimates follow by simple algebra on the basis of Bayes' theorem and can be provided more rapidly and accurately by computer facilities. The experiment being reported is actually related to both of these suggested paradigms for Bayesian systems. First, the effects of the MBT aid upon the user's a posteriori estimates can be observed. In addition, at least a preliminary judgment can be made regarding the automated paradigm suggested by Edwards because several parallel calculations of the a posteriori probabilities based upon the MBT will be available for comparison with the CO's estimates. These MBT solutions will incorporate the same data available to the human. One of these MBT solutions will be based upon P(D|H) values estimated by the ISO team.

II. CONTROL OVER THE MODIFIED BAYES THEOREM AS A VARIABLE

Dodson's modification of Bayes' theorem allows for multiple event or data categories and multiple states or conditions within each of the data categories (ref. 1). The latter modification is particularly important because it permits expression of uncertainty with respect to the state or condition observed in a data category. The characteristics of the stimulus environment structured for the current series of experiments match the Bayesian paradigm as modified by Dodson. In the experiment being reported, aggressor was allowed four response alternatives (or hypotheses from the point of view of the threat-evaluation team). Appendix I lists the 25 data classes used to describe the attributes or characteristics of the aggressor developmental groupings. Also, as Appendix I shows, each data class had several possible states or conditions. The task for the ISO team was to estimate which level or state existed in each of the 25 data classes for each developmental grouping. Observational uncertainty about the true state of each data class for each developmental grouping could be indicated by the ISOs because their responses in each data class were estimates of the probability that any (k^{th}) state of that data class was, in fact, being observed.

Equation 1 below describes Dodson's MBT with some notational modifications:

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

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 (in the present experiment there are 25 such classes). 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. Equation 1 is thus the basis both for the hypothesis-selection aid given to the CO and for the various Bayesian solutions (described in Section III of this report) calculated by the experimenter for comparison with the human estimates of $P(H_i|D)$.

Dodson has also suggested how the MBT can be made to adapt to changing environmental dynamics. This adaptation feature has as its basis parameters which regulate the rate at which information obsolesces and an expression describing 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 1 (ref. 1, p. 43). The parameters describe how these terms are to be updated on every trial or observation cycle in a sequence. In the experiment being reported the a priori probability term was fixed at .25 since each of the four response alternatives occurred an equal number of times. Since this was so, the experimenter manipulated the parameters only with respect to the $P(D|H_i)$ term. It is true, of course, that human performance under differing a priori probabilities is an extremely important issue. At this juncture in the experimental series, however, the 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. Equation 2, with notational modification, is Dodson's expression for $P(D|H_i)$ illustrating the adaptive or "learning" features allowed by the parameters.

$$P(D_{jk}|H_i)_v = \frac{P(D_{jk}|H_i)_{v-1} + K_v [P(D_{jk})_v \, P(F_i)_v \, w_{iv}]}{1 + K_v [P(F_i)_v \, w_{iv}]} \qquad (2)$$

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 specific input-output feedback sequence number, or simply, an 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. (These values are 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 input pattern. In the present experiment the term applies to the strategy (H_i) aggressor actually used in cycle v. $P(F_i)_v$ assumed 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 conditional probabilities. K_v can assume any value in the range $0 \leq K \leq \infty$. When $K = 0$, no adjustment of the preceding conditional probability (on the $v\text{-}1^{th}$ cycle) is made, i.e., equation 2 reduces 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 $P(D_{jk})_v$.* This means that $P(D_{jk}|H_i)$ on the v^{th} cycle is entirely determined by the most recent observation of $P(D_{jk})$. Large changes of $P(D_{jk}|H_i)_v$ in the direction of the most recent observation $P(D_{jk})_v$, however, can be made in the range $0 \leq K_v \leq 9.99$. (In the present experiment K_v was limited to values in the range $0 \leq K \leq 9.99$.)

---

* 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:

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

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

w_{iv} = a parameter which regulates 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. Values of w_{iv} are limited to the range $0 \leq w_{iv} \leq 1.00$. The reason for this particular range will become apparent in the discussion below when we define w_{iv} more precisely.

Let us now specify the manner in which MBT-aid control was made an experimental variable. From equations 1 and 2 above, and noting the fact that the a priori term was constant at .25 throughout the experiment, there appear to be two ways in which the user of the MBT aid could control the process of computer-implemented solutions of $P(H_i|D)$. First, one might allow the user to manipulate, by means of the parameters K_v and w_{iv}, the rate at which the probabilistic information used in the MBT solutions becomes obsolescent or the degree to which new data are allowed to influence current MBT solutions. Such manipulation would be justifiable, for example, whenever the MBT-aid user had any knowledge of the reliability or accuracy of the data presented to him or had knowledge of the discrepancy between a previous MBT-predicted hypothesis and the true one. By raising K_v or w_{iv}, the user is merely indicating "let us place greater faith in this more current information." The second method of MBT-aid control is to have the user change the $P(D_{jk}|H_i)$ values himself on a cell-by-cell basis. These $P(D_{jk}|H_i)$ values could be judged either by the CO or by members of his staff. Indeed, this is analogous to Edwards' paradigm for a PIP (Probabilistic Information Processing) system (refs. 2, 3). Direct insertion of the $P(D_{jk}|H_i)$ values in this manner represents, ignoring the a priori values as we have done, the highest level of control over the MBT aid.

We are now in a position to specify the four experimental conditions maintained in the present experiment.

Condition I (no-aid condition): In this condition probabilistic estimates of the likelihood of the various aggressor strategies given a sample of probabilistic attribute data $[P(H_i|D)]$ were made by the CO without benefit of any MBT aid. The purpose of this condition was to provide base line data in an unaided condition for later comparison with the data collected in the other three conditions of the experiment when the MBT aid was available.

Condition II (self-adapting MBT-aid condition): Bayesian solutions were available to the CO to aid him in his judgments, but he had no control over the process The constants K_v and w_{iv} were set by the experimenter. On the first day of this condition the $P(D_{jk}|H_i)$ values for every state of the 25 data classes across all four hypotheses were set at chance. Chance

probability was determined by the number of states (μ) within any of the data classes [$\sum_{k=1}^{\mu} P(D_{jk}|H_i) = 1.00$ for every data class given any of the hypotheses]. K_V was set at $K_V = .5$ by the experimenter and remained so throughout the experiment. $K_V = .5$ introduced a reasonable but not drastic automatic gross update of the $P(D_{jk}|H_i)$ values on the basis of incoming $P(D_{jk})$ information. The vernier adjustment parameter w_{iv} was changed automatically throughout the experiment according to the following equation which defines the w_{iv} parameter in condition II.

$$w_{iv} = [P(H_i|D)_v - P(F_i)_v]^2 \qquad (3)$$

where:

$P(H_i|D)_v$ = the MBT-calculated a posteriori probability that hypothesis i occurred given the probabilistic estimates that each of the states of all 25 data classes was observed in cycle v.

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

This equation says, in effect, that the closer the calculated $P(H_i|D)$ was to being correct the smaller was the change induced in the values of $P(D_{jk}|H_i)$ on the following cycle. Thus, large changes would not be induced in $P(D_{jk}|H_i)$ values which were shown to be relatively accurate. Equation 3 also shows why w_{iv} was limited to the range $0 \leq w_{iv} \leq 1.0$. Note, by observing equation 2, that a change in $P(D_{jk}|H_i)$ values could occurr only in the hypothesis category correct on that cycle, i.e., where $P(F_i)_v = 1.0$. When $P(F_i)_v = 0.0$ the right-hand side of equation 2 reduces to $P(D_{jk}|H_i)_{v-1}$.

Condition III (MBT aid with CO adjustment of MBT parameters K_V and w_{iv}: In this condition the CO received MBT aid and exerted rudimentary control over it by adjusting the parameters K_V and w_{iv}. On the first day of this condition the $P(D_{jk}|H_i)$ values were set at a chance level as in condition II. The CO was allowed to change K_V and w_{iv} on every other experimental session within the ranges $0 \leq K_V \leq 9.99$ and $0 \leq w_{iv} \leq 1.00$. The data-processing load on the computer facilities precluded a more frequent alteration of these parameters. The CO was instructed to manipulate these parameters when he wished to affect the rate of data obsolescence. Since he monitored the information-processing activities of his ISOs and was given information about the relative accuracy of the $P(D_{jk})$ estimates they produced, he had some notion of the reliability of the data being introduced into the MBT. He was instructed to apply K_V as a gross weighting factor which might vary according to his perception of the reliability of information produced by the system that experimental day. In addition, he was told to adjust w_{iv} in accordance with the discrepancy between the MBT-calculated $P(H_i|D)$ values and the true hypotheses which were given to him at the end of each experimental cycle.

Condition IV [MBT aid with CO insertion of human-estimated $P(D_{jk}|H_i)$ values]: Estimated values of $P(D_{jk}|H_i)$ were relayed to the CO by his ISOs and were revised by the CO (if necessary) before being entered into the MBT. By entering these values cell by cell into the MBT, the CO directly determined the conditional values used in the solutions of the a posteriori probabilities. This direct insertion represented the greatest degree of control given to the CO over his MBT-aid solutions.

III. EXPERIMENTAL DESIGN AND MEASURES

This experiment consisted of 60 trials or sessions each of 4-hour duration. Since the experimental problems were presented in double time, the threat-evaluation team observed 8 hours of events in the aggressor "world" during each 4-hour session. System load, defined as the number of aggressor developmental groupings terminating in a given session, was constant throughout the experiment. In each session four developmental groupings terminated. This load level was chosen on the basis of results obtained in the first experiment in the series. The assignment of the four experimental conditions (described in the preceding section) across the 60 sessions of the experiment presented several problems. A complete counterbalancing of conditions in order to take into account possible learning and perseveration effects was not possible. A compromise experimental plan involving a partially counterbalanced arrangement of conditions was chosen. In the plan (see table 1) the experimental conditions are numbered in the same manner as in the preceding section of the report.

TABLE 1

EXPERIMENTAL PLAN

Experimental Block	Sessions	Experimental Condition
1	1-12	I
2	13-24	II
3	25-36	III
4	37-48	IV
5	49-54	III
6	55-60	II

There were 10 different measures taken with respect to the performance of members of the threat-evaluation team and various MBT $P(H_i|D)$ solutions. The point at which these measures could be taken depended upon the particular experimental condition. Table 2 lists the various measures and the experimental conditions in which they were taken.

TABLE 2

PERFORMANCE MEASURES

Measure	Experimental Condition			
	I	II	III	IV
1. Unaided CO $P(H_i\|D)$ estimation accuracy	X			
2. Aided CO $P(H_i\|D)$ estimation accuracy		X	X	X
3. Self-adjusting MBT $P(H_i\|D)$ accuracy using same $P(D_{jk})$ values as CO	X	X	X	X
4. Self-adjusting MBT $P(H_i\|D)$ accuracy using correct values of $P(D_{jk})$	X	X	X	X
5. Self-adjusting MBT $P(H_i\|D)$ accuracy using CO-adjusted parameters K_v and w_{iv}			X	
6. Self-adjusting MBT $P(H_i\|D)$ accuracy using ISO-CO estimates of $P(D_{jk}\|H_i)$				X
7. Accuracy of final MBT aid used by CO		X	X	X
8. Accuracy of final estimates by ISOs of $P(D_{jk})$	X	X	X	X
9. Accuracy of $P(D_{jk})$ used by CO in the final MBT-aid solutions		X	X	X
10. Accuracy of ISO-CO estimates of $P(D_{jk}\|H_i)$				X

Some of the measures listed in table 2 should be apparent from the preceding discussion while others require further explanation. Figure 1, illustrating the location in time of various human and MBT solutions for a particular developmental grouping, should aid the reader in seeing the points at which certain of the measures were taken. Figure 1 shows that as time elapsed during the buildup of a certain developmental grouping the

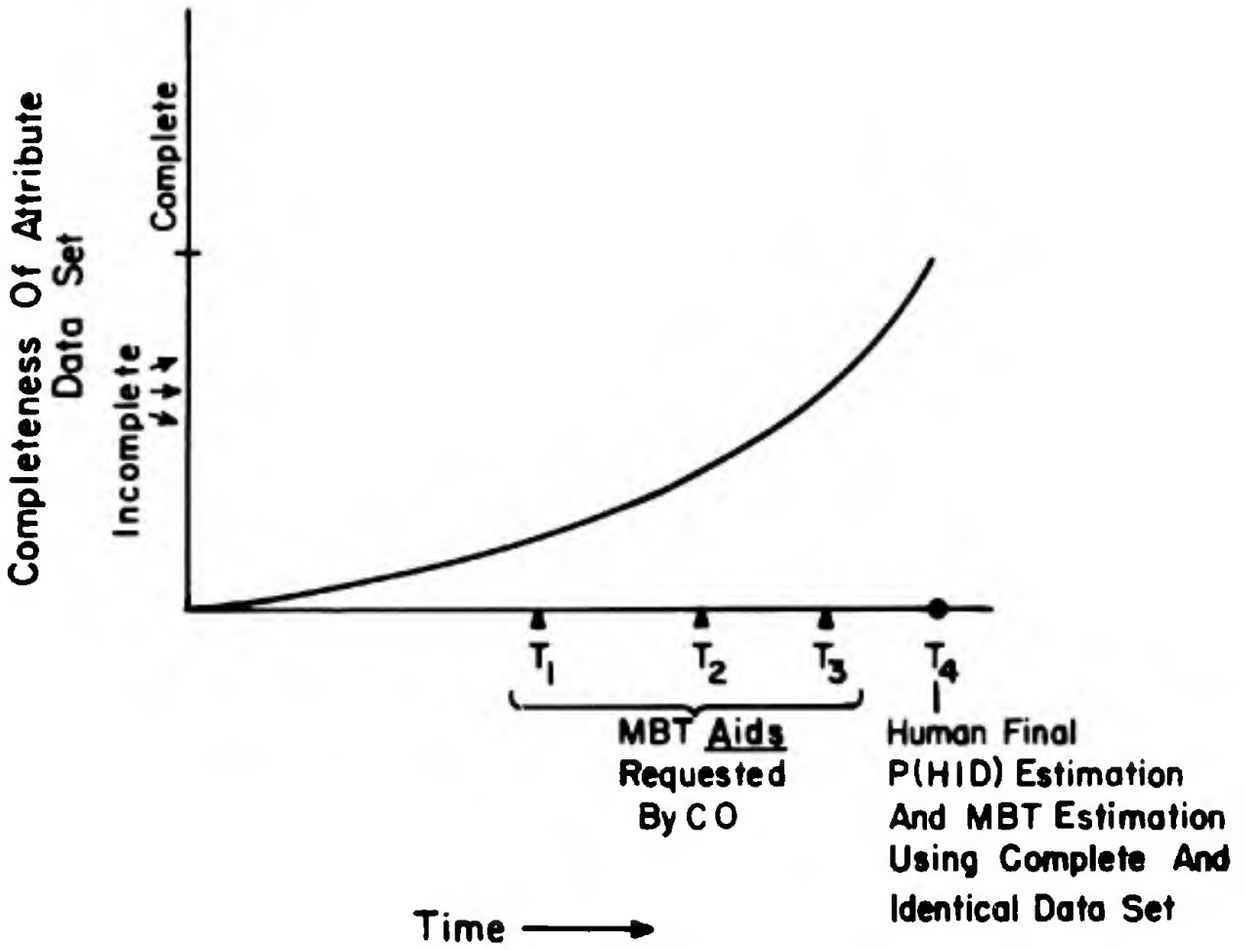


Figure 1. Location in Time of Various Human and MBT $P(H_i|D)$ Estimations for a Certain Developmental Grouping.

completeness of the attribute data set increased roughly according to the function as shown. Since MBT aid could be requested at any stage of this buildup process, MBT aid requested at times prior to termination (T_4) would be based upon incomplete and often less accurate data sets. This was an unavoidable situation due to the time-dependent character of the stimulus environment. Measures 1, 2, and 3 were all taken at T_4 on the basis of complete data sets (i.e., ISO responses in all 25 of the data classes). The same data were used in measure 3 as the CO used in his aided and unaided $P(H_i|D)$ estimates. (In the present experiment the CO made no interim $P(H_i|D)$ estimates as he did in the first experiment.) Measures 5 and 6 were taken of MBT $P(H_i|D)$ estimations also on the basis of the same final complete data sets available to the CO in the conditions in which these measures could be taken. Measure 7 was taken with respect to the last MBT aid (T_3 in figure 1) requested by the CO before he made his final aided $P(H_i|D)$ estimates. Thus, measure 7 reflects the accuracy of the MBT aid requested on the basis of the most complete set of $P(D_{jk})$ values prior to T_4. Measure 8, taken at T_4, reflects the accuracy of the data used in the various human and MBT estimates of $P(H_i|D)$ and indicated by measures 1, 2, 3, 5, and 6. Measure 9 reflects the accuracy of $P(D_{jk})$ values which the CO entered for his final MBT-aid solutions (at T_3 in

figure 1). An MBT solution of $P(H_i|D)$ using only the correct or true states or conditions of each of the 25 data classes was provided (measure 4). These $P(H_i|D)$ values, calculated on the basis of the true data states, furnished some notion of an accuracy upper limit for the MBT solutions. Measure 10 reflects the accuracy of the $P(D_{jk}|H_i)$ values estimated by the ISOs and revised by the CO.

There was a methodological peculiarity in the experiment which will bear some explanation. The original design of the stimulus environment specified 20 different aggressor response alternatives or strategies. These strategies were defined on the basis of three dimensions: (a) whether a developmental grouping of aggressor forces was to represent a forthcoming hostile action (a real attack) or whether the grouping represented rehearsals or maneuvers being conducted by aggressor for training purposes, (b) the probable endurance of the threatened action in terms of the depth to which the forces in the grouping could penetrate into friendly territory given the available logistics support, and (c) specification of the particular infantry-armoured tactic to be initiated by the ground forces in the grouping, e.g., double pincer, multiple penetration, etc. The original 20 possible response alternatives are shown in Appendix II of this report. A table of contingency rules relating every state of each of the 25 data classes to each of these 20 response alternative classes was prepared. This table was in fact the data-generation model by means of which individual samples of the 20 possible types of developmental groupings were drawn and prepared according to a procedure specified in detail in the report of the first experiment (ref. 6, Section IV). Preparation of these individual samples of aggressor developmental groupings (which were called developmental grouping scripts) was an exceedingly laborious and time-consuming task. For example, the preparation of scripts which specified large aggressor groupings required up to 30 man-hours. For this reason a limited number of samples of each aggressor response alternative was available. On the average there were four samples of each strategy type. Methodological problems were posed by this response-alternative, sample-size restriction as well as by the limited amount of time available for operation in each of the four experimental conditions. First, because the time available for each experimental condition was limited, the size of the set of response alternatives had to be reduced. If we had used the complete set of 20 response alternatives and had presented the threat-evaluation team with examples of each alternative, we could have provided the team with at most three examples of each aggressor response alternative and maintained a balanced situation across all 20 alternatives in any of the experimental conditions. (Such balancing would not have been necessary, of course, if we had not desired the frequency of occurrence of the alternatives to be equal across the alternative set.) Moreover, stable or meaningful performance measures could not have been taken in any experimental condition where there were only three examples of each of the possible alternatives. The task of estimating $P(D_{jk}|H_i)$ for the 20-alternative case would have been difficult for the ISOs to complete in the 4-hour experimental sessions. Mere reduction of the existing alternative set (say from 20 alternative to 4 or 8) was also judged to be undesirable. Although this procedure would have allowed more possible observations of each aggressor response alternative, the limited number of samples would have permitted the team

to observe the same developmental groupings on frequent occasions throughout the experiment.

A procedure was developed which, although not perfect, allowed at least a partial solution to the various problems discussed above. First, four aggressor response alternatives were defined and labeled simply A, B, C, and D. Then, for each of these four alternatives, selections were made from the 20 original response alternatives (listed in Appendix II) which defined the characteristics of the four arbitrary alternatives for varying lengths of time. Table 3 should help to clarify this procedure.

TABLE 3

AGGRESSOR RESPONSE ALTERNATIVES AND THEIR CHANGING CHARACTERISTICS

Aggressor Response Alternative							
A		B		C		D	
Original Alternative Type*	Days	Original Alternative Type	Days	Original Alternative Type	Days	Original Alternative Type	Days
12	1-9	16	1-13	3	1-15	9	1-18
13	10-21	10	14-29	2	16-25	20	19-28
17	22-34	6	30-37	14	26-42	7	29-44
5	35-47	15	38-42	18	43-50	4	45-60
1	48-55	8	43-60	19	51-60		
11	56-60						

* The original alternative type numbers in the first columns of the A, B, C, and D categories refer to the original response alternative numbers listed in Appendix II.

Table 3 shows, for example, that response alternative A was defined by original response alternative 12 for the first 9 days, by original response alternative 13 for the next 12 days, and so on. In summary, the procedure allowed four arbitrary classes of aggressor response alternatives whose characteristics or defining dimensions changed at specified times during the experiment. The decision to use four alternatives was made essentially because of the magnitude of the $P(D_{jk}|H_i)$ estimation task. With a larger number of alternatives it was felt that the ISOs could not make these estimations and perform their various other tasks. The procedure allowed the use of all samples from each of the 20 original alternative classes for a reasonable number of times each during the experiment. In addition, it permitted the experimenters to specify an equal a priori probability of occurrence of the response alternatives. Finally, each definition or rule

change shown in figure 2 was called a "category change." It was not possible to have an equal number of category changes in each of the four alternative categories. This was due primarily to the fact that the individual developmental groupings had differing buildup cycle lengths (i.e., they took either 2, 3, 4, or 5 days to develop). Fitting these groupings into a real-time environment with the constraint that four terminate each day prevented perfect counterbalance.

IV. SYSTEM TASKS AND PROCEDURES

In Section V of the report describing the first experiment* a detailed description was presented of the various tasks performed by each member of the threat-evaluation team. Since there were only a few methodological changes introduced for the present experiment, the reader can refer to the earlier report for as much additional detail as he wishes. In general, the mission of the threat-evaluation team was to produce probabilistic evaluations of the threat posed by certain developments of aggressor forces in a simulated hostile environment. These evaluations, in the form of $P(H_i|D)$ estimates, were produced by the CO of the team on the basis of data which he received from various members of his intelligence staff. These data, describing the attributes or features of the developmental groupings of aggressor's forces, were themselves probabilistic estimates of the state or condition existing in each of 25 different attribute data classes $[P(D_{jk})]$. Four aggressor developmental groupings terminated in each of the 60 experimental sessions. One-half hour before the termination time for each grouping the CO produced his estimate of $P(H_i|D)$ on the basis of a complete set of attribute data $P(D_{jk})$. This is precisely the time indicated as T_4 in figure 1. During the buildup of each developmental grouping the CO used interim estimates of $P(D_{jk})$ provided by the ISOs. In the experimental conditions in which he had access to the MBT aid, he entered these interim $P(D_{jk})$ estimates into a 1401 computer which was programed to provide the MBT solutions. In the present experiment, the CO produced no overt interim $P(H_i|D)$ estimates which were scored by the experimenter.

Figure 2 illustrates the flow of information throughout the experimental threat-evaluation system. IBM 1401 and 7090 computer facilities simulated the activities of individuals who gather and process photographic, radar, and infra-red sensor data obtained on hypothetical overflights of aggressor territory. In figure 2 these activities are represented by the primary input data-processing level. The operations liaison officer (OPNS in figure 2) planned these hypothetical overflights in cooperation with the other four intelligence staff officers (labeled A through D in figure 2). A request specifying an overflight was entered by the operations officer into the 1401. This entry was called an "acquisition request" (AR in figure 2). When an AR had been processed by the computer and notification regarding this completed request had been given to the operations officer, the other four ISOs made specific interrogations of computer storage regarding information collected on the simulated overflights. These requests were called "reconnaissance analys's requests" (RAR in figure 2). Since all events in aggressor territory were time-dependent, the ISO team was required to update its information continually by successive AR-RAR

* reference 6

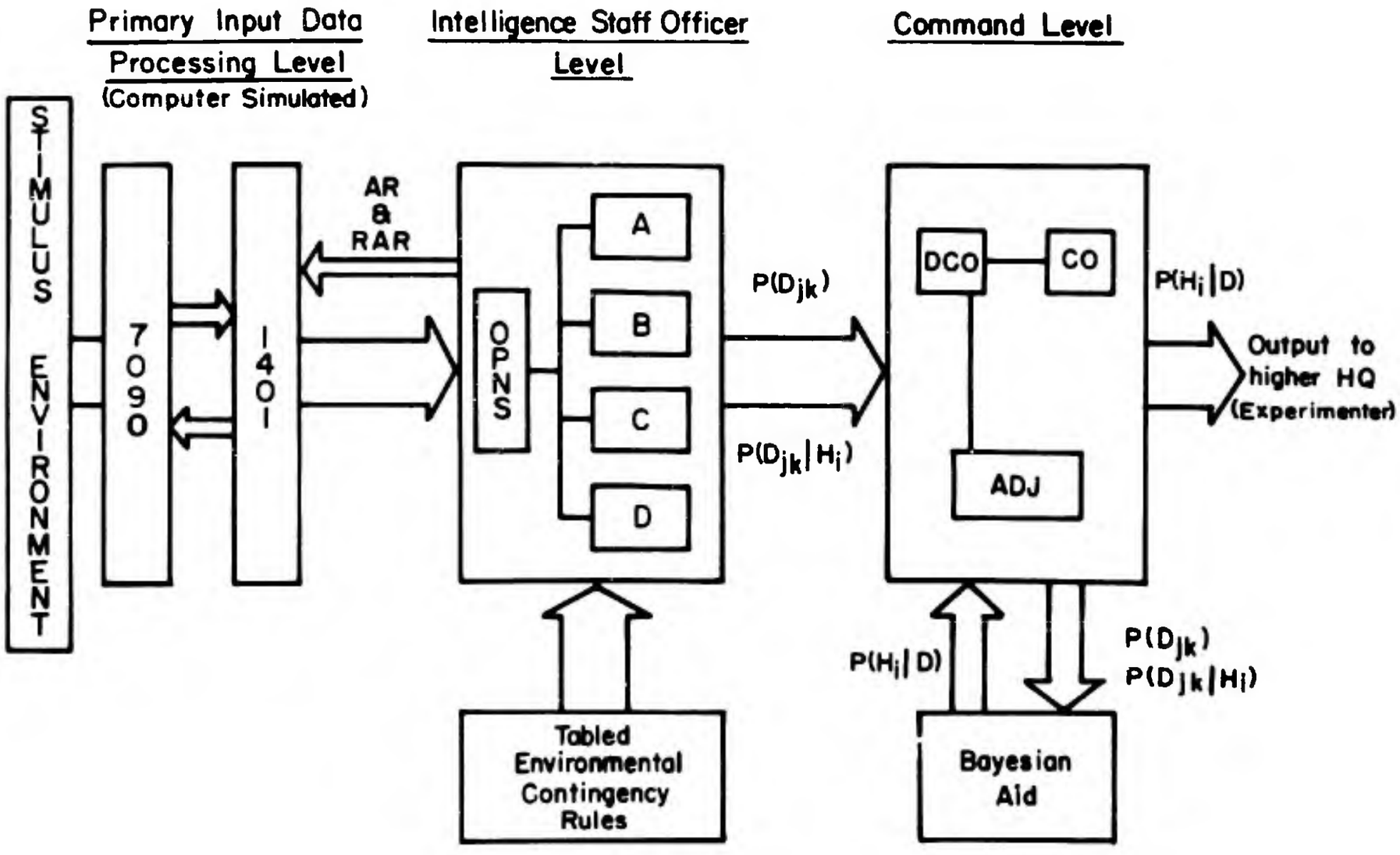


Figure 2. System Task Description (see text for explanation).

sequences. On demand from any of the ISOs (by means of the RAR) the computer facilities produced verbal descriptions of aggressor elements (mobile weapons, vehicles, and aircraft), element activity, and locus of element activity. On the basis of this input information, and making use of tabled contingency data, the ISOs identified and tracked concentrations of weapons and vehicles and produced probabilistic estimates [$P(D_{jk})$] of the state or level of each of the attributes of the specific aggressor developmental groupings. As noted earlier, each developmental grouping could be described in terms of 25 data classes (listed in Appendix I). Each of the four ISOs was responsible for a subset of these 25 data classes. In a sense, each served as a "content expert." One ISO was a logistics expert, one was an order of battle expert, and so on. The ISOs produced both interim and final estimates of $P(D_{jk})$ for each developmental grouping and passed these on to the CO. The CO used the interim estimates for the MBT aid and for his own interim $P(H_i|D)$ estimates which he recorded but did not forward to the experimenter. The final $P(D_{jk})$ estimates were given to the CO enough ahead of time so that he could produce his final $P(H_i|D)$ estimates and forward them to the experimenter before the developmental grouping terminated. In conditions II, III, and IV of the experiment, the ISOs produced daily $P(D_{jk}|H_i)$ estimates in addition to the $P(D_{jk})$ estimates. Although estimated by the ISOs in conditions II, III, and IV, the $P(D_{jk}|H_i)$ values were only utilized for MBT-aid solution in condition IV. The estimates made in conditions II and III were required in order to give the ISOs experience in producing them. The $P(D_{jk}|H_i)$ estimates could be revised by the CO before entry into the 1401 for the MBT aid in condition IV.

At the beginning of each experimental session the CO was informed about the accuracy of the $P(H_i|D)$ estimates he produced in the preceding session. The ISOs were given similar knowledge of results about their $P(D_{jk})$ estimation performance of the previous day. They were not informed, however, about the accuracy of their $P(D_{jk}|H_i)$ estimates. All eight subjects in the present experiment served in the first experiment. All had completed the training program described in the report of the first experiment. Additional training was necessary for the four ISOs and the CO who assumed the additional task of estimating $P(D_{jk}|H_i)$. During the week before the experiment began, these individuals were given practice in providing these estimates. Each ISO, as noted previously, was responsible for a subset of the data classes shown in Appendix I of this report. Based upon knowledge of the correct hypotheses describing the developmental groupings which in fact occurred on the previous day in combination with the data estimates for these developmental groupings, each ISO estimated $P(D_{jk}|H_i)$ for each of the data classes in his area of responsibility.

V. RESULTS

Three different types of scoring procedures were used with respect to the various performance measures indicated in table 1. For the $P(H_i|D)$ estimates produced by the CO and the MBT, and for the $P(D_{jk})$ estimates produced by the ISOs, a score called "verified certainty" was used. For the $P(H_i|D)$ estimates, verified certainty was simply the value of the probabilistic estimate placed in the correct hypothesis category for each

developmental grouping. For the $P(D_{jk})$ estimates, verified certainty was the value of the probabilistic estimate placed in the correct state or condition of each data class for each developmental grouping. The CO's probabilistic estimates in a hypothesis category and the ISOs' probabilistic estimates in the various states of a data class were assumed to represent their degree of certainty that the hypothesis would explain the occurrence of the observed data or that the state or condition of the data class was being observed. These certainty estimates were verified by comparing them with the true hypotheses [for the $P(H_i|D)$ estimates] or with the true level or condition [for the $P(D_{jk})$ estimates]. The verified certainty scores for the $P(H_i|D)$ estimates do not, however, show the precise number of occasions on which the highest CO or MBT estimates (for each developmental grouping) were correct or incorrect. For this reason another score, called a "dichotomous score," was used for the $P(H_i|D)$ estimates. Using this procedure, the highest $P(H_i|D)$ estimate or "first choice" among the four hypotheses was scored as being correct or incorrect. For the $P(D_{jk}|H_i)$ estimates produced by the ISOs and revised by the CO a score termed an "agreement score" (a_j) was used. The score a_j was defined as follows:

$$a_j = 1 - \frac{\sum_{k=1}^{\mu} |d_k|}{2} \qquad 0 \leq a_j \leq 1.0 \qquad (4)$$

where:

μ = the number of states or conditions in a data class
k = the k^{th} state or condition in a data class
$d = [P(D_{jk}|H_i)_{estimated} - P(D_{jk}|H_i)_{true}]$
i = the i^{th} hypothesis category
j = the j^{th} data class

Perfect agreement between the estimated conditional probabilities and the true conditional probabilities within a certain data class given a certain hypothesis yielded a score of $a_j = 1.00$. Complete lack of agreement yielded $a_j = 0.0$. There was an a_j score for each data class under each of the four hypotheses. Since $P(D_{jk}|H_i)$ was estimated once each day by the ISOs, there were 25 (data classes) × 4 (hypotheses classes) or 100 a_j scores on each day in experimental blocks 2 through 6.

There are four general classes of results to be presented with respect to the present experiment: the effects of increased control over the MBT aid upon the CO's performance, comparison of MBT and CO performance using identical data, accuracy of the $P(D_{jk}|H_j)$ estimations produced by the ISOs, and the effects of the category change procedure upon CO and MBT performance. Unfortunately, the fact that the primary data were collected from only one subject precluded an elegant statistical analysis. The reasons for failure to use more than one subject in key roles (CO for example) were discussed in the report of the first experiment.

A. The Effects of Increased MBT-Aid Control upon the $P(H_i|D)$ Estimations Produced by the CO

In figure 3 the solid line connecting the filled circles represents the CO's $P(H_i|D)$ estimation performance under the one unaided and three aided conditions of the experiment. First, as the graph indicates, a regular and systematic increase in CO's $P(H_i|D)$ performance with increased MBT-aid control (as we defined "control") was not obtained. Performance in condition II (self-adapting aid with no CO control) and condition III (aid with CO adjusting K_V and w_{iV}) were both significantly superior to performance in the

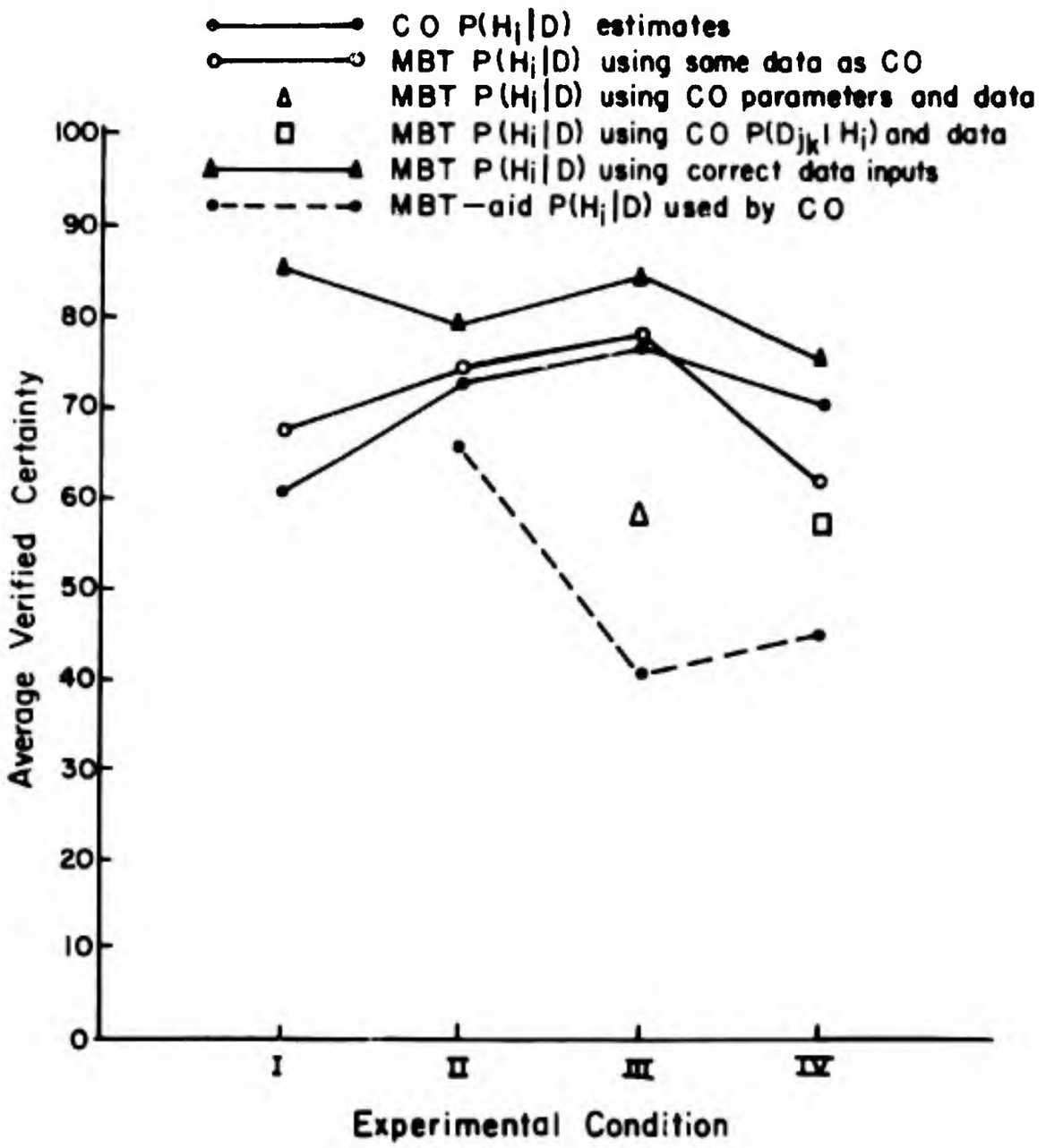


Figure 3. Various $P(H_i|D)$ Verified Certainty Scores by Experimental Condition.

unaided condition (I): II > I with $Z = -1.62$, $p < .05$; III > I with $Z = -2.27$, $p < .01$ (using the Sum of Ranks Test described by Walker and Lev, ref. 7). There were no other significant differences among the four conditions. In particular, the condition representing highest MBT-aid control (IV) was not significantly superior either to the unaided condition or to other aided conditions. The apparent superiority of conditions II and III over conditions I and IV can be accounted for in terms of learning and experience on the part of the CO rather than in terms of MBT-aid configuration. Consider figure 4 and recall that there was a partially counterbalanced experimental design consisting of six blocks; blocks 2 and 6 were performed under

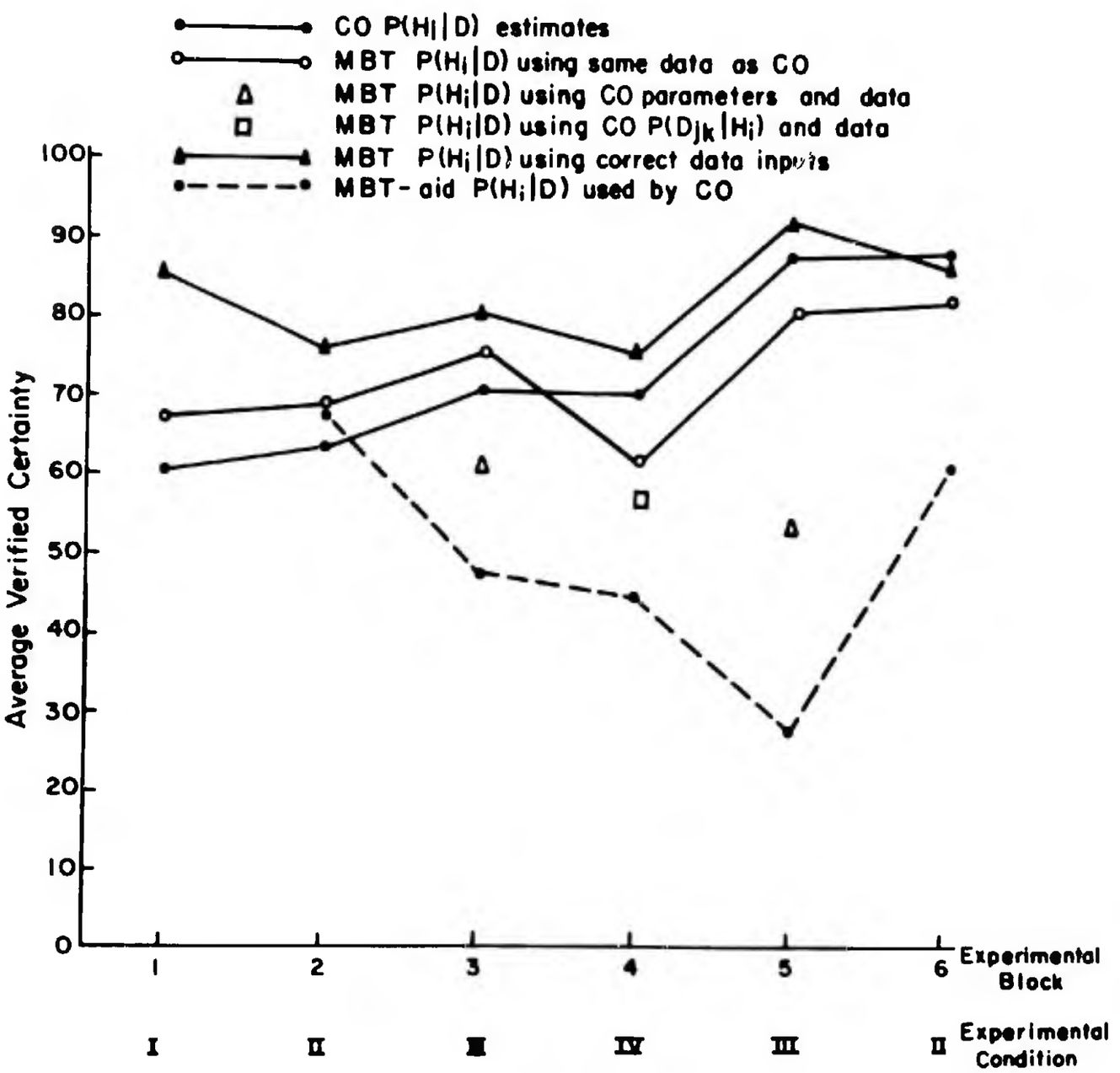


Figure 4. Various $P(H_i|D)$ Verified Certainty Scores by Experimental Block.

condition II and blocks 3 and 5 were performed under condition III. Combining the data in blocks 5 and 3 and combining those in blocks 6 and 2 resulted in the apparent superiority of conditions II and III in figure 3. Notice that blocks 5 and 6 were completed at the very end of the experiment (sessions 49 through 60). Throughout blocks 1 through 6 the CO's performance shows an increase even though the performance of his MBT aid dips sharply in the conditions under which he exerted control over the aid. Therefore, a learning or experience phenomenon independent of MBT aid seems to be indicated since there was no concomitant increase in MBT-aid accuracy (dashed line) throughout these blocks.

The fact that MBT-aid accuracy was inferior to CO accuracy in $P(H_i|D)$ estimation (except in block 2) can be accounted for in two ways. First, the CO made his final estimates on the basis of a more complete and a more accurate set of data than the set he entered into the 1401 for his final MBT aid. The reason for this was discussed in Section III of the report and illustrated in figure 1. The difference in accuracy of the data used by the CO in his final estimates and those he entered into the 1401 for his final MBT aid is shown in figure 5. Second, as figures 3 and 4 indicate, any manipulation of the MBT aid on the part of the CO, either by adjustments of K_v and w_{iv} or by direct insertion of $P(D_{jk}|H_i)$, decreased the controlled MBT-aid accuracy over the self-adapting (uncontrolled) version.

B. Comparison of CO and Various MBT $P(H_i|D)$ Estimations on the Basis of Identical Data

Any comparison between the CO's performance and that of the MBT aid he used would not be a particularly fair one because, as shown in figures 1 and 5, more and better information was usually available to the CO when he made his final estimates than when he requested MBT aid. In order to see how the MBT would perform using the latest data to which the CO was given access, calculations of $P(H_i|D)$ on the basis of these latest data using the self-adapting version of the MBT (see Section II, p. 7) were provided for the experimenter. The accuracy of these calculations is shown in figures 3, 4, and 5 by the open circles. Figure 4 shows that, with respect to the verified certainty scores, this self-adapting MBT was slightly superior early in the experiment, but became inferior to the CO beginning in block 4. The sudden decrease in MBT performance for the developmental groupings in block 4 cannot easily be accounted for. Several recalculations were made of the $P(H_i|D)$ values in this block but the same results were always obtained. One possibility is that the MBT was more sensitive than the CO to disparate situations between observed data and established contingency rules. Note also that the upper limit in accuracy for the MBT (the filled triangles in figure 4) was lowest in block 4. This means that, in a broad sense, the sample of developmental groupings seen in this block was slightly less discriminable. This may indicate that the human tended to assign somewhat higher certainty estimates than his data justified. The CO was, nevertheless, below the theoretical upper limit for the MBT.

Two other versions of MBT solutions were provided on the basis of the same final data used by the CO. One solution, the MBT using the CO's adjustments of K_v and w_{iv} and the same final data as the CO, is shown by the

open triangles in figures 3 and 4. This solution could only be provided in the experimental blocks when the CO adjusted these parameters. As the graphs indicate, these solutions were inferior both to the CO and to the self-adjusting MBT. The same can be said for the MBT solutions provided by the ISO-CO estimates of $P(D_{jk}|H_i)$ and the CO's final data. These solutions, shown as the open square in figures 3 and 4, were also inferior to the human and self-adapting MBT solutions.

The verified certainty score distributions for the CO and the self-adapting MBT using identical data are quite interesting. These are shown

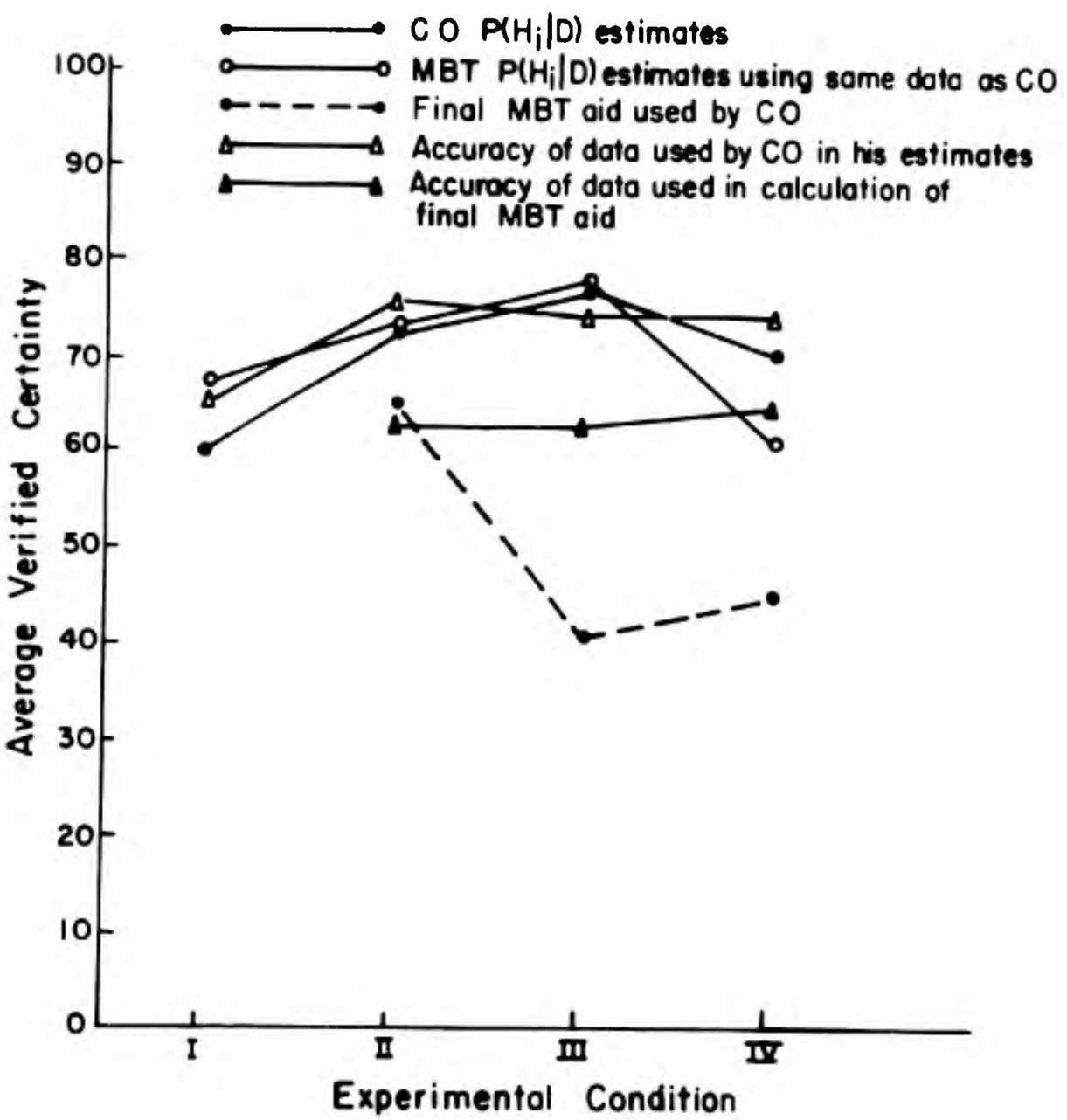


Figure 5. $P(H_i|D)$ and $P(D_{jk})$ Verified Certainty.

in figures 6 and 7. The distributions shown in these figures are verified certainty (probabilistic estimate in the correct hypothesis category) score distributions which were accumulated over the first, middle, and last 12-session periods of the experiment. Notice in figure 6 that in the first 12-session period the CO's scores tend to pile up in the center of the range of possible values. This indicates an apparent conservatism on his part in these early trials. By session 36, however, the distributions begin to show accumulations of very high and very low $P(H_i|D)$ verified certainty scores indicating a shift from conservative to definite commitment-type responses.

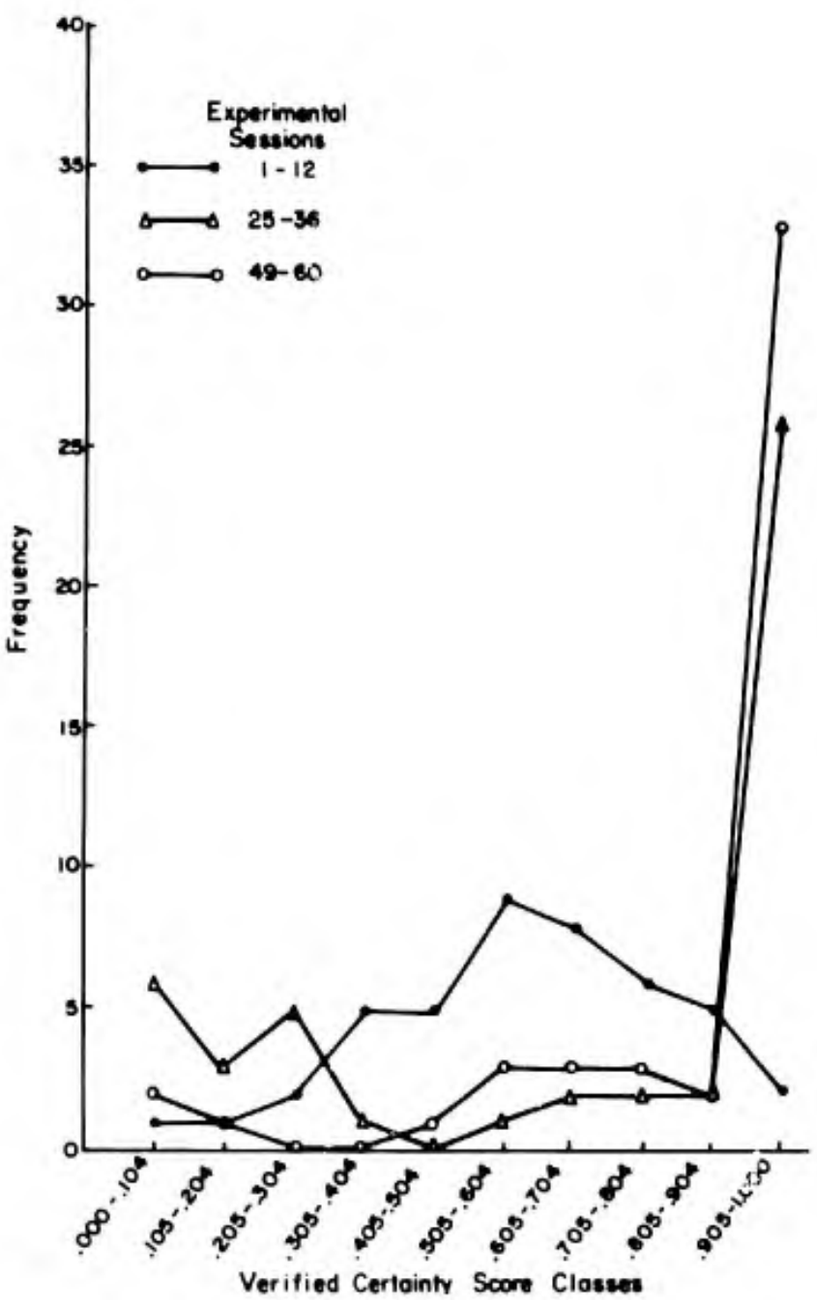


Figure 6. Distributions of $P(H_i|D)$ Verified Certainty scores for the CO.

Figure 7 illustrates the same distributions for the self-adapting MBT. The MBT provided a large number of very high $P(H_i|D)$ estimates from the beginning of the trial series. In order to provide more evidence for the early conservatism on the part of the CO (which later disappears to a great extent), distributions of all $P(H_i|D)$ estimates across all hypothesis categories were graphed. These are shown in figures 8 and 9. Again, early but diminishing conservatism on the part of the CO is apparent.

In comparing the CO and self-adapting MBT thus far, only the verified certainty scores have been considered. Let us now consider the dichotomous

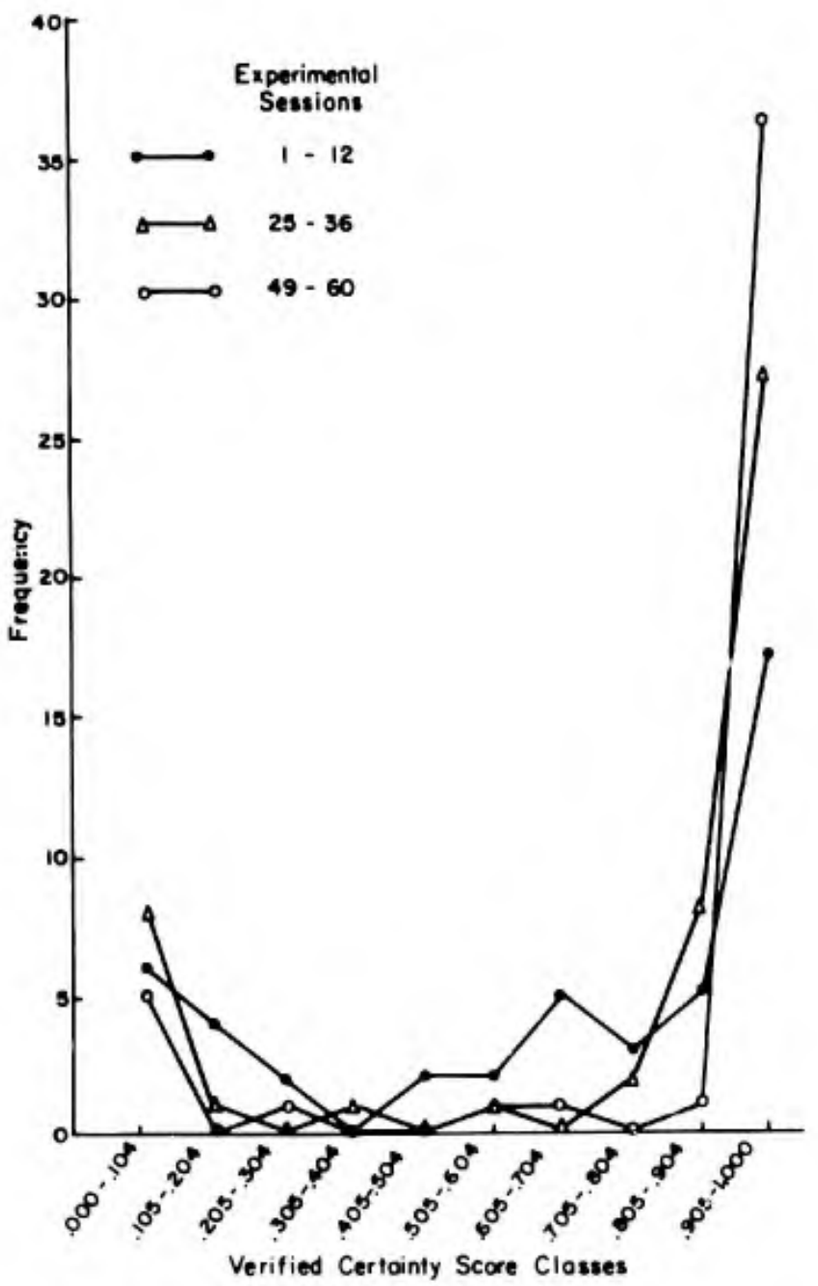


Figure 7. Distributions of $P(H_i|D)$ Verified Certainty Scores for the MBT Using the Same Data as the CO.

TABLE 4

DICHOTOMOUS SCORES FOR CO AND SELF-ADAPTING MBT USING IDENTICAL DATA

	Experimental Condition				Totals
	I	II	III	IV	
CO	31	57	55	35	178
MBT	34	55	57	31	177

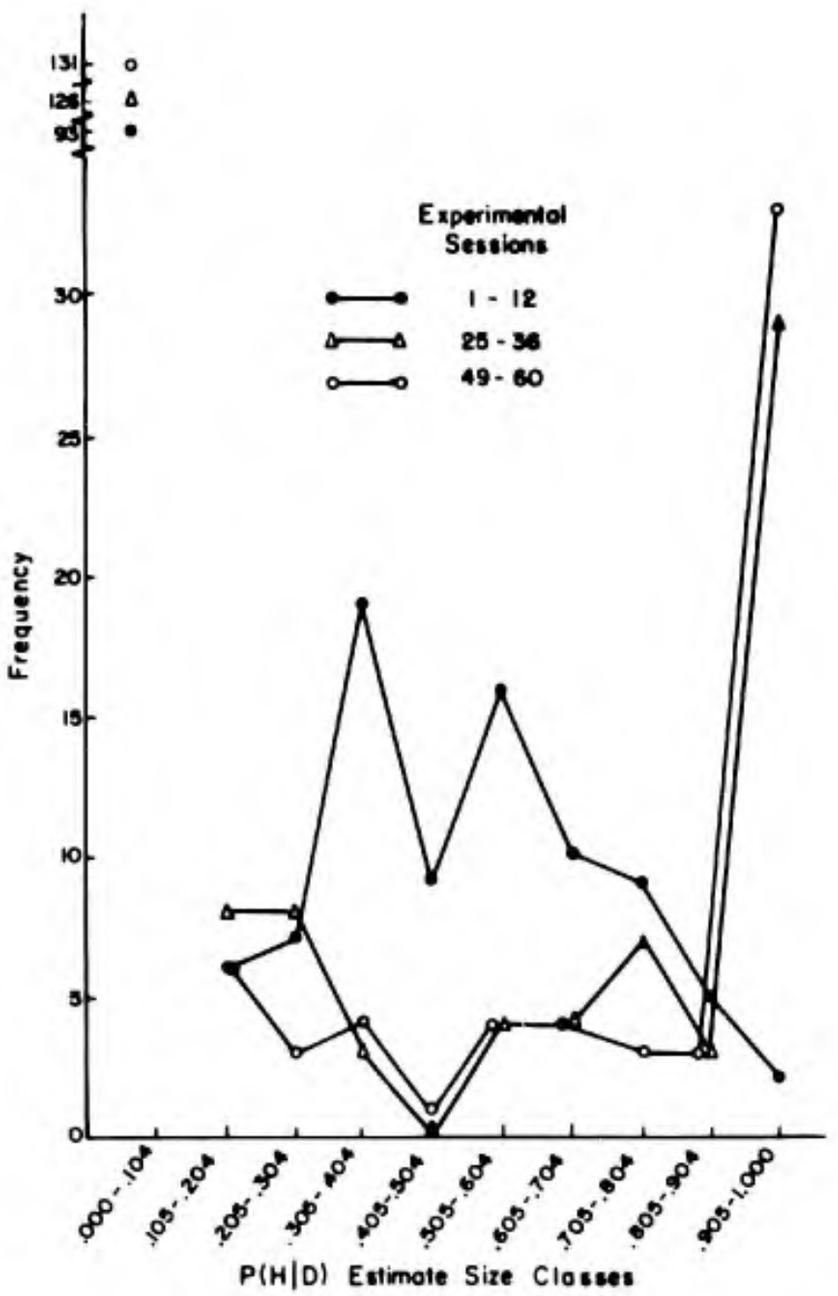


Figure 8. Distributions of all $P(H_i|D)$ Values Estimated by the CO.

scores which indicate the number of occasions on which the highest (or first choice) $P(H_i|D)$ estimate was correct. Table 4 indicates the number of occasions on which the first choice estimates made by the CO and the self-adapting MBT (using identical data) were correct in each experimental condition.

The similarity between CO and MBT scores in each condition is striking. A further breakdown of these dichotomized scores is provided in table 5. The $P(H_i|D)$ estimates for the 236 developmental groupings seen in the experiment were scored in the dichotomized manner. On seven occasions the

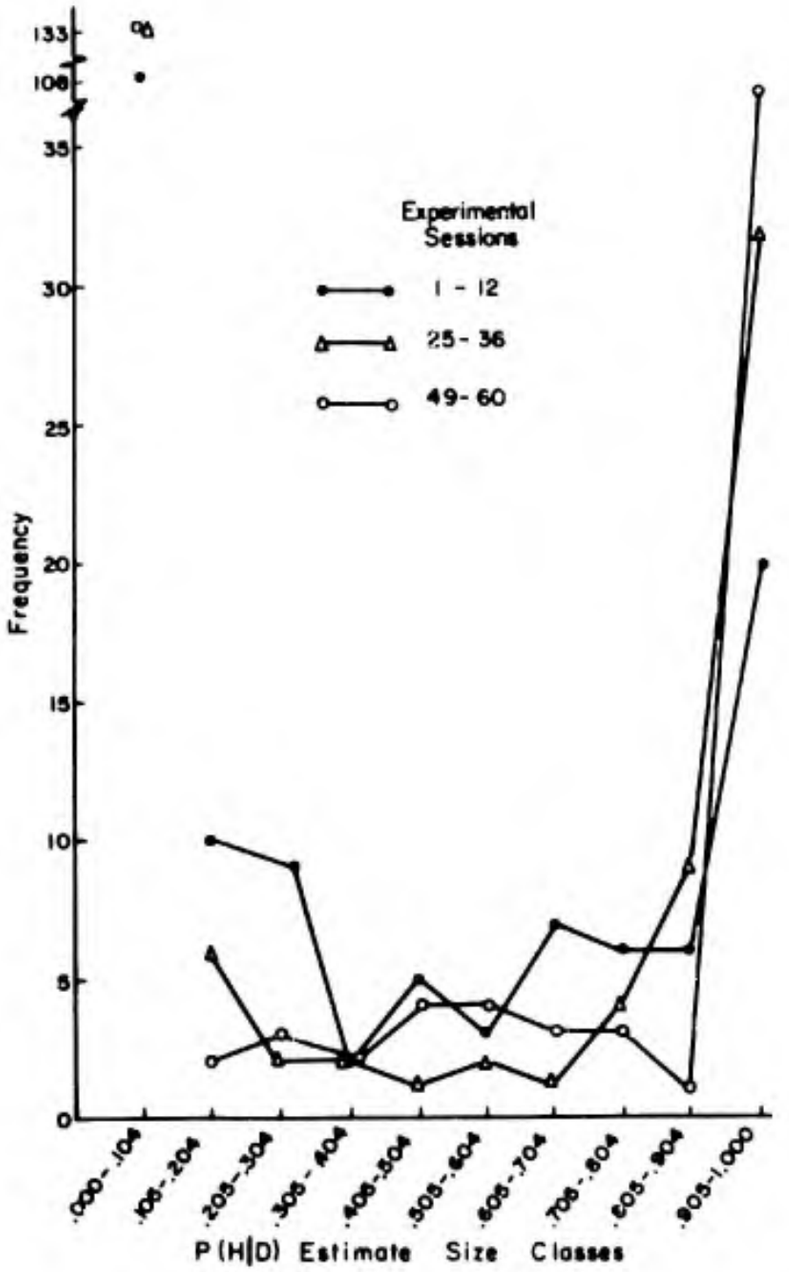


Figure 9. Distributions of all $P(H_i|D)$ Values Estimated by the MBT Using Same Data as CO.

TABLE 5

CO AND SELF-ADAPTING MBT DICHOTOMIZED SCORE BREAKDOWN

	Number of Cases	Percentage of Total
I. CO and MBT Agree	164	69.49
Both were correct	149	63.13
Both were incorrect	15	6.36
II. CO and MBT Disagree	65	27.54
CO was correct	29	12.29
MBT was correct	27	11.44
Neither was correct	9	3.81
III. Equivocal CO Judgments	7	2.97
MBT was correct	1	0.43
MBT was incorrect	6	2.54

CO's first choice could not be determined because of equal probability assignment. These were termed "equivocal" judgments.

In summary, the CO was correct in his first choice or highest $P(H_i|D)$ estimate on 178 out of 236 possible cases or 75.42% of the time. The self-adapting MBT using the same data as the CO produced 177 out of 236 correct first choice estimates or 75% of the time

C. The Effects of the Category Change Procedure upon CO and Self-Adapting MBT $P(H_i|D)$ Estimation

One might expect that the contingency rule alteration occasioned by the category change procedure (see Section III) would have a definite decremental effect on $P(H_i|D)$ estimations produced after the category change. What is interesting, of course, is the number of examples of the new category required to elevate the $P(H_i|D)$ estimates again for the hypothesis category in question. Figure 10 shows this relationship both for the CO and the self-adapting MBT (using identical data). In the upper graph N is the number of changed categories (original aggressor response alternatives as listed in Appendix B). The graph relates N to the number of examples (or individual developmental groupings) seen of a new category after it had changed. For example, in all 16 of the changed categories (other than the first four which were the original definitions) there were at least six examples, in eight of the categories there were at least 14 examples, and in one category there were 18 examples of it following the change. The lower graph shows how the $P(H_i|D)$ estimates produced by the CO and the

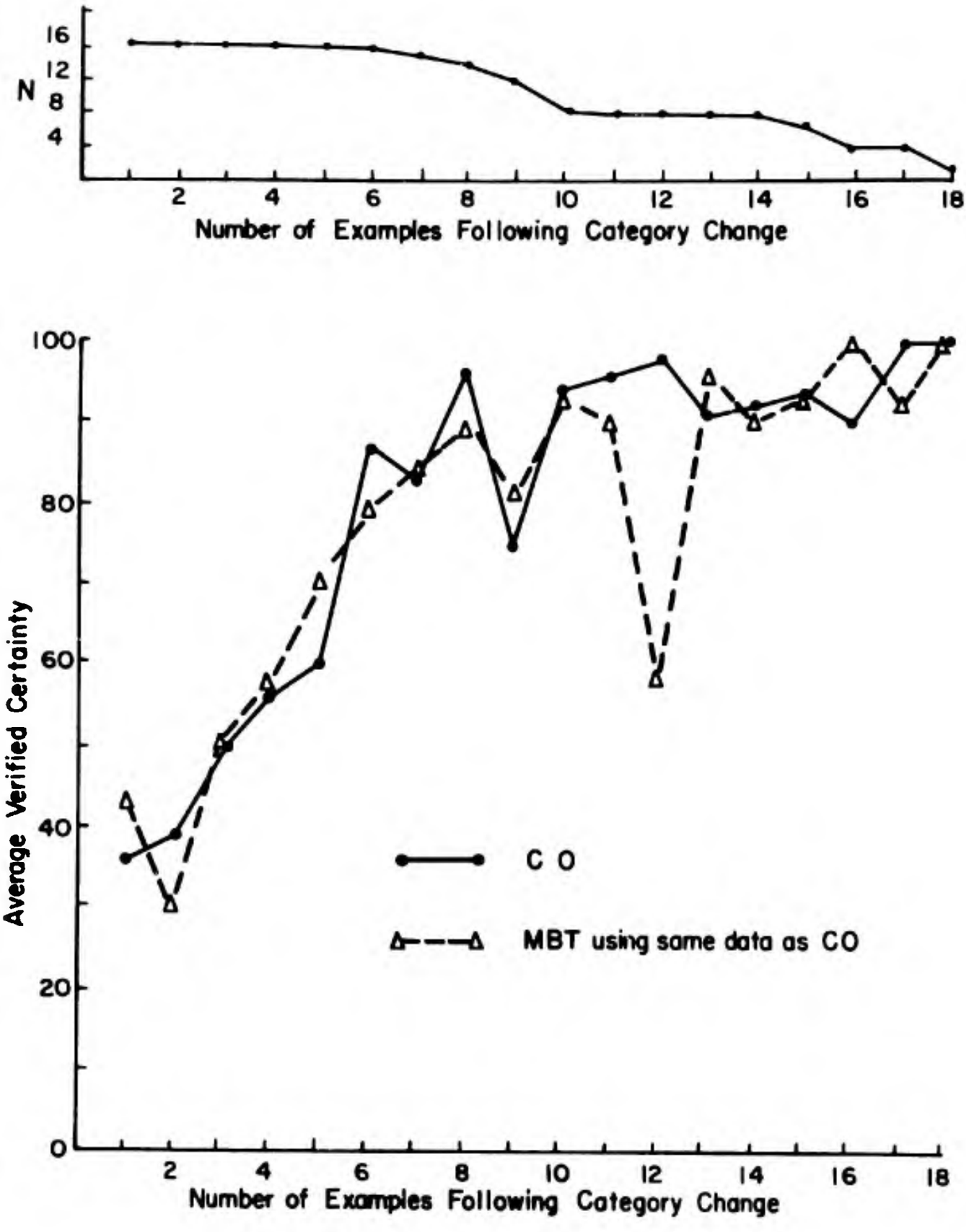


Figure 10. Average $P(H_i|D)$ Verified Certainty for CO and MBT as a Function of the Number of Exposures to Examples of a Response Alternative Category.

self-adapting MBT increased as the number of examples of a category increased. Again, the similarity of performance by the CO and the MBT is striking. The one irregular point on the MBT graph, even though rechecked, remains unaccounted for.

D. $P(D_{jk}|H_i)$ Estimation Accuracy

Unfortunately, there can be no simple statement of the overall accuracy the ISOs or the CO demonstrated in estimating these values. The primary reasons are: (a) since the different data classes had differing numbers of alternative states or conditions, the probabilistic structure underlying distributions of scores in these classes will be different, and (b) the scale properties of the a_j score are somewhat difficult to specify; if a_j has only ordering properties, combination of a_j across different data classes may not be too defensible. For these reasons a_j scores are presented with respect to each of the 25 data classes individually. The form of the presentation of these values judged most meaningful at present is a distribution of a_j averaged over the four hypothesis categories for each data class. Because there are 25 data classes the distributions are included in Appendix III of this report. The mean a_j score for each data class is also shown with each of the distributions.

VI. DISCUSSION AND INTERPRETATION OF RESULTS

Increased control over the MBT aid on the part of the CO, as we defined control, seems to have had little or no effect on the CO's performance in estimating $P(H_i|D)$. The data, presented in figure 4, indicate that the CO went through the required motions of increased MBT-aid manipulation, then in the final analysis relied upon his own judgment. His judgment improved throughout the experiment apparently independent of the MBT-aid configuration. The MBT-aid accuracy was highest in the self-adapting configuration and suffered under both conditions in which the CO exerted control over it. The adjusted parameters K_v and w_{iv} in condition III and the inserted $P(D_{jk}|H_i)$ values in condition IV both served to decrease MBT-aid accuracy. One might argue, of course, that to be maximally effective, MBT aid should have been provided for the CO at the exact time he made his final $P(H_i|D)$ estimates (i.e., at T_4 in figure 1). The fact that all four developmental groupings terminated in each session at nearly the same time imposed data-processing restrictions which prevented MBT aid simultaneously with required decision times. We are in a position, however, to observe the accuracy of the MBT aid under circumstances when the final and complete data set was available to the CO. In figure 4 the open circles in blocks 2 and 6 and the triangles and square in blocks 3, 4, and 5 indicate what the MBT-aid accuracy would have been in these circumstances. In only one block (2) would the verified certainty scores of such an aid have been superior to the CO's estimates. This assumes, of course, that the aid under these latter circumstances would have had no decremental effect on the CO performance.

The verified certainty data show only slight differences between the CO and self-adapting MBT estimates on the basis of identical attribute data.

The dichotomous score data show an even greater overall similarity between these two estimates. Conceivably, the 4-alternative hypothesis situation presented no real challenge to the human inspite of the large amount of data to be processed and the category change procedure. What is most interesting, however, is the shift from conservative human certainty estimates in early trials to very large or definite commitment-type estimates in later trials. It may be necessary to qualify Edwards' conclusions regarding human $P(H_i|D)$ estimation conservatism (ref. 3, p. 16). More evidence is necessary on this issue, however, since we offer data obtained from only one subject.

The category change procedure, instituted out of necessity, seems not to have been particularly bothersome either to human or MBT. Indeed, it even allowed observation of the effects on $P(H_i|D)$ estimation of varying numbers of examples of each H_i category. The similarity exhibited by the CO and self-adapting MBT in adjusting to the differing contingency rules with category change is most interesting.

The discussion will now be concluded with the following general remarks:

1. In two experiments thus far, compelling evidence for the efficacy of a Bayesian hypothesis-selection aid has not been obtained. In part, rather severe methodological difficulties to which the "aid" concept has led, have been responsible. The aid concept has effectively reduced the number of key subjects from whom performance data can be collected. In addition, efficacy of the aid within the context of a time-contingent stimulus environment is difficult to estimate because of the continual obsolescence of information. The precise extent to which the aid influences human judgment on any occasion is difficult, if not impossible to specify.

2. Given a small hypothesis set and an environment in which a meaningful and fairly accurate accumulation of past history can be maintained (i.e., a frequentistic environment), the present data suggest that human $P(H_i|D)$ estimates may approach, if not exceed, the size of those produced by the MBT.

3. Statements regarding the unwillingness of humans to estimate extreme probabilities (i.e., conservatism) may have to be qualified to allow for learning and experience. In the present experiment the data suggest decreasing conservatism with increasing experience.

4. Because the concept of the MBT as an aid leads to rather intractible experimental situations and because there is not yet enough data available to form an intelligent judgment of the efficacy of automated hypothesis selection, it is proposed that the systems research vehicle be used to examine, under a wide range of conditions, more basic issues concerning unaided human commerce with the probabilistic data implied by the Bayesian paradigm.

REFERENCES

1. Dodson, J. D., Simulation System Design for a TEAS Simulation Research Facility, PRC Report R-194, Planning Research Corporation, Los Angeles, Calif., 15 November 1961.

2. Edwards, W., "Dynamic Decision Theory and Probabilistic Information Processing," Human Factors, Vol 4, pp 59-72, 1962.

3. 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. (ASTIA Document No. AD 3780-12-T)

4. Edwards, W., Probabilistic Information Processing by Men, Machines, and Man-Machine Systems SDC Technical Memorandum 1418/000/01, System Development Corporation, Santa Monica, California, 13 August 1963.

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

6. Southard, J. F., 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, June 1964.

7. Walker, H. M., and J. Lev, Statistical Inference, Henry Holt and Company, New York, N. Y., 1953.

BLANK PAGE

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 Armoured 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 armoured 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

ORIGINAL ALTERNATIVE AGGRESSOR STRATEGIES

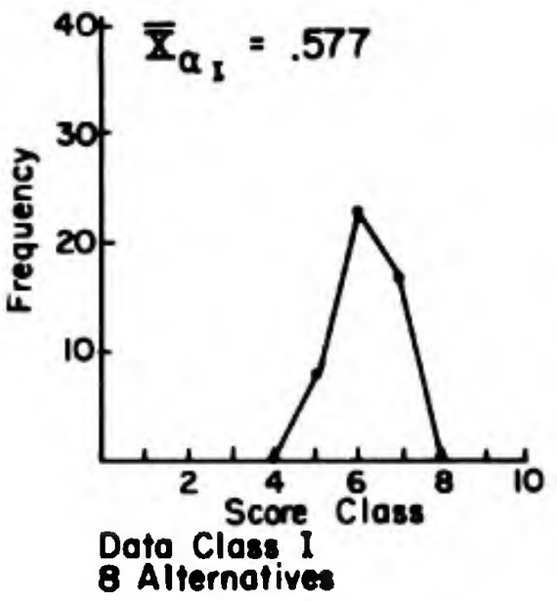
Strategy No.	Penetration Depth (miles)	Strategy
	Actual Attacks	
1	100	Double Pincer
2	100	Multiple Penetration
3	50	Double Pincer
4	50	Multiple Penetration
5	25	Double Envelopment
6	25	Single Envelopment
7	25	Penetration
8	10	Double Envelopment
9	10	Single Envelopment
10	10	Penetration
	Rehearsals	
11	100	Double Pincer
12	100	Multiple Penetration
13	50	Double Pincer
14	50	Multiple Penetration
15	25	Double Envelopment
16	25	Single Envelopment
17	25	Penetration
18	10	Double Envelopment
19	10	Single Envelopment
20	10	Penetration

APPENDIX III
DISTRIBUTIONS OF $P(D_{jk}|H_i)$ AGREEMENT SCORES (α_j)

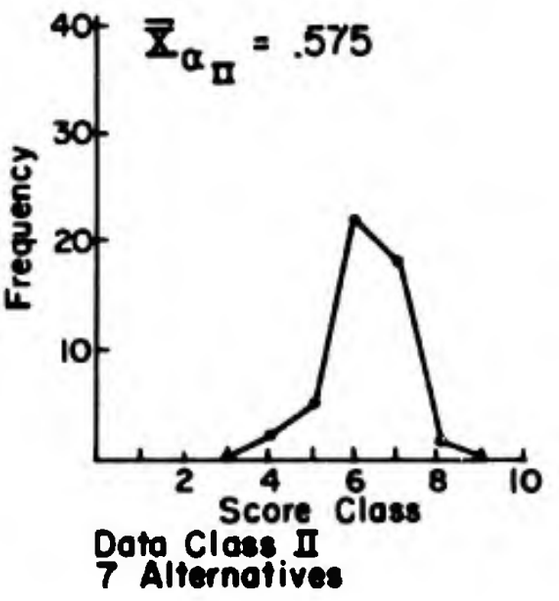
1. α_j is defined on page 16 of this report.

2. Score class code used in the distributions below and on the next 6 pages:

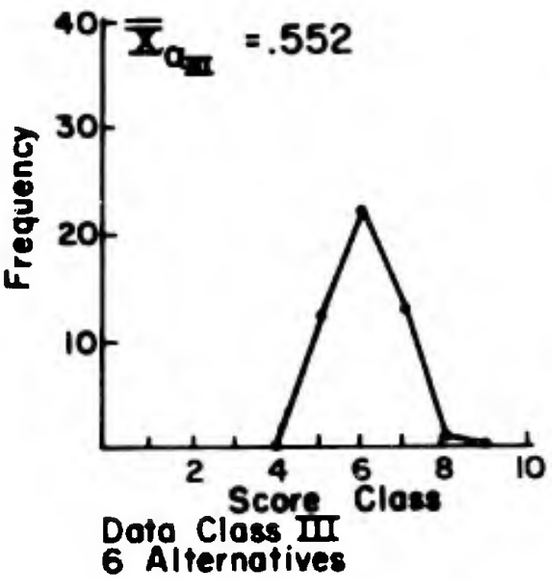
Score Class Code Number	Range of α_j Values
1	.000 - .104
2	.105 - .204
3	.205 - .304
4	.305 - .404
5	.405 - .504
6	.505 - .604
7	.605 - .704
8	.705 - .804
9	.805 - .904
10	.905 - 1.000



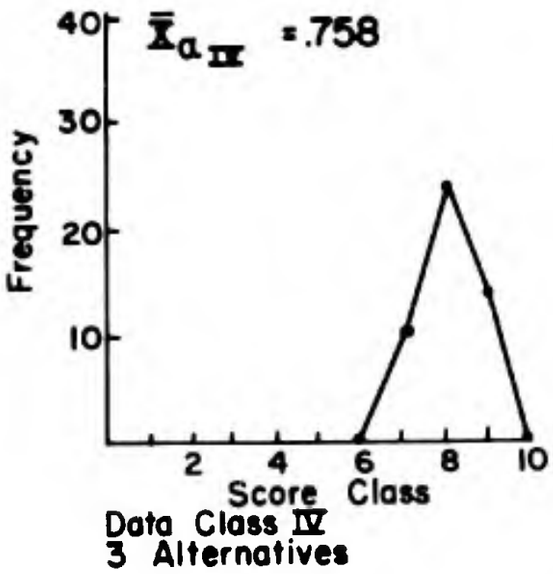
Data Class I
8 Alternatives



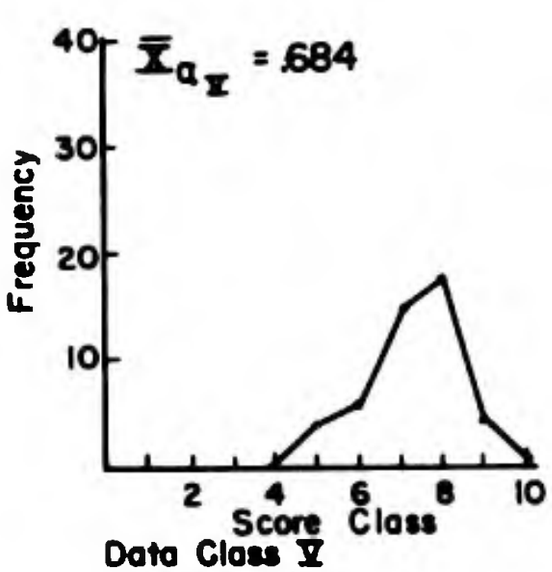
Data Class II
7 Alternatives



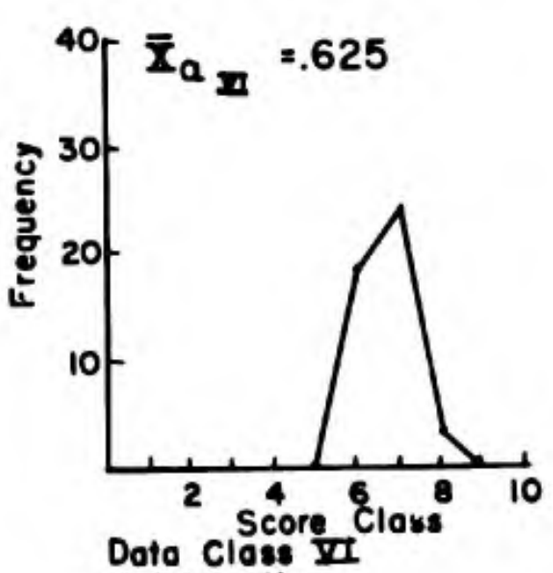
Data Class III
6 Alternatives



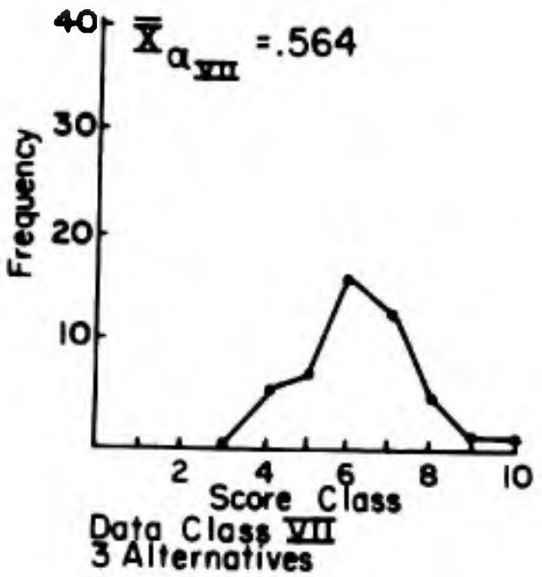
Data Class IV
3 Alternatives

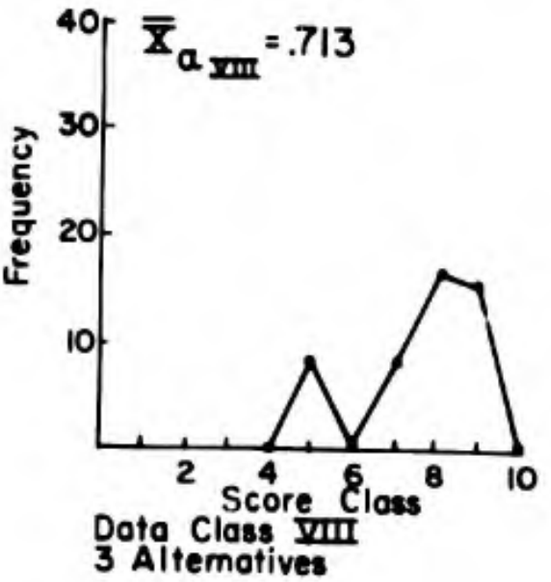


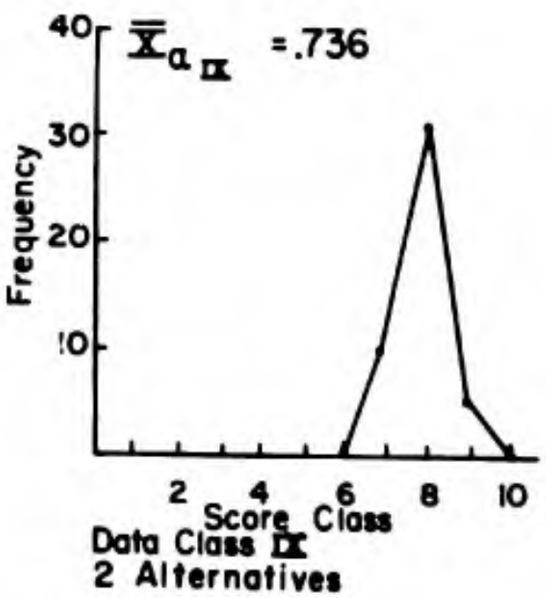
Data Class V
3 Alternatives

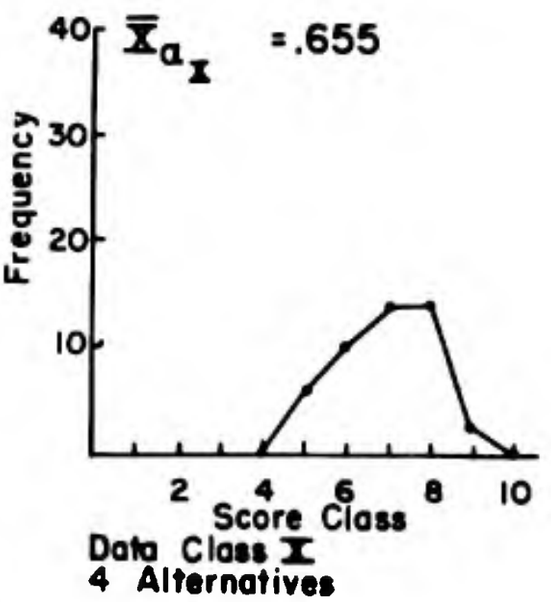


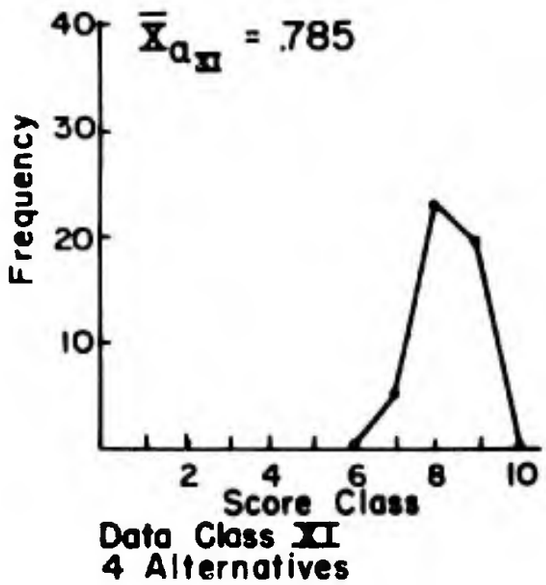
Data Class VI
3 Alternatives

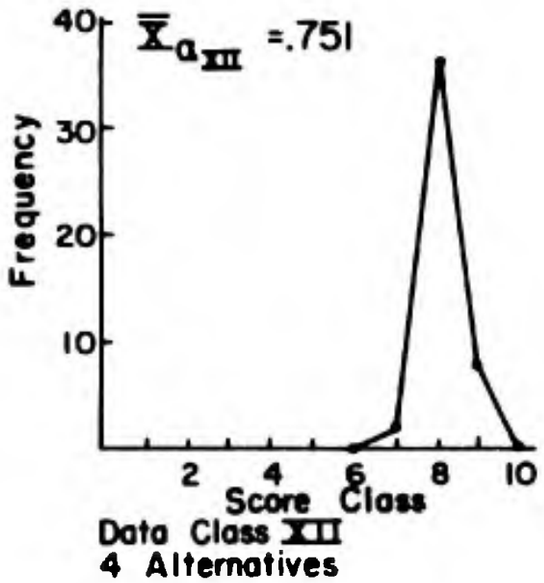

$\bar{X}_{a_{VII}}$ =.564
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class VII
3 Alternatives

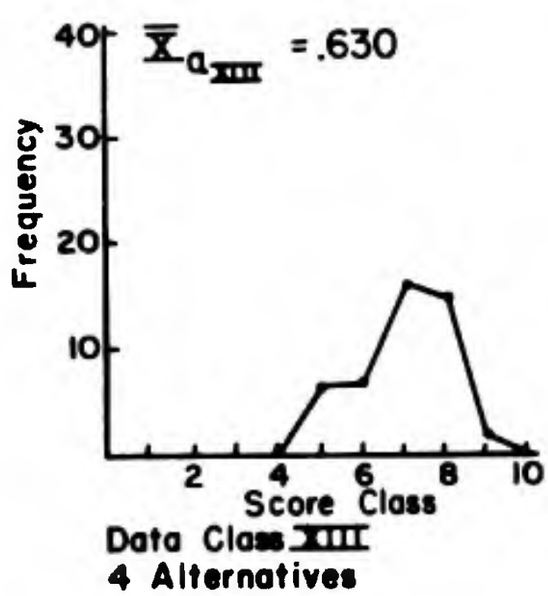

$\bar{X}_{a_{VIII}}$ =.713
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class VIII
3 Alternatives

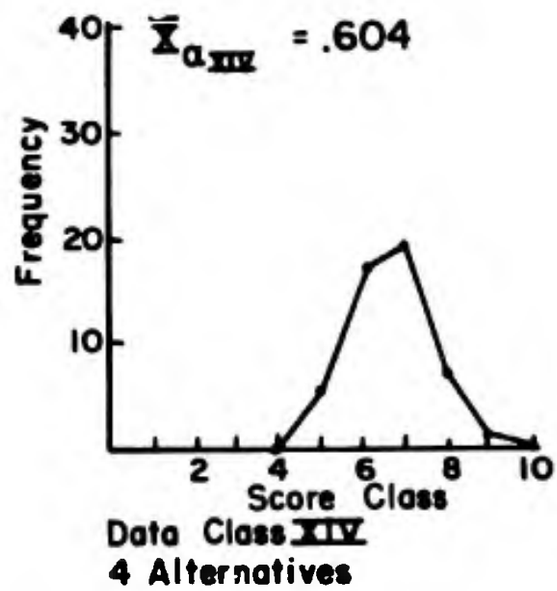

$\bar{X}_{a_{IX}}$ =.736
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class IX
2 Alternatives

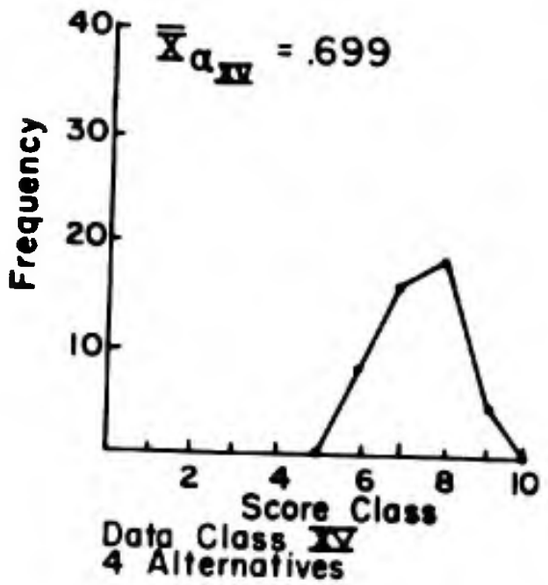

$\bar{X}_{a_{X}}$ =.655
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class X
4 Alternatives

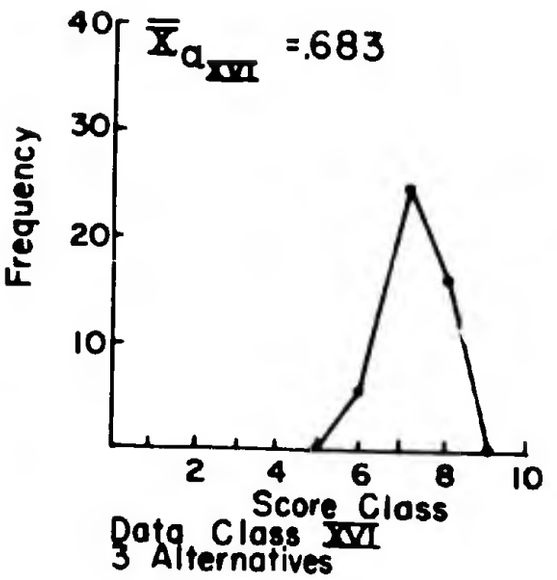

$\bar{X}_{a_{XI}}$ = .785
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XI
4 Alternatives

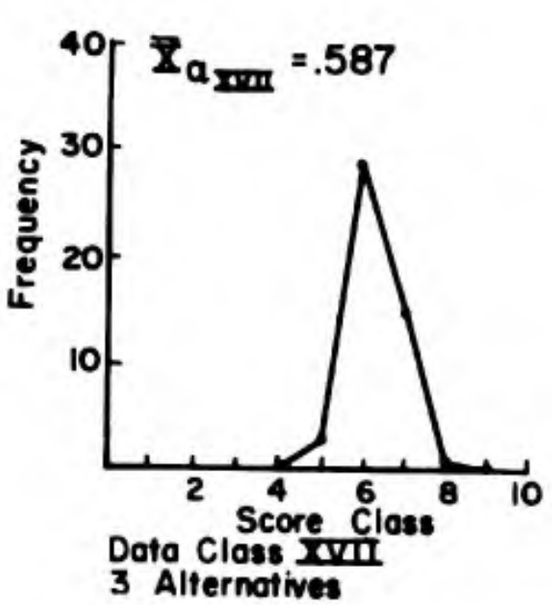

$\bar{X}_{a_{XII}}$ =.751
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XII
4 Alternatives

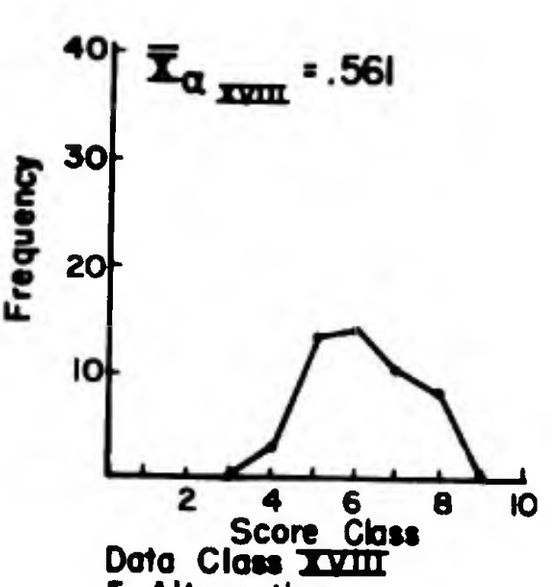

$\bar{X}_{a_{XIII}}$ = .630
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XIII
4 Alternatives

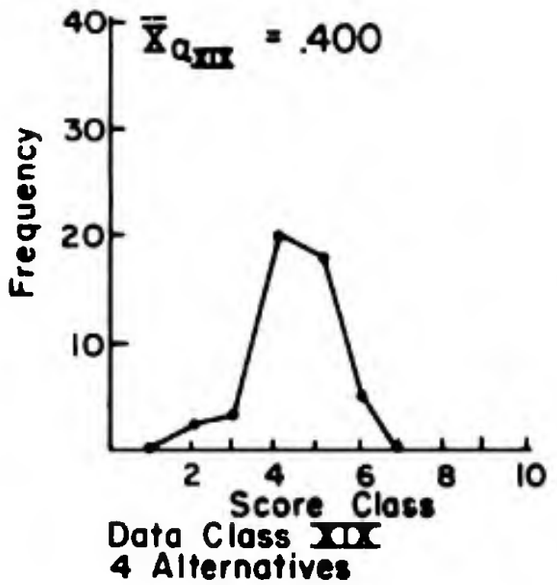

$\bar{X}_{a_{XIV}}$ = .604
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XIV
4 Alternatives


$\bar{X}_{a_{XV}}$ = .699
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XV
4 Alternatives

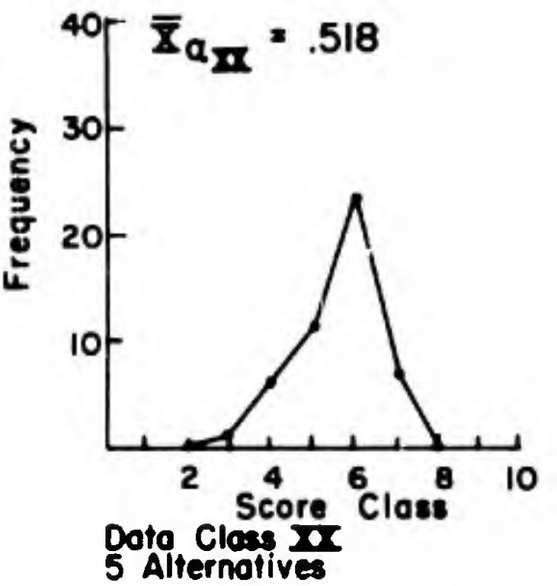

$\bar{X}_{a_{XVI}}$ =.683
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XVI
3 Alternatives


$\bar{X}_{a_{XVII}}$ =.587
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XVII
3 Alternatives

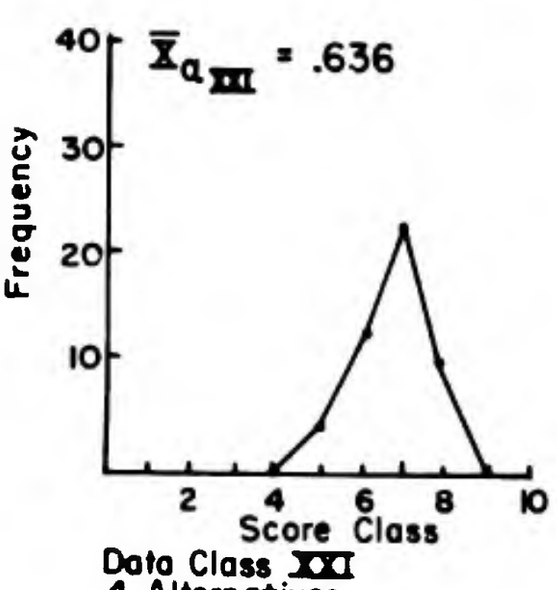

$\bar{X}_{a_{XVIII}}$ = .561
Frequency
40
30
20
10
2
4
6
8
10
Score Class
Data Class XVIII
5 Alternatives



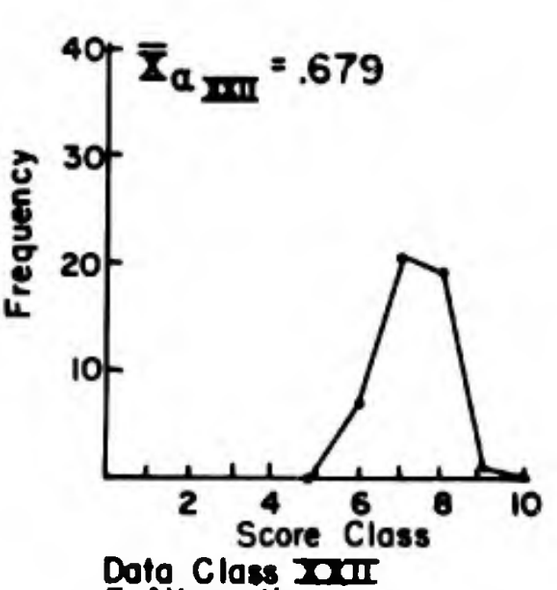
Data Class XIX
4 Alternatives



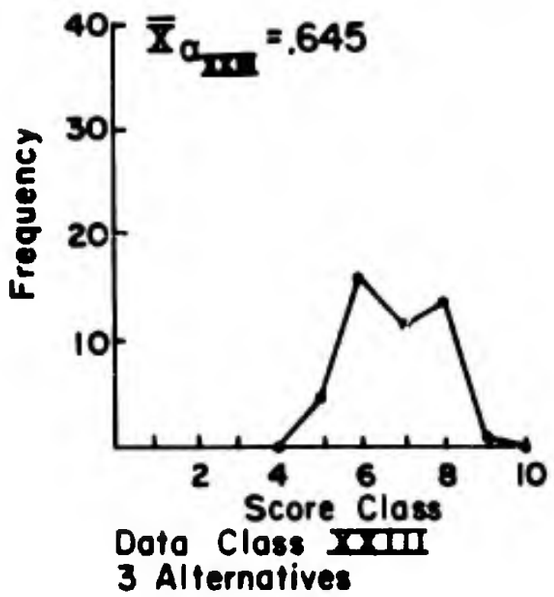
Data Class XX
5 Alternatives



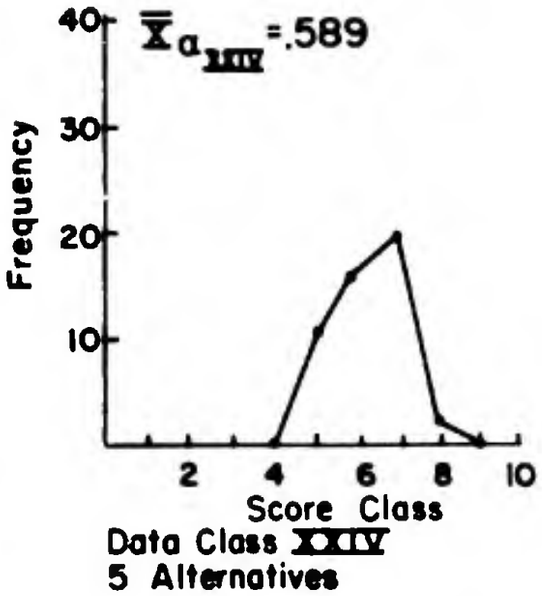
Data Class XXI
4 Alternatives



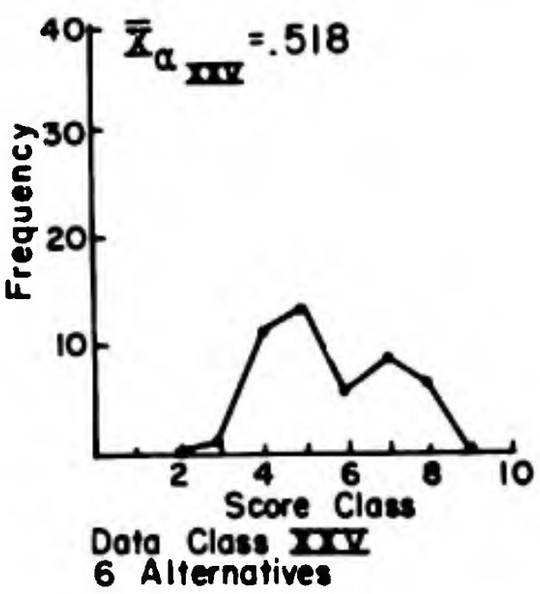
Data Class XXII
3 Alternatives



Data Class XXIII
3 Alternatives



Data Class XXIV
5 Alternatives



Data Class XXV
6 Alternatives

BLANK PAGE

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA · R&D

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

1 ORIGINATING ACTIVITY (Corporate author)

Ohio State University
Columbus, Ohio

2a REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b GROUP

N/A

3 REPORT TITLE

SUBJECT CONTROL OVER A BAYESIAN HYPOTHESIS-SELECTION AID IN A COMPLEX INFORMATION-PROCESSING SYSTEM

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final report, 1 June 1963 - 15 April 1964

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

Southard, Jack F.
Schum, David A.
Briggs, George E.

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
September 1964	47	7

8a CONTRACT OR GRANT NO

AF 33(657)-10763

b PROJECT NO

7184

c Task No.

718403

d

9a ORIGINATOR'S REPORT NUMBER(S)

9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

AMRL-TR-64-95

10 AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC.
Available, for sale to the public, from the Office of Technical Services, U.S. Department of Commerce, Washington, D. C. 20230.

11 SUPPLEMENTARY NOTES

12 SPONSORING MILITARY ACTIVITY

Aerospace Medical Research Laboratories
Wright-Patterson Air Force Base, Ohio

13 ABSTRACT The second experiment in a series devoted to estimating the effectiveness of automated hypothesis selection in man-machine systems in which threats are evaluated is described. An 8-man team produced evaluations of various threats posed by a hypothetical aggressor. These evaluations were made on the basis of intelligence information gathered on simulated reconnaissance overflights of the homeland area of the aggressor. IBM 1401 and 7090 computer facilities generated the complex stimulus environment. The primary output from this threat evaluation team was a series of a posteriori probabilities estimations produced by the team's commanding officer (CO). In three of the four experimental conditions the CO was provided with a hypothesis-selection aid based upon a modification of Bayes' theorem (MBT). In these three conditions the CO was permitted to exert an increasing amount of control over the MBT-aid mechanism. He exerted control either by adjustment of certain parameters in the MBT model or by direct insertion of conditional probabilities into the model. The purpose of the experiment was to observe whether increasing control over the MBT-aid mechanism would increase the user's acceptance of the aid and improve his threat-diagnosis performance. The CO's threat-evaluation performance improved, but independently of the MBT-aid configuration. Solutions of a posteriori probabilities based upon the MBT were calculated by the experimenter for comparison with human estimates, which were strikingly similar. The overall difference between the accuracy of the CO and MBT estimations was negligible.

DD FORM 1 JAN 64 1473

AF-WP-B-AUG 64 400

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Decision making, psychology						
Simulation						
Mathematical prediction						
Information Retrieval						
Human engineering						
Military psychology						
Computers						
Bayes Theorem						

INSTRUCTIONS

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

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

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

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

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

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

BLANK PAGE