THE UNIVERSITY OF MICHIGAN

INDUSTRY PROGRAM OF THE COLLEGE OF ENGINEERING

CONTINUOUS HUMAN ESTIMATION OF A TIME-VARYING, SEQUENTIALLY DISPLAYED, PROBABILITY

Gordon H. Robinson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
University of Michigan
Department of Aeronautical and Astronautical Engineering
1962

August, 1962

IP-578

Doctoral Committee:

Professor Paul M. Fitts, Co-chairman
Associate Professor Elmer G. Gilbert, Co-chairman
Associate Professor Frederick J. Beutler
Associate Professor Ward D. Edwards

ACKNOWLEDGEMENTS

The author is greatful to Professor Elmer G. Gilbert for his guidance and instruction during the author's graduate program, and to Professor Paul M. Fitts for his considerate leadership into the field of Engineering Psychology.

Professor Arthur W. Melton, Director of the Engineering Psychology Group, provided support and direction for the laboratory under which this work was performed.

Special gratitude must be extended to Professor Ward Edwards for his direct supervision of the work and his many suggestions concerning interesting directions of investigation. The author also greatly appreciates the intellectual stimulation and leadership that Professor Edwards has provided during the past three years.

Support for this work was provided by Project MICHIGAN under Department of the Army Contract DA - 36-039 SC - 78801, administered by the U.S. Army Signal Corps.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF SYMBOLS

a	a constant
C	convergence, a measure made on the subject's response (see page 17)
$\overline{C}$	no convergence, a measure made on the subject's response (see page 18)
D	detection, a measure made on the subject's response (see page 17)
$\overline{D}$	no detection, a measure made on the subject's response (see page 17)
$e(n)$	a model's error at n, $e(n) = r(n) - P_2$
$\overline{e^2(n)}$	the expected value of $e^2(n)$, an ensemble average
$\langle \overline{e^2} \rangle_A$	the average value of $\overline{e^2(n)}$ over some set of samples A
$e_M(n)$	the descriptive models error at n, $e_M(n) = r(n) - P_2$
$e_{MS}(n)$	the error between the subject and the descriptive model at n, $e_{MS}(n) = R_n - r(n)$
$e_S(n)$	the subject error at n, $e_S(n) = R_n - P$
$\langle e_S^2 \rangle_P$, $\langle e_M^2 \rangle_P$, $\langle e_M e_{MS} \rangle_P$	the average values of $e_S^2(n)$, $e_M^2(n)$ and $e_M(n)e_{MS}(n)$ over a problem
$E(x)$	the expected value of x
FAR	false alarm rate, a measure made on the subject's response (see page 18)
fps	flashes per second, the flash rate
IC	initial convergence, a measure made on the subject's response (see page 18)
k	$\frac{1}{T} \sum_{i=1}^{i=s} (P_{i-1} - P_i)^2$
k_1	the decision criterion level in the descriptive model
k_2	the fractional response adjustment in the descriptive model

k_3	the number of flashes in u(n) in the descriptive model
k_4	the flash shift between the subject's and the descriptive model's responses
M	the last flash in a sub-problem
M_i	the length of sub-problem i
ME,C	mean error after convergence, a measure made on the subject's response (see page 18)
n	a flash index, n = 1 is the first flash in a sub-problem
N	the total number of flashes in the model's summation
P	the probability of a 1 in a 0,1 binary series
P_i	the probability of a 1 in the binary series i
Pr(x)	the probability of x
r	the geometric ratio in the geometric model
r(n)	a model's response at n
$\overline{r(n)}$	the expected value of r(n), an ensemble average
R_n	the subject's response at n
RMSE	the square root of the mean squared error, a measure made subject's response (see page 18)
RMSE,C	the square root of the mean squared error after the point of convergence, a measure made on the subject's response (see page 18)
ρ	a correlation coefficient
S	the total number of sub-problems in a problem
s_n	the n^{th} flash in a binary series
σ_i^2	the variance of a sample from a population with $P = P_i$
$\sigma_r^2(n)$	the variance of r(n) at n, an ensemble average
$\sigma_s^2(n)$	the variance of s_n at n, an ensemble average

T	the total number of flashes in a problem
$u(n)$	an average of k_3 flashes in the descriptive model
v	$\frac{1}{S}\sum_{i=1}^{i=s}\sigma_i^2$
w_i	a weight attached to sample s_{n-i+1}

I. INTRODUCTION

Human decision tasks can be described as static or dynamic. In a dynamic decision task some of the relevant stimuli vary as a function of time, or of past decisions, or both. The decision maker must keep track of these changes in order to perform satisfactorily.

This experiment examines the human ability to follow or estimate a time-varying probability. This probability could be an important input to a dynamic decision task. This experiment attempts to isolate the estimation of the probability from the utilization of the estimate in making decisions. The task selected was the estimation of the mean of a binary (Bernoulli) distribution. Samples (0 or 1) from the distribution were displayed sequentially and the subject made a continuous estimate of the distribution mean. The mean varied with time. The experimental details are described in Chapter II.

The study of probability estimation isolated from decision making is important for two reasons. In decision making under uncertainty the estimation of probabilities is always at least an implicit part of the task. The ability of the decision maker to produce decisions according to a maximum expected value criterion will depend directly on his ability to estimate the probabilities of the various alternative courses of action.

There is also an applied interest in the human ability to turn uncertainty into probability. In many systems involving stochastic inputs it is relatively easy to automate the application of

decision rules. It is far harder, however, to find automatic means for supplying the probabilities and payoffs necessary for the application of the rules. Probability estimation is a candidate for inclusion as a human task in semi-automatic information processing and decision making systems in which the subsequent choice of a course of action is performed automatically.

Research on Estimation and Prediction

Human binary choice behavior has been studied extensively. This experiment complements previous studies by isolating the estimation function and by introducing dynamic probabilities.

Most of the studies on human binary choice have not included estimation as an explicit part of the subjects' task. These studies generally generate prediction data which are averaged over blocks of trials (decisions, choices) to produce prediction frequencies. A prediction frequency of 0.67 on trials 121 through 150 would indicate that the predictions during these 30 trials were distributed about two to one between the two choices. The subject may, or may not, be told the correct choice after each trial. The correct choices are drawn in some manner from a stationary binary distribution.

Examples of these experiments, often called probability learning experiments, are Grant, 1953; Hake and Hyman, 1953; Hake, 1954; Estes, 1957; and Neimark and Shuford, 1959. Most of these studies report prediction frequencies approaching asymptotically the frequency of correct choice or the generating probability. This phenomenon has been named "probability matching". This behavior is

not optimum. The optimum strategy, under instructions to maximize correct choices, is to predict consistantly the event having the larger probability. This event can be inferred from the relative frequency of previous events.

Behavior significantly different from matching has been reported by Gardner, 1959 and Edwards, 1961. The number of trials may have been insufficient in some of the experiments in which matching was found. An unpublished experiment by Tannenbaum and Edwards at The University of Michigan indicates an interaction between the amount of reward for a correct choice and the prediction frequency. Some subjects used near-optimum strategies.

A few studies have looked at the estimation ability of the binary decision maker. Grant (1953) reports an experiment by Hornseth in which the subject was asked to guess, at the end of 150 choice trials, which event had been the more frequent. The prediction frequencies for the last 30 trial block were close to the matching level. The data on guessing the overall frequency was plotted as the percentage of guesses that one of the events was the more frequent against the actual event frequency. These data showed the percentage of correct guesses at a particular event frequency to be higher than the event frequency. (i.e., an event frequency of 0.70 would be guessed to be the more frequent more than 70 percent of the time.)

Grant concludes from this experiment and presumably from other probability learning experiments that the processes of estimation and prediction are distinct and that prediction is the more

accurate. The processes may indeed be distinct but the accuracy of estimation was not measured by Hornseth's experiment.

Hake (1954) surveys a major portion of the probability learning literature including experiments by Estes, Grant, Hornseth, Hake, and Hyman and concludes that estimation is not accurate enough to be the basis for binary predictions.

The choice behavior in these experiments should have been based on an estimate of the event frequencies or generating probabilities. To conclude that the non-optimum performance was an indicator of probability estimation ability is unjustified, however.

Neimark and Shuford (1959) included estimation as an explicit part of the task in a probability learning experiment. In addition to making a choice at each trial some subjects were required to estimate the proportions of the past events. The events frequency was 0.67. These subjects gave an unbiased estimate and had prediction frequencies significantly higher than the matching level. Subjects who predicted only had prediction frequencies at the matching level. These results suggest the explicit estimation improved prediction.

Erlick (1959) looked at estimation without a decision task. He presented 100 binary events at a rate of five per second and asked for an estimate of the more frequent event and for an actual estimate of the event frequency on a continuous scale. Four event frequencies were used: 0.50 - 0.50, 0.48 - 0.52, 0.45 - 0.55, and 0.43 - 0.57. The data indicated that the more frequent event would be selected

correctly 75 percent of the time if the event frequency difference was approximately 0.08 (0.46 - 0.54). The median estimate of the frequency was within 0.01 for 0.50 and 0.52; 0.55 and 0.57 were each estimated about 0.02 high.

All of the experiments reviewed above used stationary processes to generate the binary events. A few experiments have used a dynamic generating process. These experiments have all used prediction as the required task, however, and thus give only indirect evidence on estimation.

Grant (1953) reports an experiment in which the generating probability changed periodically as a square wave. The probability values always differed by 0.50 with higher values: 1.00, 0.90, 0.80, and 0.70. The period was 40 events and two and one-half cycles were presented. A prediction frequency was calculated by averaging over five trials and about 40 subjects. This prediction frequency followed the cyclic change only when the higher probability was 1.00 or 0.90; reaching 0.95 in 20 trials at 1.00 and 0.70 in 20 trials at 0.90. No systematic performance changes apparently occured during the two and one-half cycles.

The subjects were evidently not instructed about the non-stationarity of the generating process. Such instructions would probably have an appreciable effect.

Goodnow and Pettigrew (1955) performed a binary prediction experiment in which a change from 0.50 - 0.50 to 0 - 1.00 occurred. They found that the response to such a change was more rapid when the

subjects had initially experienced a 0 - 1.00 series prior to the 0.50 - 0.50 series. Again, however, no specific instructions were given concerning the non-stationarity of the generating process.

In both the Grant and Goodnow experiments there is evidence that a change in the generating probability of 0.50 will produce an appropriate change in the prediction frequency if the change is to an extreme probability. These extreme probabilities (1.00 and 0.90) evidently represent obvious changes even with no expectancy for change induced by the instructions.

Flood (1954) discusses the strategy of a subject who may not be convinced that the probabilities are stationary. No particular results were obtained in an experiment designed to induce certainty versus uncertainty in the stationarity of a stationary generating process.

The human ability to give a direct magnitude estimate of a stationary binary probability is uncertain. Most experimenters have postulated estimation only as an intervening variable between the display and a decision task. Decision behavior was improved in one experiment by including explicit estimation in the task. Two questions seem appropriate: What role does estimation play in a decision task, and how well can this estimation be performed. This experiment sheds light on the second question as well as providing a fairly comprehensive look at the continuous estimation of dynamic probabilities.

II. THE EXPERIMENT

Task Selection

The task was estimating the mean of a binary distribution as samples (individual drawings) from that distribution were sequentially displayed. This task was selected for two reasons. The binary distribution is completely described by one parameter, its mean. It is thus easily understood by people unfamiliar with the mathematical aspects of probability. The second reason for a binary presentation was its relevance to the binary decision and estimation literature discussed in Chapter I. Except where specifically defined otherwise, the word "probability" will refer to the current probability of drawing a right sample.

Samples from the binary distributions were presented at a fixed rate by flashing either the right or left light of the apparatus shown in Figure 1. Directly beneath the two lights is an illuminated dial indicating the position of the manual response lever. A continuous response mechanism was selected as appropriate for the estimation of a continuously varying stimulus. The display and response mechanisms were designed for convenient and effective interpretation and control. Details of the apparatus are given in Appendix A.

The lever was free to move between stops at 0 and 100 on the scale. The lever and associated mechanisms contained enough Coulomb friction to retain a setting without constant force. Neither springs nor viscous friction was used. The smallest scale division was 2. It was possible to position the pointer precisely to one half of the least division corresponding to a probability change of 0.01. A main scale division occurred at every fifth small division and was marked: 0, 10, 20, ..., 90, 100.

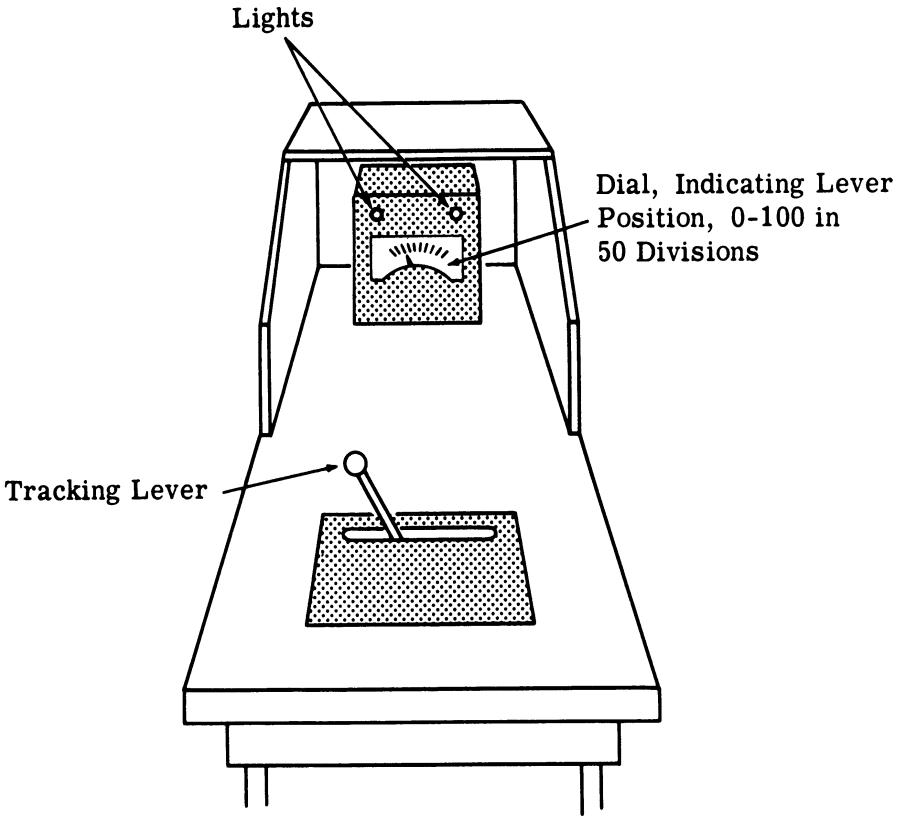


Figure 1. Tracking Console.

The position of the lever was recorded by two means, Friden punched paper tape and a Sanborn continuous strip recorder. The paper tape was punched in a Grey, or Cyclically Permuted binary code using six channels of an eight channel punch to encode 101 symbols, 0 through 100. Pilot studies indicated that the rate of output sampling necessary to recover the response information was dependent on the flash rate and that a response sampling rate equal to the flash rate would be adequate. A sample was thus taken every two seconds at the slowest presentation rate, 0.5 flashes per second, and every 0.125 seconds at the fastest presentation rate, 8 flashes per second. The punched tape record was later transferred to IBM cards on a modified IBM Tape to Card Converter and the data analyzed on an IBM 709 Data Processing System. The Sanborn records were used to make qualitative judgements about the response and to select appropriate criteria for the computer analysis. They also permitted continuous monitoring of the task as it was performed.

The subject and his console were housed in a small isolated room. The subject wore noise insulating ear muffs. He had a two-way communication system with the experimenter. A low level white noise was presented by the earphones during the experimental run. His microphone was always on, and comments during the experimental run were permitted.[1] When the experimenter spoke to the subject the noise was automatically switched off.

The task has a strong resemblance to a standard unidimensional manual tracking task. The main difference is the presentation of the

---

[1]Few comments were actually made. Most of these were not printable.

target. Instead of being displayed explicitly as a dot or a line it exists only as a parametric description of the method used to select the flash sequence. In this experiment the generating process was time-variant and the target could be defined as the mean of the distribution from which the last flash was drawn. It is impossible to recover the precise target from the information available to the subjects. The cursor, or 0-100 dial pointer in this case, is pointed at an estimated value of the target. The system is essentially open loop since the lack of an explicit target prohibits the formation of an error signal. The dynamics are almost entirely in the mental computation and there was no indication that motor skill was a limiting factor.

The use of a tracking lever as a response means is unique in research on probability estimation. It is appropriate to the task and permits an easy understanding of the response scale by the subjects. Both end points are well fixed in the same sense that events which are impossible and certain are fixed in value on a personal or subjective probability scale. The 50 point on the scale might also be considered as an anchor point in the sense that all subjects clearly understood that 50 per cent meant equally frequent flashes.

Input Selection

The input probability changed in a series of discrete steps. This input form permitted visual, qualitative interpretations to be made from the data as well as the more extensive analysis done by the computer. It also permitted static as well as dynamic measurements to be made. The step change sizes and their directions as well as the number of flashes

between steps were selected randomly from a finite set of values. The mechanism for the generation of the sequences is described in detail in Appendix B. The sequence of steps so generated is called a problem.

Preliminary investigations revealed that step changes ranging from 0.06 to 0.64 in eight values would adequately cover the interesting range of probability change.[2] The number of flashes between step changes was selected from a set ranging from 34 to 89 flashes. Thirty-four flashes was considered the smallest number required to minimize interaction between successive step changes. The range between 34 and 89 was considered sufficient to prevent any performance improvement due to the learning of inter-step length. A step change and the flashes up to the next change is called a sub-problem.

Flash Series Generation

The flashes were drawn from finite populations without replacements. The population size was an experimental variable and is discussed below. Finite populations were selected to fix the average value of the flashes for each sub-problem. The effects of finite population sampling on variances are shown in Appendix C.

Experimental Variables

Five independent variables were used in the experiment: the rate at which the flashes were presented, the magnitude and sign of the step change, the probability that the step changed to, a constraint on the randomness of the flash series, and subjects. The number of flashes between step changes was not studied as a variable. The values of these

---

[2] A pilot experiment with a simplified apparatus was run before the main console was built in order to establish the general form of the response and reasonable ranges for the independent variables. It is discussed in more detail in Chapter III.

variables were as follows:

Rate; 0.5, 1.0, 2.0, 4.0, and 8.0 flashes per second, (fps)

Step size; 0.06, 0.12, 0.16, 0.18, 0.24, 0.32, 0.48, 0.64; both + and -

Probability; 0.02, 0.08, 0.14, 0.18, 0.26, 0.32, 0.34, 0.44, 0.50, and the complementary values between 0.50 and 1.00.

The step changes and probability values were arranged in two problem types as described in detail in Appendix B. For one type, the small step problems, the mean step change is approximately 0.15; for the other, the large step problems, the mean step change is approximately 0.40. Both types contain the entire range of probabilities and are symmetric about 0.50.

The constraint variable had two values leading to the random and the constrained problem types. The random problems were generated from finite populations which were the length of the respective sub-problems being generated. These finite populations were thus of size 35 through 89 flashes. It was felt that these sizes would be large enough to yield experimental results fairly close to those which would result from infinite populations.

The constrained problems were generated from finite populations of 17 flashes. The lengths of the sub-problems were arranged in whole number multiples of 17: 34, 51, 78, and 85 flashes. It was assumed that this constraint would be sufficient to indicate those aspects of performance that constraint would affect. It is not a severe enough constraint to be readily perceived from inspection of the flash series, however. The same series of steps and probabilities were used in the random and in the constrained problems.

Each of the four subjects performed the task in 15 sessions. Each session consisted of two or three problems separated by a short rest period. Each session lasted for about an hour. Each subject saw the same series of problems in the same order.

Rates, small and large step problems, subjects, and constraints were exhaustively combined. The order of presentation was chosen at random under the constraint that the tracking sessions were of about the same length. Appendix D gives the sequence used. The pilot experiment had indicated that about 25 minutes at two flashes per second was the maximum time that a subject could be expected to track without a significant decrement in his performance. The problems presented at 0.5 and 1.0 flashes per second were given in four and two separate sessions respectively in order to limit all sessions to a maximum of 25 minutes of continuous tracking.

Task Instructions

Careful attention was paid to the instruction of the subjects prior to the recorded experimental sessions. This effort was repaid with an excellent consistency in the tracking behavior of the eight subjects, four in the pilot and four in the main experiment. The standard instructions used are shown in Appendix E. These served only as the initial, formal introduction, however. Actually about 10 minutes was spent discussing the task to be performed and the purpose of the experiment. Instruction was concluded when the experimenter was satisfied that all important concepts were understood.

A 45 minute practice session was used preceding the 15 hours of data recording. During this session the response was continuously monitored

and the subject was assured of the quality of his performance. The lack of error feedback made it difficult for the subject to evaluate his own performance until he had some experience with the task.

The instructions were as complete as the subject seemed to need in all but one important area. He was told nothing about the dynamics of the input sequence aside from the fact that there would be changes in the probability. He was told to expect both rapid and slow changes. He was instructed that the pay he would receive would be a constant rate per minute of tracking minus the accumulated squared error during the same interval. The amount was computed automatically on an analog computer operating during the tracking session. The circuit used for the pay scheme is shown in Appendix F.

III. EXPERIMENTAL RESULTS

The Pilot Experiment

A pilot experiment was run prior to the main experiment in an attempt to answer three questions. The first concerned the general form and quality of the response. The responses found were qualitatively similar to the response in Figure 2. Both the response to change and the estimation of probability were better than expected. The second question answered by the pilot experiment concerned changes in response with continued performance of the task, reflecting learning or fatigue. The identical problems were presented to the four subjects in each of six sessions about two days apart. There was no indication of a significant change in performance after the first session. To test for specific problem learning the problem which had been presented for six sessions was presented again only backwards. No decrement in performance was observed on this reversed problem. It was concluded that no specific problem learning had occurred. None of the subjects recognized that the problem had been the same in each of the six sessions nor were they able to describe the changes which had occurred in the probabilities. Tracking sessions up to 15 minutes caused no particular fatigue or boredom and it was concluded that sessions of 25 minutes would be permissible on the more isolated, impressive, and comfortable console used in the main experiment. The third question answered by the pilot experiment concerned the kind and amount of instruction needed to bring the subjects up to a reasonably consistent level of performance. The instructional method described in Chapter II was the result. The subjects in the main experiment performed consistently after the instruction

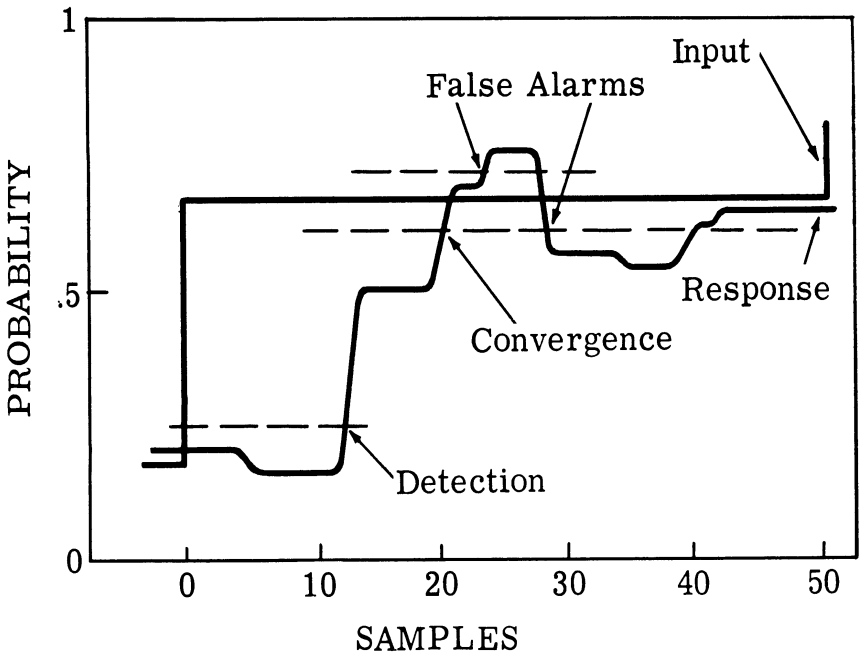


Figure 2. A Typical Response to a Sub-Problem.

and practice. Appendix G presents two interesting exceptions. The important task learning evidently occurs during the first few minutes of performance and the 45 minute practice session was sufficient.

Response Measures

The response measures were chosen after study of the Sanborn records from the main experiment. The form of these responses was the same as in the pilot experiment and shown in Figure 2. The response was characterized by fairly rapid changes separated by periods of little or no change. This discontinuous form indicates that the behavior might be described in terms of a series of decisions concerning changes in the probability. A descriptive model with this characteristic is developed in Chapter IV. Several of the response measures were chosen to fit this response form. All of the response measures refer to individual sub-problems. The following measures were calculated:

1. DETECTION, D: The number of samples from the step change to the point where the response has changed 0.05 in the direction of the new probability from its value at the point of the step change. If R_n is the response at point n in a sub-problem which starts at n = 1, the point of detection is where $R_n = R_0 \pm 0.05$, the plus sign for an increasing step and the minus for a decreasing step.

2. NO DETECTION, $\bar{D}$: The number of sub-problems in which detection did not occur; R_n never changed 0.05 in the direction of the new probability.

3. CONVERGENCE, C: The number of samples from the step change to the point where the response is within 0.05 of the new probability.

The point of convergence is where $R_n = P \pm 0.05$ where P is the probability following the step change. The point of convergence is the first entry into this region from either side.

4. NO CONVERGENCE, $\overline{C}$: The number of sub-problems in which convergence did not occur; R_n was always outside the 0.05 region about P.

5. INITIAL CONVERGENCE, IC: The number of sub-problems in which the response was within the convergence region about the new probability at the point of the step change. $P - 0.05 \leq R_0 \leq P + 0.05$

6. ROOT MEAN SQUARE ERROR, RMSE: The square root of the mean squared error over the entire sub-problem. Error equals the response minus the probability. The response was measured on a 0 to 1 scale corresponding to the probability measure and the error can thus be considered to be an error in probability. For a sub-problem of length M; RMSE =

$$\left[\frac{1}{M}\sum_{n=1}^{n=M}(R_n - P)^2\right]^{\frac{1}{2}}$$

7. ROOT MEAN SQUARE ERROR AFTER C, RMSE,C: The square root of the mean squared error from the point of convergence to the end of the sub-problem. $RMSE,C = \left[\frac{1}{M-C}\sum_{n=C+1}^{n=M}(R_n - P)^2\right]^{\frac{1}{2}}$ This measure and the following two were only made when either convergence or initial convergence was measured.

8. MEAN ERROR AFTER C, ME,C: The mean error from the point of convergence to the end of the sub-problem. $ME,C = \frac{1}{M-C}\sum_{n=C+1}^{n=M}(R_n - P)$

9. FALSE ALARM RATE, FAR: The number of times per sample that the response left the 0.05 convergence region between the point of convergence

and the end of the sub-problem. If $P - 0.05 \leq R_{n-1} \leq P + 0.05$ and $R_n < P - 0.05$ or $R_n > P + 0.05$ then the point n would be a false alarm point.

Detection and convergence were measures designed to describe the discontinuous response form. The 0.05 criterion used in these measures was selected after an extensive study of the data. A sudden response to the new probability occured shortly following a step change in about 80 percent of the sub-problems. This movement was interpreted to be the result of the perception of the change in the probability. The 0.05 detection criterion was selected as measuring this point with fair consistency. For step changes greater than about 0.15 this measure is relatively insensitive to the choice of the 0.05 criterion since the sudden response was characteristically 0.10 or greater.

Convergence is more dependent on the selection of 0.05 as a criterion. The point of convergence was most useful, however, in determining the beginning of measures 7, 8, and 9. These measures were all averaged over flashes and the location of the convergence point did not affect their values. Detection and convergence, as measured with the 0.05 criterion, are not particularly informative for the smallest step change, 0.06.

Measures 7, 8, and 9, the three starting at the point of convergence, indicate the subjects' static estimation ability. The subject is operating under what might be called a dynamic set, however. He has an expectancy for changes in the probability. Changes in his responses during this period could be called "microstructure" tracking since the subject was not aware that the probability was constant. No measures were made of the persistence of this microstructure tracking on the longer sub-problems.

This behavior might begin to diminish with long presentations of a constant probability.

RMSE was the only measure made on all sub-problems regardless of their response form. It indicates the overall quality of performance. RMSE is a common measure in continuous tasks of this kind largely because it is easily derived and manipulated in mathematical expressions.

Data Analysis

There was one sub-problem for each rate, step size, step direction, probability, constraint, and subject, 3440 in all. The combinations of variables presented here was judged to be the most informative set from among the total available from the computer analysis. These quantitative performance measures were the intended output of this experiment and since no testable hypothesis were generated no tests of statistical significance were made.

Experimental Data

Differences Between Subjects. No qualitative differences existed among the four subjects used in the main experiment. The four subjects in the pilot experiment behaved similarly. Inspection of the data indicated that for general performance information it would be best to average the data over subjects. Appendix H presents some of the subject-by-subject data.

Detection, D. Figures 3 through 6 show the effects of the independent variables on detection. The data on step direction showed no appreciable difference between positive and negative directions and they are averaged together in all figures. The interaction of step size and rate shown in

Figure 3 was the most interesting relation found. Detection decreases fairly linearly with step size and increases fairly linearly with the logarithm of rate.

The linear increase in detection with the logarithm of the rate probably reflects a combination of factors influencing the response. A small linear increase with rate would be caused by a constant reaction and movement time. For the usual tracking tasks this might be expected to be on the order of 0.5 seconds yielding a lag of 2 flashes at 4 fps and 4 flashes at 8 fps. The more important factor is probably a change in the method of performing the task as rate changes. At rates of 0.5 and 1 fps the subjects reported counting the flashes at times, occasionally counting the number of flashes of the lower frequency and comparing this to an estimate of the total number of flashes. They did not use any procedure of this sort consistently, however; at least not one apparent to them. They all reported that the rate of 2 fps was the most difficult. Evidently the methods which they had used effectively at 0.5 and 1 fps became difficult if not impossible at 2 fps. Beginning at 4 fps it is clearly impossible to respond to separate flashes and the series is probably perceived in groups of flashes. The task becomes similar to a continuous tracking task at these rates. Reese (1943) postulated a change in the mechanism of counting light flashes at about 4 flashes per second.

Figure 3 shows an effect due to the presentation of the step changes in two separate series, the small and the large step problems. There is a region of overlap in step size between these two problems.

The smallest change in the large step problem is 0.16 and the largest in the small step problem is 0.24. In this overlapping region the small step problem yields detections of from one to six flashes higher than the large step problem at all rates. The subjects were evidently modifying their tracking method according to the type of problem being presented. The large and small step problems were ordered randomly, of course, and the subjects had no prior indication that there were two problem types. This change is perhaps not surprising considering the difference between the two problem types. The average step changes were 0.15 in the small step problem and 0.40 in the large step problem. Step changes of about 0.30 and larger are readily noticed.

The subjects appear to have made larger, more decisive, response changes on the large step problem than on the small step problem. This change to a more responsive behavior is appropriate in quickly reducing the large errors following the larger step changes.

"No detection" as a percentage of the total sub-problems is shown in Figure 4 as a function of step size and rate. About 90 percent of the "no detections" occurred with the combination of rate above 4 fps and step size below 0.15. Some of the "no detections" were probably caused by occasional lapses of attention. At 4 fps a 42 flash sub-problem is over in 11 seconds. In 25 minutes of continuous tracking a few 11 second lapses are certainly reasonable.

The effect of probability on detection is shown in Figure 5. Perhaps the most interesting finding is that detection is not appreciably smaller for the extreme probabilities. In Chapter IV it will be seen that

responses generated by simple running averages produce detections which are similarly independent of probability.

The variability among a set of detections of a particular step size and rate will depend on the probability, however. Detections of central probabilities, those near 0.5, will have more variability than detections of extreme probabilities, those nearer 0 or 1.

The effect of the flash generation constraint on detection is shown in Figures 5 and 6. Constraint has no particular effect on average detection. As in the case of probability, however, constraint will effect the variability of a set of detections. The constrained problems will yield less variable detections.

Convergence, C. Figures 7 through 11 show the effects of the independent variables on convergence. The interesting effects are again with step size and rate. The effect of rate on convergence is similar to its effect on detection; a linear increase in convergence with the logarithm of rate. The effect of increasing step size is to increase convergence, although the increase is small. The number of flashes between detection and convergence increases as step size increases. This probably reflects the size of the response more than any other factor. The majority of sub-problems show a response successively approaching the new probability rather than one that overshoots.

Convergence shows a similar difference to that noted in detection between the small and the large step problem in the region of overlapping step size.

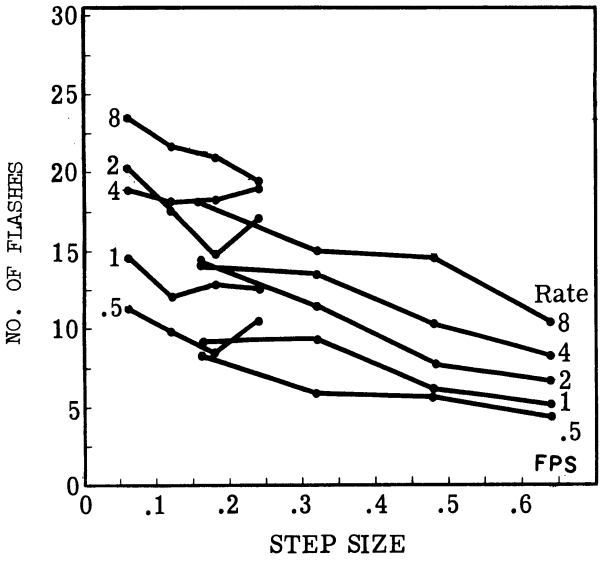


Figure 3. Detection as a Function of Step Size and Flash Rate. The small and large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64. Detection is measured in flashes.

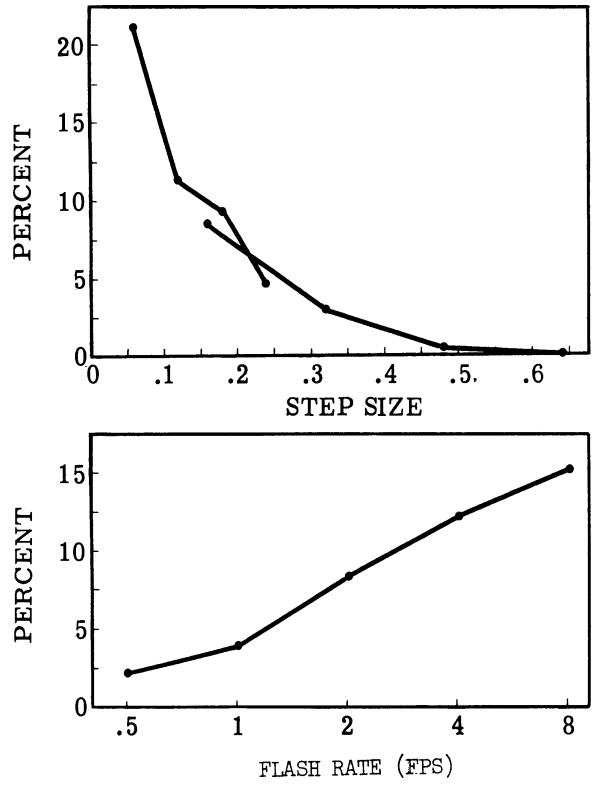


Figure 4. Percentage of Sub-Problems in which "No Detection" Occurs as a Function of Step Size and Flash Rate. The small and the large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64.

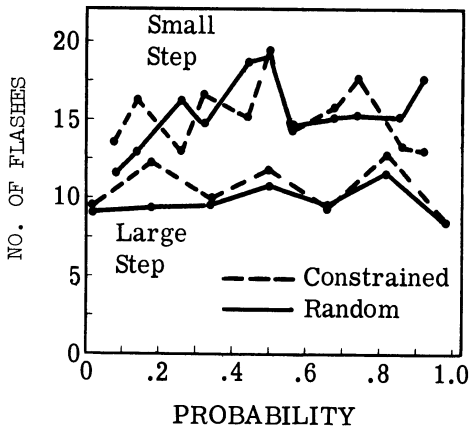


Figure 5. Detection as a Function of Probability, Constraint, and Small and Large Step Problems. Detection is measured in flashes.

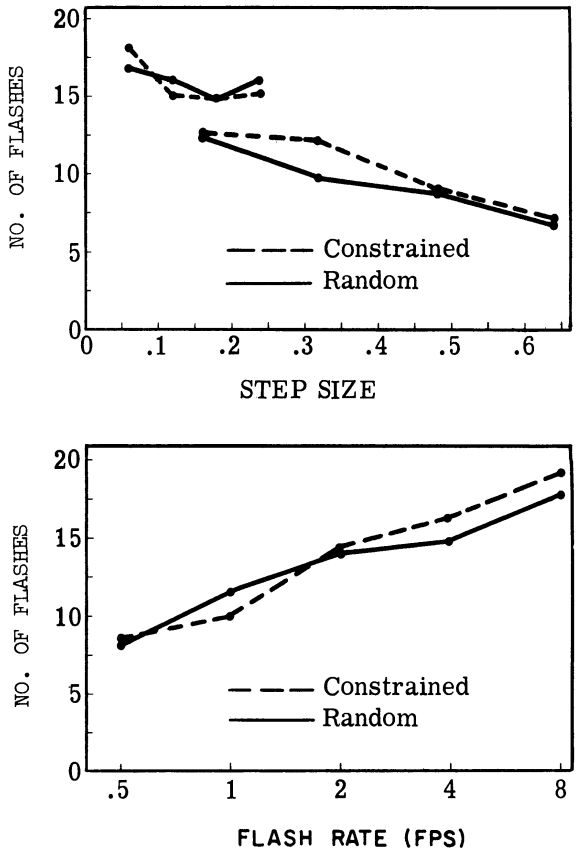


Figure 6. Detection as a Function of Step Size, Sample Rate, and Constraint. The small and large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64. Detection is measured in flashes.

"No convergence," expressed as a percentage of sub-problems, is shown in Figure 8. "No convergence" remains relatively insensitive to changes in step size except for the largest step, 0.64, where it is zero. It is approximately 10 per cent for the large step problem and 12.5 per cent for the small step problem. "No convergence" rises sharply with increasing rate, reaching about 28 per cent at 8 fps. This is consistent with the data which show convergence equal to 35 flashes at 8 fps; about the length of the shortest sub-problem.

"Initial convergence" has a high of 35 per cent for a step change of 0.06 and goes to zero for steps of 0.48 and 0.64. It increases slightly dwith rate from about 8 to 12 per cent.

The relationship between probability and convergence is shown in Figure 10. Convergence is relatively insensitive to probability as was detection.

The effects of constraint on the sample generation are shown in Figures 10 and 11. Again, as with detection, there is little if any effect.

Root Mean Square Error, RMSE. This measure was introduced to provide a single, overall indicator of the task performance. The most informative variation of RMSE is with rate as shown in Figure 12. RSME increases linearly with rate from 1 to 8 fps. It is interesting to evaluate this performance measure on a time basis as it might be in a situation where it was desirable to perform the estimation in the shortest time possible. RSME divided by fps yields values of error-seconds per flash which decrease as rate increases, going from 0.134 at 1 fps to 0.022 at 8 fps. This decrease might well continue with even higher rates as the task becomes the

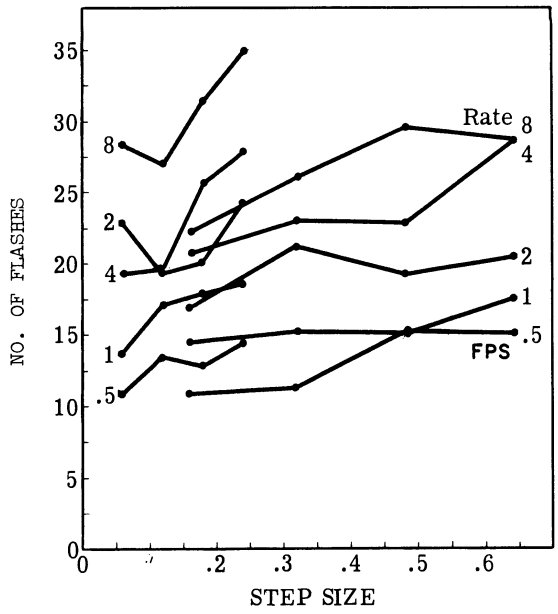


Figure 7. Convergence as a Function of Step Size and Flash Rate. The small and large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64. Convergence is measured in flashes.

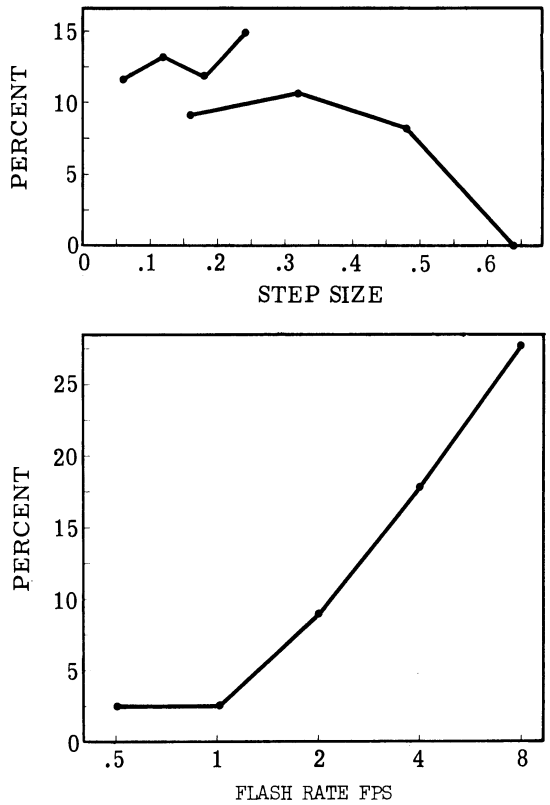


Figure 8. Percentage of Sub-Problems in which "No Convergence" Occurs as a Function of Step Size and Flash Rate. The small and large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64.

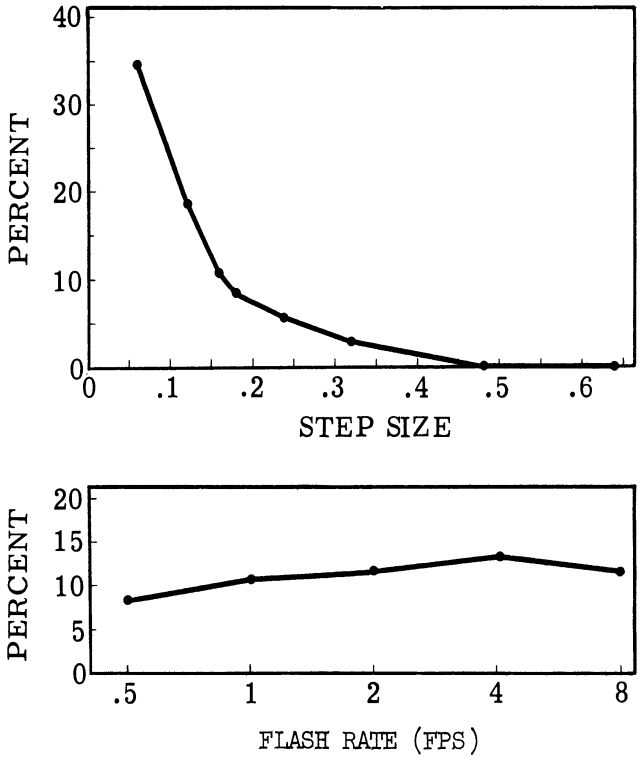


Figure 9. Percentage of Sub-Problems in which "Accidental Initial Convergence" Occurs as a Function of Step Size and Flash Rate.

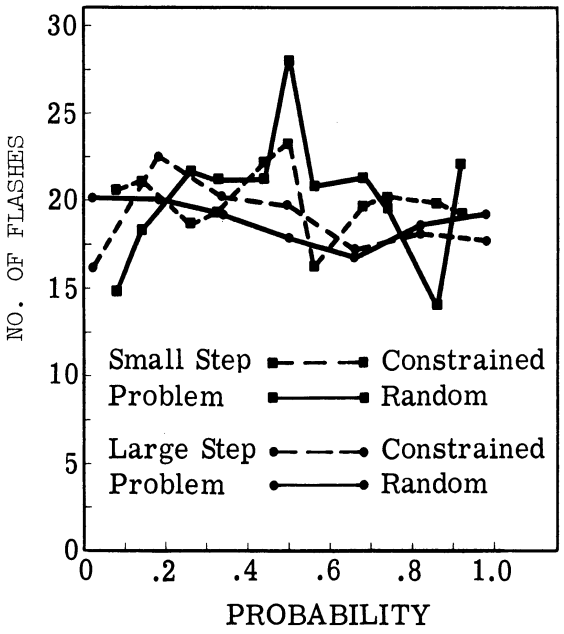


Figure 10. Convergence as a Function of Probability, Constraint, and Small and Large Step Problems. Convergence is measured in flashes.

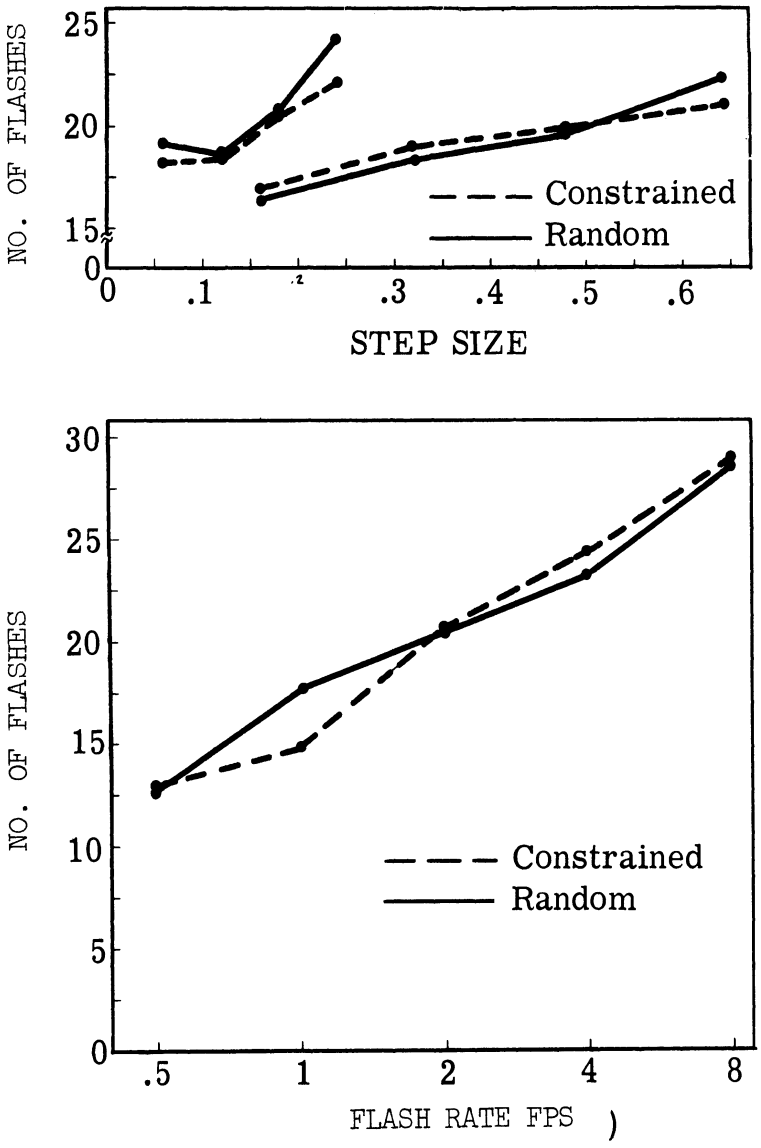


Figure 11. Convergence as a Function of Step Size, Sample Rate, and Constraint. The small large step problems are plotted separately, the small extending from 0.06 to 0.24 and the large from 0.16 to 0.64. Convergence is measured in flashes.

tracking of the relative brightness of the lights. Either the limitations on the judgment of relative brightness or simple reaction time would finally limit the performance. This performance index must be viewed with caution. The error itself has a meaningful upper bound at the level where the lever is left stationary or is moved in some manner independent of the flashes. As this error level is approached further increases in rate would continue to decrease the index of error-seconds per flash but the index would have little meaning.

The following three measures were made from the point of convergence to the end of the sub-problem or from the beginning of the sub-problem if "initial convergence" occurred. They are therefore measures made on an average of 85 to 90 per cent of all of the sub-problems and on about 95 per cent of those sub-problems with step changes above 0.15 at rates below 4 fps.

Mean Error After Convergence, ME, C. The mean error is shown as a function of probability in Figure 13. The average estimate is essentially unbiased at all probabilities. The largest error is smaller than the least scale division on the subject's response indicator (0.02). Mean error was not significantly affected by rate, constraint, step size, or subjects.

This finding contradicts a body of conjecture based in part on the results of static estimation and choice experiments. Neither the overestimation of high nor the underestimation of low probabilities appears. The two distinctive features of this task were the dynamic estimation and the tracking lever as a response mechanism. The excellence in static estimation was undoubtedly due at least in part to these two features.

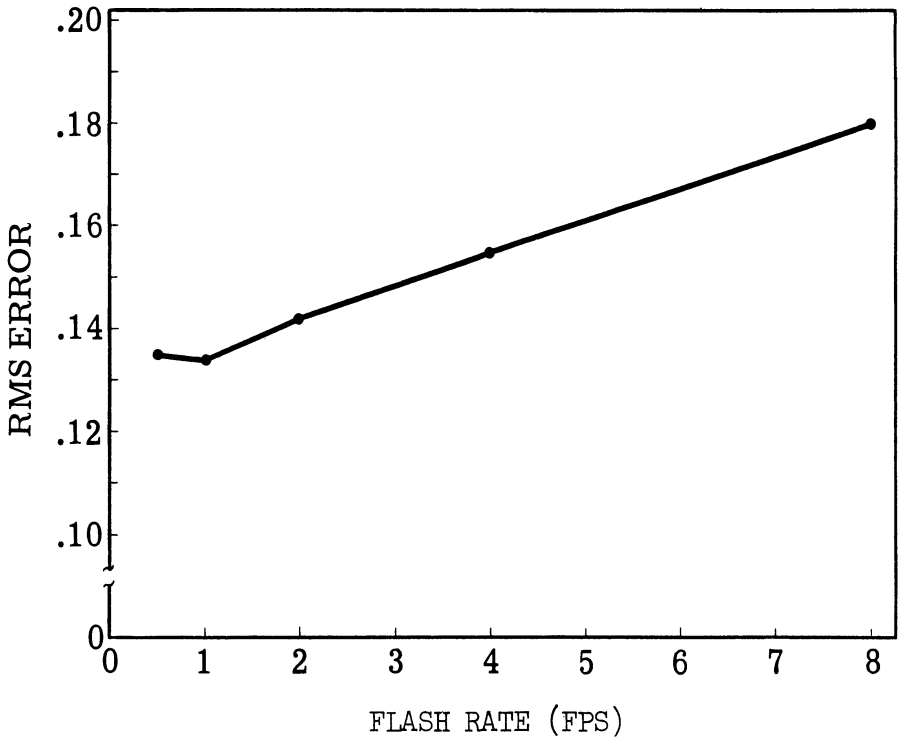


Figure 12. Root Mean Square Error Over the Whole Sub-Problem as a Function of Flash Rate.

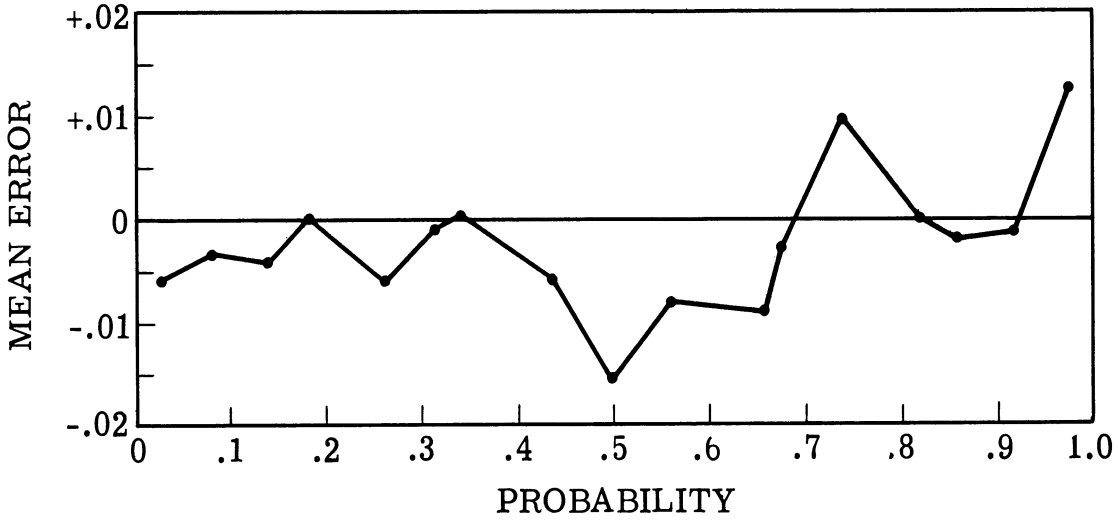


Figure 13. Mean Error as a Function of Probability. This measure is made from convergence to the end of the sub-problem.

Root Mean Square Error After Convergence, RSME, C. RMSE, C is shown in Figures 14 through 16. The only independent variable not affecting RMSE, C is step size. This indicates that the period after the point of convergence is not affected by step size. The constraint on the generation of the flashes reduced the RMSE, C by about 0.014 and does not appear to interact with either step size or rate. RMSE, C decreases with increasing rate from 0.5 to 2 fps and thereafter remains relatively constant.[3] Considered together with the data indicating smaller detection values at the lower rates it is high probable that the number of decisions concerning changes in the probability on a per flash basis, is highest at the lowest rate. Thus the additional decision time available at the lower rates permitted smaller detection values but resulted in larger RMSE, C when the probability was constant.

The effect of probability on the RMSE, C is shown in Figure 16. The "random" problems are consistently higher than the "constrained" problems at all probabilities. The N = 17.3 line is the RMS error, or standard deviation, of a 17.3 flash average. The subject's response is about this good or better at all probabilities.

False Alarm Rate, FAR. The number of false alarms per flash is shown in Figures 17 through 19. Its behavior is similar to RMSE, C. It is similarly insensitive to the size of the step change. Increasing rate causes a decrease in FAR up to 4 fps with an apparent leveling off above 4 fps. These data lend additional support to the hypothesis concerning an increase in

---

[3]Inter-subject variation was high for 4 and 8 fps. See Figure 27, Appendix H.

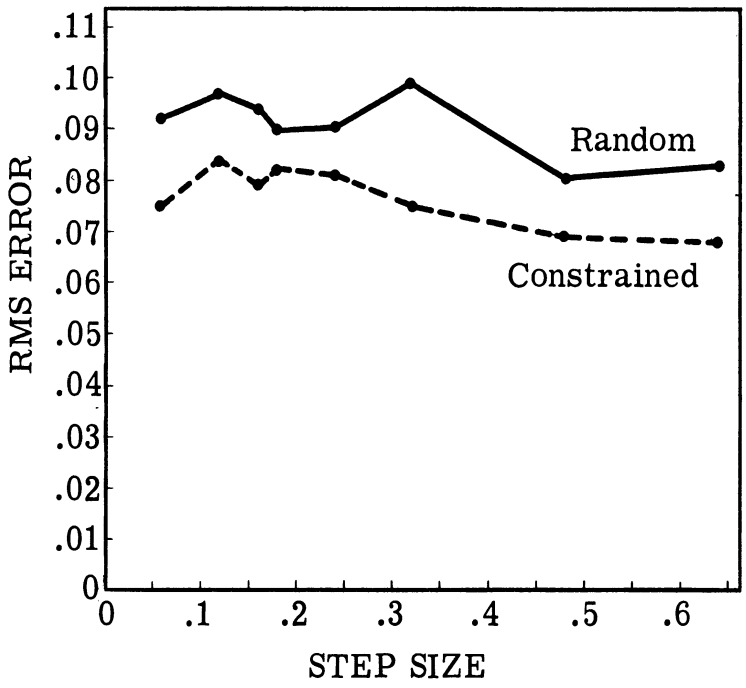


Figure 14. Root Mean Square Error as a Function of Step Size and Constraint. This measure is made from convergence to the end of the sub-problem.

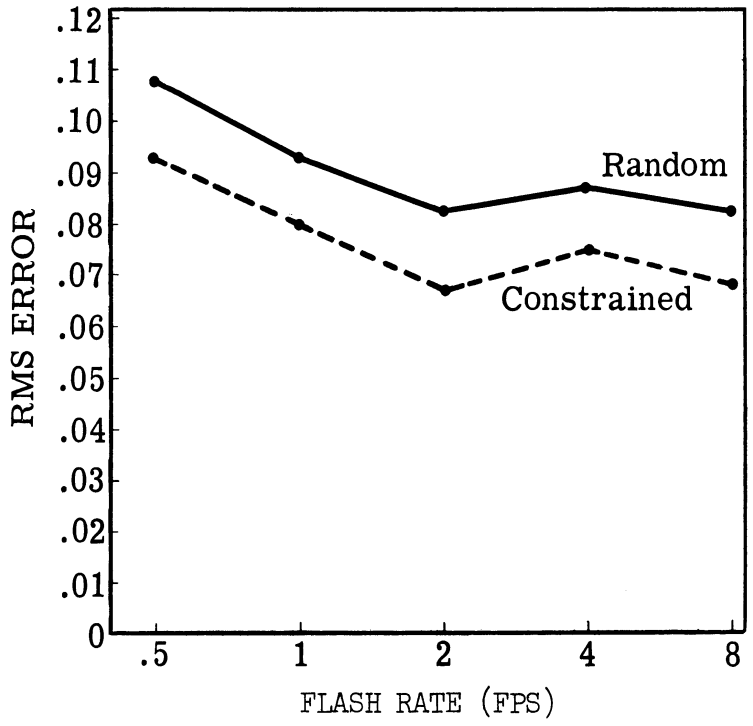


Figure 15. Root Mean Square Error as a Function of Flash Rate and Constraint. This measure is made from convergence to the end of the sub-problem.

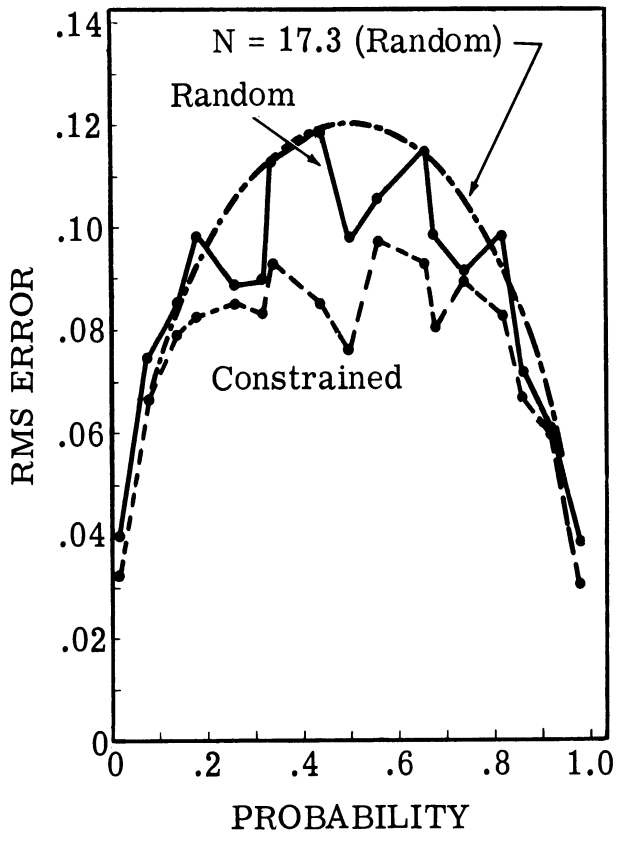


Figure 16. Root Mean Square Error as a Function of Probability and Constraint. This measure is made from convergence to the end of the sub-problem. The standard deviation for a 17.3 sample mean is shown for the random problem.

number of decisions per flash at the lower rates. False alarms can be considered as indicating decisive changes in the estimate.

FAR remains constant over the entire probability range with the exception of the extreme values: 0.02 and 0.98. These probabilities were usually estimated as 0 or 1 with an excursion away from 0 or 1 only following an occurrence of the infrequent flash. Since FAR did not change with probability it appears that the rate of "decisive" movements (greater than 0.05 from the probability) remained constant for all probabilities. The reduction in RMSE, C as the probability tends to extreme values therefore indicates that the time spent at these "erroneous" estimates decreased with extreme probabilities. This hypothesis is supported by observations made during the tracking sessions. The lever movements appeared larger although less frequent at the more extreme probabilities. The increase in magnitude evidently compensated for the decrease in frequency to maintain the FAR at a constant level.

The constrained series produced a slightly higher false alarm rate than the random series. The constrained series has a greater number of runs of right or left flashes and would be expected to yield a higher decision rate. All of the FAR data will be dependent on the false alarm criterion level. A larger criterion could well reverse the constraint finding, for example, since the random series probably produces larger decision movements than the constrained series.

Summary of Results

The response to a step change in probability can be described in three regions: The period before any response to the change, before the

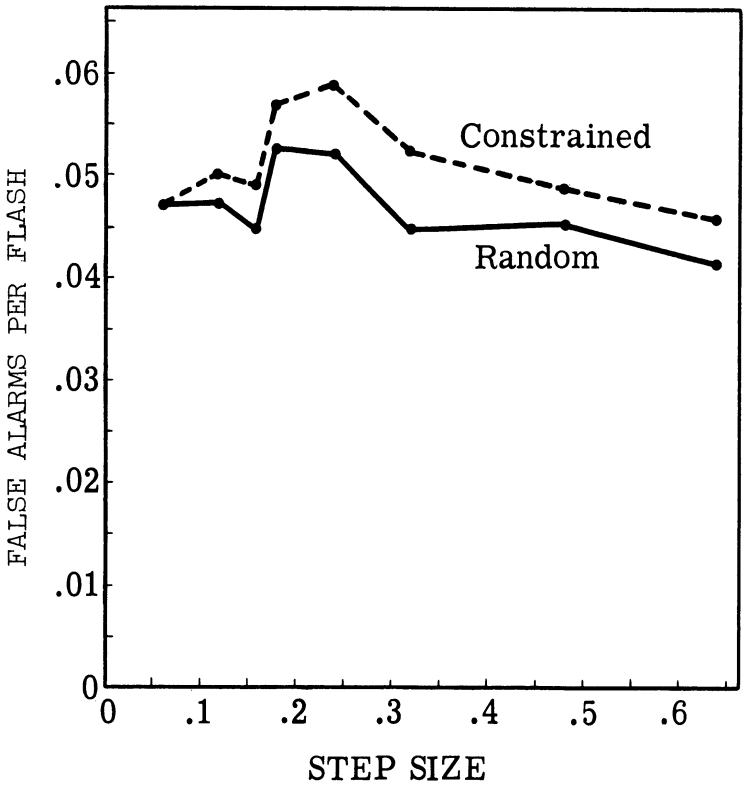


Figure 17. False Alarm Rate in False Alarms Per Flash as a Function of Step Size and Constraint. This measure is made from convergence to the end of the sub-problem.

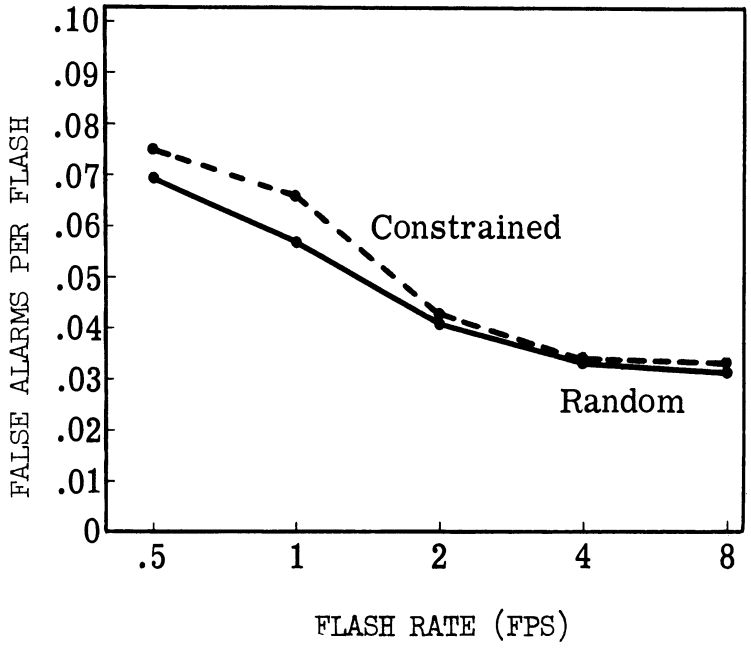


Figure 18. False Alarm Rate in False Alarms Per Flash as a Function of Flash Rate and Constraint. This measure is made from convergence to the end of the sub-problem.

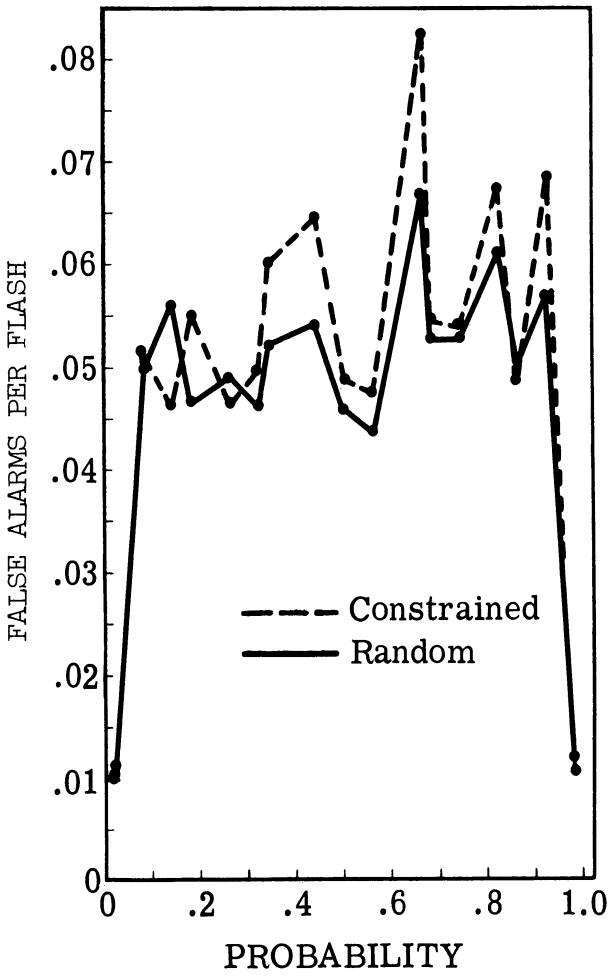


Figure 19. False Alarm Rate in False Alarms Per Flash as a Function of Probability and Constraint. The measure is made from convergence to the end of the sub-problem.

point of detection; the period before the convergence on a new estimate; and the period from the convergence point to the end of the sub-problem. These regions were defined mathematically as functions of probability response form, and somewhat arbitrary constants in order to achieve a complete description of the response.

Detection increases with increasing rate and decreases with increasing step size. The range was from 4 to 24 flashes for a rate range of 0.5 to 8 flashes per second and a step size range of 0.06 to 0.64. Detection was approximately nine flashes for a step of 0.32 at 1 fps.

Convergence increases with both rate and step size. The range was from 11 to 35 flashes for the same step and rate ranges stated above. Convergence was approximately 15 flashes for a step of 0.32 at 1 fps.

Both detection and convergence were independent of the constraint imposed on the generation of the flash series. Both were independent of probability.

After the point of convergence the average estimate was unbiased at all probabilities. This unbiased estimate had an RMS error, or standard deviation, of about 0.08.

The overall task performance was measured by the RMS error over the whole sub-problem. RMSE increased linearly with rate from 0.135 at 1 fps to 0.180 at 8 fps.

IV. MATHEMATICAL MODELS

Three mathematical models will be derived in this chapter. Two of these will be called normative in that the purpose for their derivation is to provide standards by which to compare the data presented in Chapter III. The third is a descriptive model designed to simulate the human performance.

The normative models are somewhat arbitrary in their form. They are optimized within this form, however, to provide a best RMS error fit to the various inputs used in the probability tracking task. One form selected is a constant weighted average over a finite number of past flashes. The simplicity of this model makes it ideal for intuitive comparisons with the subjects' performance. The number of flashes in the running average is selected to give the best fit. The other model has geometrically decreasing weight for flashes extending into the past. This model is more appealing from the standpoint of response to the step inputs. It also corresponds to assumptions often made concerning the human immediate memory function. The best fit is found by selecting the appropriate geometric ratio.

More sophisticated linear, and certainly some non-linear, models would undoubtedly perform this task with a lower RMS error than the two models selected. The value of more complex models for providing simple standards is marginal, however.

The descriptive model was derived from thoughts on how the task was performed by the subjects. Its form arises from the qualitative aspects of the data and from observations of the subjects' behavior.

It has four parameters which are adjusted to yield a minimum RMS error fit to a subject's response.

The normative models to be considered have the form

$$r(n) = \sum_{i=1}^{i=N} w_i \, s_{n-i+1} \tag{4.1}$$

where $r(n)$ is the models response or output at the point n in the sample series and w_i is a weight attached to the sample s_{n-i+1}. The response at n is thus the weighted average of the sample at n and its $N-1$ immediate predecessors. This is an averaging or smoothing model intuitively appropriate to this task. It is limited to samples at and prior to the response point, considering only a finite number of these, and is therefore physically realizable. w_i is not a function of n and could be described as sample-invariant.

The random variables s_n are drawn from an infinite population and are independent. They have values 0 or 1 corresponding respectively to left and right on the subject's display. The probability of a 1 is P and the probability of a 0 is therefore $1-P$.

In the situation where the N samples are all generated from a static distribution described by the probability P it is desirable that the estimate be unbiased or that

$$\overline{r(n)} = P \tag{4.2}$$

where $\overline{r(n)}$ is the expected value of $r(n)$, an ensemble average. This simply requires that

$$\sum_{i=1}^{i=N} w_i = 1 \tag{4.3}$$

The responses of the two model forms will be derived for a sub-problem beginning with $n = 1$ as the first sample of the new probability and ending with $n = M$. The previous probability will be P_1 and the sub-problem probability P_2. The step change is therefore $P_2 - P_1$. For $N < n \leq M$ the response will be called steady-state since the samples are all from a static distribution. For $1 \leq n \leq N$ the response will be called transient.

A Model With Geometric Weighting

The first model to be considered is one having a weighting function

$$w_i = ar^{i-1} \tag{4.4}$$

where a and r are constants and $0 < r < 1$. This function assigns geometrically decreasing weights to the samples. Limiting r to the range 0 to 1, exclusive, confines the function to one assigning monotonically decreasing weights to samples receding from n.

The value of N, the number of samples included in one computation, will be selected as a number large enough to assure the relative unimportance of the weight at $n = N$, ar^{N-1}, compared to the weight at $n = 1$, a. This merely implies that the function's memory extends smoothly to the point of essentially complete "forgetting". The exact value of N in any particular model of this form is relatively unimportant to the considerations that follow. It will simply be assumed that

$$r^N \ll 1 \tag{4.5}$$

and all quantities of this magnitude will be dropped.

Of primary interest is the selection of r to produce an optimum model in the least mean squared error sense. This particular measure of performance was the same one used as a measure of the subject's performance and used in the payoff scheme.

The constant a is selected to satisfy Equation (4.3) which becomes

$$\sum_{i=1}^{i=N} ar^{i-1} = a + ar + ar^2 + \ldots + ar^{N-1} = a\,\frac{1-r^N}{1-r} = 1 \qquad (4.6)$$

or

$$a = 1-r, \quad r^N \ll 1 \qquad (4.7)$$

We will be concerned with two quantities; $\overline{r(n)}$, the expected value of $r(n)$ at the point n, and $\sigma_r^2(n)$, the variance of $r(n)$ at the point n. These are ensemble averages.

For the expected value of r we have

$$\overline{r(n)} = E\Big[\sum_{i=1}^{i=N} w_i\, s_{n-i+1}\Big] \qquad (4.8)$$

and since the w_i are constant over the ensemble

$$\overline{r(n)} = \sum_{i=1}^{i=N} w_i\, E(s_{n-i+1}) = \sum_{i=1}^{i=N} w_i \qquad (4.9)$$

For the step function input $\overline{r(n)}$ will depend on P_1 and P_2 during the transient phase and on P_2 alone during the steady-state phase.

For $1 \leq n \leq N$ we have

$$\begin{aligned} r(n) &= \frac{(1-r)(1-r^n)}{1-r}\,P_2 + \Big[1 - \frac{(1-r)(1-r^n)}{1-r}\Big]\,P_1 \\ &= (1-r^n)\,P_2 + r^n P_1 \\ &= P_2 - r^n(P_2 - P_1) \end{aligned} \qquad (4.10)$$

For $N < n \leq M$

$$\overline{r(n)} = P_2 \tag{4.11}$$

For the variance we have the variance of the sum of the $w_i s_{n-i+1}$ terms. Since the s_n are independent we have the sum of the variances of the individual terms

$$\sigma_r^2(n) = a^2\sigma_s(n) + a^2r^2\sigma_s(n-1) + a^2r^4\sigma_s^2(n-2) + \ldots + a^2r^{2(N-1)}\sigma_s^2(n-N+1) \tag{4.12}$$

where $\sigma_s^2(n)$ is the variance of the sample s_n. Since $\sigma_s^2(n) = P(1-P)$ where P is the probability with which s_n was generated, $\sigma_s^2(n)$ will be a constant for constant P and in particular it will have two values during the transient phase, $\sigma_1^2 = P_1(1 - P_1)$ and $\sigma_2^2 = P_2(1 - P_2)$.

For the transient phase we then have

$$\sigma_r^2(n) = (1-r)^2 \frac{(1-r^{2n})}{(1-r^2)} \sigma_2^2 + [(1-r)^2 \frac{(1-r^{2N})}{(1-r^2)} - (1-r)^2 \frac{(1-r^{2n})}{(1-r^2)}] \sigma_1^2$$

$$= \frac{(1-r)}{(1+r)} [\sigma_2^2 - r^{2n}(\sigma_2^2 - \sigma_1^2)], \quad r^N \ll 1 \tag{4.13}$$

and for the steady-state phase

$$\sigma_r^2(n) = \frac{(1-r)}{(1+r)} \sigma_2^2 \, . \tag{4.14}$$

We can now proceed with the formulation of the models performance in terms of its mean squared error. During the transient phase the error can be written as

$$e(n) = r(n) - P_2$$

$$= [\overline{r(n)} - P_2] + [\overline{r(n)} - r(n)] \tag{4.15}$$

and the squared error is then

$$e^2(n) = [\overline{r(n)}-P_2]^2 + [r(n)-\overline{r(n)}]^2 + 2[\overline{r(n)}-P_2][r(n)-\overline{r(n)}]. \quad (4.16)$$

The expected value of the squared error is then

$$\overline{e^2(n)} = [\overline{r(n)} - P_2]^2 + \sigma_r^2(n) \quad (4.17)$$

since $[\overline{r(n)} - P_2]$ is a constant for a particular n and $E[r(n) - \overline{r(n)}]$ is 0.

The average value of this mean squared error over the transient phase of the sub-problem will be $\langle\overline{e^2}\rangle_T$, representing the average over the ensemble and also over samples in the sub-problem. We will then have

$$\langle\overline{e^2}\rangle_T = \frac{1}{N}\sum_{n=1}^{n=N}\overline{e^2}(n) = \frac{1}{N}\sum_{n=1}^{n=N}[\overline{r(n)}-P_2]^2 + \frac{1}{N}\sum_{n=1}^{n=N}\sigma_r^2(n) \; . \quad (4.18)$$

Using Equation (4.10) the first term on the right side of Equation (4.18) becomes

$$\frac{1}{N}\sum_{n=1}^{n=N}[\overline{r(n)} - P_2]^2 = \frac{1}{N}\sum_{n=1}^{n=N}[r^n(P_1 - P_2)]^2$$

$$= \frac{(P_1 - P_2)^2}{N}\;\frac{r^2}{1 - r^2}, \quad r^N \ll 1 \quad (4.19)$$

Using Equation (4.13) the second term on the right side of Equation (4.18) becomes

$$\frac{1}{N}\sum_{n=1}^{n=N}\sigma_r^2(n) = \frac{1}{N}\sum_{n=1}^{n=N}\frac{(1-r)}{(1+r)}[\sigma_2^2 - r^{2n}(\sigma_2^2 - \sigma_1^2)]$$

$$= \frac{(1-r)}{N(1+r)}[N\sigma_2^2 - \frac{r^2}{1-r^2}(\sigma_2^2 - \sigma_1^2)] \quad (4.20)$$

$$= \frac{(1-r)}{(1+r)}[\sigma_2^2 - \frac{r^2}{N(1-r^2)}(\sigma_2^2 - \sigma_1^2)], \quad r^N \ll 1$$

The average, mean squared error during the transient phase, $\overline{\langle e^2 \rangle}_T$ is then

$$\overline{\langle e^2 \rangle}_T = \frac{(P_1-P_2)}{N} \frac{r^2}{1-r^2} + \frac{(1-r)}{(1+r)} \left[\sigma_2^2 - \frac{r^2}{N(1-r^2)}(\sigma_2^2 - \sigma_1^2)\right] . \qquad (4.21)$$

The average, mean squared error during the steady-state phase, $\overline{\langle e^2 \rangle}_{SS}$, is simply the variance, $\sigma_r^2(n)$, as given by Equation (4.14)

$$\overline{\langle e^2 \rangle}_{SS} = \frac{(1-r)}{(1+r)} \sigma_2^2 . \qquad (4.22)$$

The average, mean squared error over the whole sub-problem is then

$$\overline{\langle e^2 \rangle}_{SP} = \frac{N}{M} \overline{\langle e^2 \rangle}_T + \frac{M-N}{M} \overline{\langle e^2 \rangle}_{SS}$$

$$= \frac{(P_1-P_2)^2}{M} \frac{r^2}{1-r^2} + \frac{N}{M} \frac{(1-r)}{(1+r)} \left[\sigma_2^2 - \frac{r^2}{N(1-r^2)}(\sigma_2^2 - \sigma_1^2)\right]$$

$$+ \left(\frac{M-N}{M} \frac{(1-r)}{(1+r)} \sigma_2^2\right.$$

$$= \frac{(P_1-P_2)^2}{M} \frac{r^2}{(1-r^2)} + \frac{(1-r)}{(1+r)} \left[\sigma_2^2 - \frac{r^2}{M(1-r^2)}(\sigma_2^2 - \sigma_1^2)\right] . \qquad (4.23)$$

We are interested in the performance of this model over the same types of problems given to the subjects. The average, mean squared error over a problem is given by

$$\overline{\langle e^2 \rangle}_P = \sum_{i=1}^{i=S} \frac{M_i}{T} \left\{\frac{(P_{i-1} - P_i)^2}{M_i} \frac{r^2}{1-r^2} + \frac{1-r}{1+r}\left[\sigma_i^2 - \frac{r^2}{M_i(1-r^2)}(\sigma_i^2 - \sigma_{i-1}^2)\right]\right\} \qquad (4.24)$$

where M_i is the length of sub-problem i, T is the total problem length in samples, and S is the total number of sub-problems.

This expression may be simplified by making the following assumptions based on the methods used for generating the problems. (Appendix B.) The M_i were selected randomly, without replacement, from a set of equally frequent values and assigned to the sub-problems. The sum of $(\sigma_i^2 - \sigma_{i-1}^2)/M_i$ therefore approaches zero for long series of sub-problems. Similarly the term $M_i\ \sigma_i^2/T$ will approach simply σ_i^2/S. Equation (4.24) can therefore be written as

$$\overline{\langle e^2 \rangle}_P = \frac{r^2}{1-r^2} \frac{1}{T} \sum_{i=1}^{i=S} (P_{i-1} - P_i)^2 + \frac{1-r}{1+r} \frac{1}{S} \sum_{i=1}^{i=S} \sigma_i^2 \qquad (4.25)$$

The large step problem had values of $|P_1 - P_2|$ of 0.16, 0.32, 0.48, and 0.64 occurring in 12, 10, 8, and 6 sub-problems respectively, yielding

$$\frac{1}{T} \sum_{i=1}^{i=S} (P_{i-1} - P_i)^2 = 0.00251 \ .$$

σ_i^2 had values of 0.250, 0.224, 0.148, and 0.020 occurring in 6, 12, 10, and 8 sub-problems respectively, yielding

$$\frac{1}{S} \sum_{i=1}^{i=S} \sigma_i^2 = 0.162$$

T was 2241 samples and S was 36 sub-problems. The average, mean squared error over the large step problem is then

$$\overline{\langle e^2 \rangle}_{LSP} = 0.00251 \frac{r^2}{1-r^2} + 0.162 \frac{1-r}{1+r} \qquad (4.26)$$

The small step problem had values of $|P_1 - P_2|$ of 0.06, 0.12, 0.18, and 0.24 occurring in 12, 10, 16, and 12 sub-problems respectively, yielding

$$\frac{1}{T} \sum_{i=1}^{i=S} (P_{i-1} - P_i)^2 = 0.000465 \ .$$

σ_1^2 had values of 0.250, 0.246, 0.217, 0.192, 0.120, and 0.085 occurring in 6, 10, 12, 10, 6, and 6 sub-problems respectively, yielding

$$\frac{1}{S}\sum_{i=1}^{i=S}\sigma_i^2 = 0.194 \quad .$$

T was 3000 samples and S was 50 sub-problems. The average, mean squared error over the small step problem is then

$$\overline{\langle e^2\rangle}_{SSP} = 0.000465\,\frac{r^2}{1-r^2} + 0.194\,\frac{1-r}{1+r} \tag{4.27}$$

It will be of interest for comparative purposes to evaluate this model for the case in which only one value of r is used for both the large and small step problems. This model will be called non-discriminating in Chapter V. In this case the sums in Equation (4.25) are over both problem types with T being equal to 5241 samples and S being 86 sub-problems. We have

$$\frac{1}{T}\sum_{i=1}^{i=S}\left(P_{i-1} - P_i\right)^2 = 0.00134$$

and

$$\frac{1}{S}\sum_{i=1}^{i=S}\sigma_i^2 = 0.181 \quad .$$

The average, mean squared error over the large plus the small step problems, is then

$$\overline{\langle e^2\rangle}_{S+L} = 0.00134\,\frac{r^2}{1-r^2} + 0.181\,\frac{1-r}{1+r} \tag{4.28}$$

We are interested in the selection of an optimum value of r for these three problem types. Equation (4.25) can be written as

$$\overline{\langle e^2\rangle}_P = k\,\frac{r^2}{1-r^2} + v\,\frac{1-r}{1+r} \tag{4.29}$$

where k and v are the constants for the specific problem type. The minimum of this function over r can then be found by setting

$$\frac{d\,\overline{\langle e^2\rangle_p}}{dr} = \frac{-2vr^2 + (2k + 4v)r - 2v}{(1 - r^2)^2} = 0 \qquad (4.30)$$

which yields two roots

$$r_{1,2} = \frac{k + 2v}{2v} \pm \left[\left(\frac{k + 2v}{2v}\right)^2 - 1\right]^{1/2} \qquad (4.31)$$

The minus sign yields a value of r between 0 and 1 and is also the minimum.

For the large step problem Equation (4.31) gives an optimum $r = 0.883$. Using this value of r in Equation (4.26) we have a corresponding minimum mean squared error or 0.0190.

For the small step problem Equation (4.31) gives an optimum $r = 0.953$. Using this value of r in Equation (4.27) we have a corresponding minimum mean squared error of 0.00931.

For the large plus small step problems Equation (4.31) gives an optimum $r = 0.915$. Using this value of r in Equation (4.28) we have a corresponding minimum mean squared error of 0.0147.

A Model with Constant Weighting

The second model to be considered is a model giving a constant weight to each of N samples, a simple averaging model. The derivation of the response and errors for this model will parallel that for the geometric model and some of the detailed explanations will be omitted.

The weighting function is

$$w_i = 1/N \qquad (4.32)$$

where N is the number of samples in the average and the weight is $1/N$ to satisfy Equation (4.3).

In this case the transient response will be

$$\overline{r(n)} = \frac{n}{N} P_2 + \frac{N-n}{N} P_1 = P_1 + \frac{P_2 - P_1}{N} n \tag{4.33}$$

and the steady-state response

$$\overline{r(n)} = P_2 \quad . \tag{4.34}$$

The variance of $r(n)$ during the transient phase will be the variance of the sum of N terms each with weight $1/N$

$$\sigma_r^2(n) = \frac{N-n}{N^2} \sigma_1^2 + \frac{n}{N^2} \sigma_2^2$$

$$= \frac{1}{N} [\sigma_1^2 + \frac{n}{N} (\sigma_2^2 - \sigma_1^2)] \quad . \tag{4.35}$$

The variance of $r(n)$ during the steady-state phase will simply be

$$\sigma_r^2(n) = \frac{\sigma_2^2}{N} \tag{4.36}$$

Following the same procedures and arguments developed in the derivation of Equations (4.15) through (4.19) we have

$$\frac{1}{N} \sum_{n=1}^{n=N} [\overline{r(n)} - P_2]^2 = \frac{1}{N} \sum_{n=1}^{n=N} [P_1 + \frac{P_2 - P_1}{N} n - P_2]^2$$

$$= \frac{(P_2 - P_1)^2}{N} \sum_{n=1}^{n=N} (\frac{n}{N} - 1)^2$$

$$= (P_2 - P_1)^2 (\frac{1}{3} - \frac{1}{2N} + \frac{1}{6N^2}) \tag{4.37}$$

and analogous to Equation (4.20) we have

$$\frac{1}{N}\sum_{n=1}^{n=N}\sigma_r^2(n) = \frac{1}{N^2}\sum_{n=1}^{n=N}\left[\sigma_1^2 + \frac{n}{N}(\sigma_2^2 - \sigma_1^2)\right]$$

$$= \frac{1}{N}\sigma_1^2 + \frac{N+1}{2N^2}(\sigma_2^2 - \sigma_1^2) \quad . \tag{4.38}$$

The average, mean squared error during the transient phase is then

$$\overline{\langle e^2\rangle}_T = (P_2 - P_1)^2\left(\frac{1}{3} - \frac{1}{2N} + \frac{1}{6N^2}\right) + \frac{1}{N}\sigma_1^2 + \frac{N+1}{2N^2}(\sigma_2^2 - \sigma_1^2) \quad . \tag{4.39}$$

The average, mean squared error during the steady-state phase is the variance given by Equation (4.36).

$$\overline{\langle e^2\rangle}_{SS} = \frac{\sigma_2^2}{N} \quad . \tag{4.40}$$

The average, mean squared error over the whole sub-problem is then

$$\overline{\langle e^2\rangle}_{SP} = \frac{N}{M}\overline{\langle e^2\rangle}_T + \frac{M-N}{M}\overline{\langle e^2\rangle}_{SS}$$

$$= \frac{N}{M}(P_1 - P_2)^2\left(\frac{1}{3} - \frac{1}{2N} + \frac{1}{6N^2}\right) + \frac{1-N}{2MN}(\sigma_2^2 - \sigma_1^2) + \frac{\sigma_2^2}{N} \tag{4.41}$$

As with the geometric weighted model we are interested in the performance of this model over the problems given to the subjects. The average, mean squared error over a problem is given by

$$\overline{\langle e^2\rangle}_P = \sum_{i=1}^{i=S}\frac{M_i}{T}\left\{\frac{(P_{i-1} - P_i)^2}{M_i}\left(\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}\right) + \frac{1-N}{2M_iN}(\sigma_i^2 - \sigma_{i-1}^2) + \frac{\sigma_i^2}{N}\right\}$$

(4.42)

With the same arguments leading to Equation (4.25) we have

$$\overline{\langle e^2\rangle}_P = \frac{1}{T}\left(\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}\right)\sum_{i=1}^{i=S}(P_{i-1} - P_i)^2 + \frac{1}{NS}\sum_{i=1}^{i=S}\sigma_i^2 \tag{4.43}$$

These sums are the same as those calculated for the geometric weighted model. For the large step problem we have

$$\langle\overline{e^2}\rangle_{LSP} = 0.00251\ (\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}) + 0.162\ \frac{1}{N} \tag{4.44}$$

For the small step problem

$$\langle\overline{e^2}\rangle_{SSP} = 0.000465\ (\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}) + 0.194\ \frac{1}{N} \tag{4.45}$$

And for the large plus small step problems

$$\langle\overline{e^2}\rangle_{L+S} = 0.00134\ (\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}) + 0.181\ \frac{1}{N} \tag{4.46}$$

The minima can again be selected by letting k and v be the constants for the particular problem type.

$$\langle\overline{e^2}\rangle_{P} = k\ (\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}) + \frac{v}{N} \tag{4.47}$$

and solving

$$\frac{d\ \langle\overline{e^2}\rangle_P}{dN} = \frac{k}{3} - \frac{k}{6N^2} - \frac{v}{N^2} = 0 \tag{4.48}$$

yielding

$$N = [\frac{1}{2} + \frac{3v}{k}]^{1/2} \tag{4.49}$$

For the large step problem Equation (4.49) gives an optimum $N = 14.1$. Using this value of N in Equation (4.44) we have a corresponding minimum, mean squared error of 0.0220.

For the small step problem Equation (4.49) gives an optimum $N = 35.3$. Using this value of N in Equation (4.45) we have a corresponding minimum, mean squared error of 0.0107.

For the large plus small step problems Equation (4.49) gives an optimum $N = 20.1$. Using this value of N in Equation (4.46) we have a corresponding minimum, mean squared error or 0.0172.

A Descriptive Model

Upon inspection of the subjects' response data it is evident that the estimation task was not performed in the smooth manner of the two normative models. The responses were characterized by rapid adjustments separated by periods of little or no movement. This evidence, together with thoughts on how this task might be performed, led to the postulation of the following model as an attempt to describe the human performance.

This model operates in the following manner: The subject maintains a short, running average of the previous k_3 flashes. This average is of exactly the same type as the second normative model discussed above. At each flash this average is compared with the existing setting of the response lever and the difference noted. If this distance measure is greater than a prescribed criterion level, the response is changed to a new value at some point intermediate between the old response and the running average. If the difference is less than the criterion level, the response remains unchanged.

This model has several features making it attractive from a descriptive standpoint. It uses the lever as a memory device, moving it only a fraction of the distance to the new average and thus preserving some of the information in the previous setting. This memory function permits a smaller number of flash in the running average than would

otherwise be required to produce the levels of mean squared error measured from the subject's responses. The criterion level corresponds to the concept of the subjects' smallest perceptable difference between the running average and the lever position. It permits the response to remain stationary through periods of small deviation of the running average from the response.

This model's operation can be thought of as a form of hypothesis testing. At each flash it is testing the hypothesis that the running average is from a population described by the response lever setting, using the criterion level as a form of significance measure. The subject's performance is thus viewed as a succession of decision making situations. This framework is appropriate with the inclusion of more higher mental processes than are in the usual manual tracking task.

This model can be described mathematically as follows:

$$u(n) = \frac{1}{k_3} \sum_{i=1}^{i=k_3} s_{n-i+1} \tag{4.50}$$

where $u(n)$ is the running average of k_3 flashes, s_n. With $r(n)$ as the current lever setting if

$$|r(n) - u(n)| \leq k_1 \tag{4.51}$$

where k_1 is the criterion band, then

$$r(n+1) = r(n) \ . \tag{4.52}$$

If, however,

$$|r(n) - u(n)| > k_1 \tag{4.53}$$

then

$$r(n+1) = r(n) + k_2 [u(n) - r(n)] \tag{4.54}$$

where k_2 is the fractional lever adjustment.

A fourth parameter, k_4, was also considered, representing a time (flash) shift between the subject's and the model's responses. The subject's response at n was compared to the model's at $n - k_4$.

The four parameters are constrained to the following ranges:

$$0 < k_1 < 1 \tag{4.55}$$

where 0 yields adjustment decisions at each sample and 1 would yield no adjustment decisions.

$$0 < k_2 \leq 1 \tag{4.56}$$

where 0 yields no response changes and 1 represents simply the following of the running average whenever an adjustment decision is made:

$$1 \leq k_3 \leq K, \quad k_3 \text{ an integer} \tag{4.57}$$

where K is some reasonable maximum number of flashes that the subject could be expected to assimilate in one averaging calculation. No definite values for K are known for this task. It is certainly reasonable to assume that the flashes are not simply remembered as a succession of binary symbols but are encoded into a larger symbol set, perhaps one depending on the lengths of runs of one of the binary symbols. Considering the nature of the task and its difficulty it would seem unreasonable that more than 20 flashes could be used in an averaging calculation and that a value closer to 10 would be more appropriate.

There are no particular constraints on k_4 except that $k_4 < 0$ implies subject prediction with respect to the model.

Having thus chosen the model form the task is now to select parameter sets (k_1, k_2, k_3, k_4) which will make the model best describe the human performance. The criterion used for this selection was the minimization of the mean squared error between the subject's and the model's responses over particular problems. This measure was selected as providing the best single measure of performance, as was discussed in Chapter II. The selection of the minimum mean squared error for the criterion assures a fairly close fit to the transient portion of the response where the error is large, at the possible expense of fit to the steady-state portion.

The actual minimization process was carried out as follows. The model was programmed on an IBM 709 computer. The computer was then fed four of the input problems used in the experiment plus the responses of one of the four subjects to these problems. At each sample point the squared difference between the subject's and the model's responses was calculated and accumulated. These values were simply printed out at the end of each particular problem-parameter set combination. The large number of parameter sets possible and the possibility of numerous minima precluded the possibility of an automatic searching technique for the minima. Several computer runs were made in which previously selected parameter ranges were either extended or filled in according to the results of the previous run. The total variation of the parameters was through the following ranges.

$$0.02 \leq k_1 \leq 0.20 \quad \text{(6 values)}$$

$$0.10 \leq k_2 \leq 0.90 \quad \text{(7 values)}$$

$$1 \leq k_3 \leq 28 \quad \text{(12 values)}$$

$$-2 \leq k_4 \leq 4 \quad \text{(6 values)}$$

The four problems investigated were the large and small step problems, random constraint, at 1 and 4 fps. The subject was S-2.

Several parameter sets with approximately equal minimum error measures were found for each problem type. In each case these minima represented either a valley in the error function or fairly distinct minima separated by regions of higher error. Table I shows the various parameter sets and their corresponding minimum errors, $\langle e_{MS}^2 \rangle_P$. In each group of parameter sets various trade offs can be seen among the parameters yielding the approximately equal error measures.

The following method was devised as a means for selecting the best descriptive model from among these parameter sets with approximately equal $\langle e_{MS}^2 \rangle_P$.

The subject's error, $e_S(n)$, can be written as

$$e_S(n) = e_M(n) + e_{MS}(n) \tag{4.58}$$

where $e_M(n)$ is the model's error and $e_{MS}(n)$ is the error between the subject and the model. Squaring this error we have

$$e_S^2(n) = e_M^2(n) + e_{MS}^2(n) + 2e_M(n)\, e_{MS}(n) \quad . \tag{4.59}$$

The average value of this squared error over a particular problem is then

$$\langle e_S^2 \rangle_P = \langle e_M^2 \rangle_P + \langle e_{MS}^2 \rangle_P + \langle 2e_M e_{MS} \rangle_P \tag{4.60}$$

The minimization process used to select the parameter sets was concerned with finding minimum values of $\langle e_{MS}^2 \rangle$. The computer also calculated values for $\langle e_M^2 \rangle_P$, the model's error. $\langle e_S^2 \rangle_P$ was, of course, one of the measures made on the subject's performance. The term $\langle 2e_M e_{MS} \rangle_P$ can therefore be calculated from Equation (4.60).

These error terms can be interpreted in the following manner. Consider the subject's error at any point in the sample sequence to be composed of two components; one dependent in some manner on the actual input samples and the remainder on other phenomena not related to the input. The first part of this error might be termed coherent and the remainder noise. Consider now a descriptive model and its relationship to these two error measures. If it performs the task in exactly the same manner as the subjects, it will have an error, $\langle e_M^2 \rangle_P$, which is equal to the subject's coherent, or sample dependent, error and the error between this model and the subject, $\langle e_{MS}^2 \rangle_P$ would be equal to the subject's noise or sample independent error. The subjects' noise can be considered to be random fluctuations in the response about the coherent value. Since $\langle e_M \rangle_P$ approaches zero over a large set of sub-problems then $\langle e_M e_{MS} \rangle_P$ will also approach zero.

If, on the other hand, the model does not represent the entire coherent part of the subject's response, that is if it is not a complete descriptor of the subject's coherent behavior, then $e_{MS}(n)$ will be partially dependent on the sample series and therefore correlated with $e_M(n)$. In this case the term $\langle 2e_M e_{MS} \rangle_P$ will not approach zero. This correlation can therefore be used as an additional selection device. It

can be written in the normalized form

$$\rho = \frac{\langle e_M e_{MS} \rangle_P}{(\langle e_M^2 \rangle_P \langle e_{MS}^2 \rangle_P)^{1/2}} \tag{4.61}$$

Table I shows the values of $\langle e_M^2 \rangle_P$, $\langle 2 e_M e_{MS} \rangle_P$, and ρ.

The normalized correlation, ρ, provides a measure giving good discrimination among the parameter sets for the large step problems. For the large step problem at 1 fps parameter set 2 has a value of ρ which is essentially zero. At 4 fps parameter set 4 has a very low value for ρ. Neither of the small step problems, however, produce a correlation which discriminates among the parameter sets or which is as small as that found for the large step problem. It would appear on the basis of this evidence that the postulated descriptive model better represents the subject's performance on the large than on the small step problems.

Zero correlation, as defined by Equation (4.61), does not necessarily imply a complete lack of dependence of $e_{MS}(n)$ on the sample series. Two hypotheses could be used to explain the fairly large $\langle e_{MS}^2 \rangle_P$ which remained even for $p \simeq 0$. One would simply be that this level of noise did exist in the subjects' performance. The other would be that this "noise" component had at least some portion which was related to the sample series but which was uncorrelated with $e_M(n)$. Perhaps one reason for a fairly large noise component would be variations in the subjects' method of performing the task during the problem run.

TABLE I

PARAMETER SETS FOR THE DESCRIPTIVE MODEL YIELDING MINIMUM VALUES OF $\langle e^2_{MS} \rangle_P$

	Criterion k_1	Fractional Adjustment k_2	Memory k_3	Lag k_4	$\langle e^2_{MS} \rangle_P$	$\langle e^2_M \rangle_P$	$\langle 2e_M e_{MS} \rangle_P$	ρ
Large Step, 1 fps								
1.	0.05	0.20	8	0	.0104	.0185	.0030	.109
2.	0.12	0.60	12	0	.0121	.0198	.0001	.003
Small Step, 1 fps								
1.	0.05	0.10	6	0	.0076	.0130	-.0066	-.332
2.	0.05	0.20	12	0	.0079	.0140	-.0079	-.375
3.	0.08	0.20	10	0	.0083	.0130	-.0073	-.351
4.	0.12	0.20	8	0	.0086	.0130	-.0076	-.359
5.	0.15	0.40	10	0	.0092	.0120	-.0072	-.343
Large Step, 4 fps								
1.	0.05	0.10	8	2	.0098	.0240	.0062	.209
2.	0.05	0.10	12	0	.0098	.0210	.0092	.321
3.	0.10	0.10	8	2	.0106	.0240	.0054	.170
4.	0.15	0.10	8	2	.0117	.0280	.0003	.008
5.	0.15	0.30	12	2	.0121	.0300	-.0021	.060
Small Step, 4 fps								
1.	0.05	0.10	16	2	.0102	.0140	-.0102	-.426
2.	0.05	0.10	24	0	.0102	.0160	-.0122	-.480
3.	0.10	0.10	16	1	.0102	.0140	-.0102	-.426
4.	0.10	0.30	24	0	.0106	.0140	-.0106	-.418
5.	0.10	0.50	24	1	.0106	.0150	-.0116	-.460

A Normative Variation

The descriptive model discussed above was constructed as an approximation to the human performance on this task. It is interesting, however, to see how well this model form can do if the parameters are selected to give a minimum $\langle e_M^2 \rangle_P$: to be normative in the same sense as the models with geometric and constant weighting. Normative parameter sets for the large and small step problems were found using the computer to calculate $\langle e_M^2 \rangle_P$ and converging on the minimum value by successive selection of the parameter sets as in the selection of the minimum $\langle e_{MS}^2 \rangle_P$ for the descriptive models. k_4 was set equal to zero for this selection.

The large step problem yielded one distinct and interesting minimum; $k_1 = 0.02$, $k_2 = 0.10$, and $k_3 = 1$. All three of these parameters are the smallest values examined and this minimum is in one corner of the error surface. This model would operate as follows: with $k_3 = 1$, the running average would have values of either 0 or 1, depending on the most recent sample, with $k_2 = 0.02$ there would be a response adjustment at every sample except when the response was within 0.02 of either 0 or 1. This adjustment would be 0.10 of the distance between the previous response and 0 or 1. The root mean squared error for this model was 0.124.

The best normative parameter set for the small step problem was found to be: $k_1 = 0.20$, $k_2 = 0.10$, and $k_3 = 6$. Again we have the minimum occurring at the smallest value of k_2 but in this case the criterion for changing the response is fairly high and we have six samples in the memory. The root mean squared error for this model was 0.099.

Both the geometric and constant weighted models are included as special cases of this descriptive model. With $k_3 = 1$ and $k_1 = 0$ the descriptive model is identical with the geometric weighted model with $r = 1 - k_2$. With $k_1 = 0$ and $k_2 = 1$ the model is identical with the constant weighted model with $N = k_3$. The best normative parameter set for the large step problem deviates from the simple geometric form only when the response is within 0.02 of either 0 or 1. The best normative parameter set for the small step problem does not yield as low an error as either the optimum geometric or constant weighted models. The equivalent parameter sets for these models were outside the parameter range investigated, however.

It would seem that this decision model, within its restricted parameter sets, represents a reasonable method for performing this task when the step changes are large, but not when they are small. It is interesting in this light to note that the decision model did not seem to describe the subject's performance on the small step problem as well as on the large step problem.

Figures 20 and 21 show a small representative portion of three responses to the same input sample sequence. The normative model at the top is the parameter set selected above as the best normative set for the decision model. Note the rapid response changes of the large step model. The center response is that of two of the descriptive parameter sets and the lower response, the subject's. The descriptive parameter set for the large step problem is the one with the low value of ρ. The set for the small step problem was selected somewhat arbitrarily as one of the five sets that seemed like a reasonable description. The fairly high coherent subject's error is clearly evident in these figures.

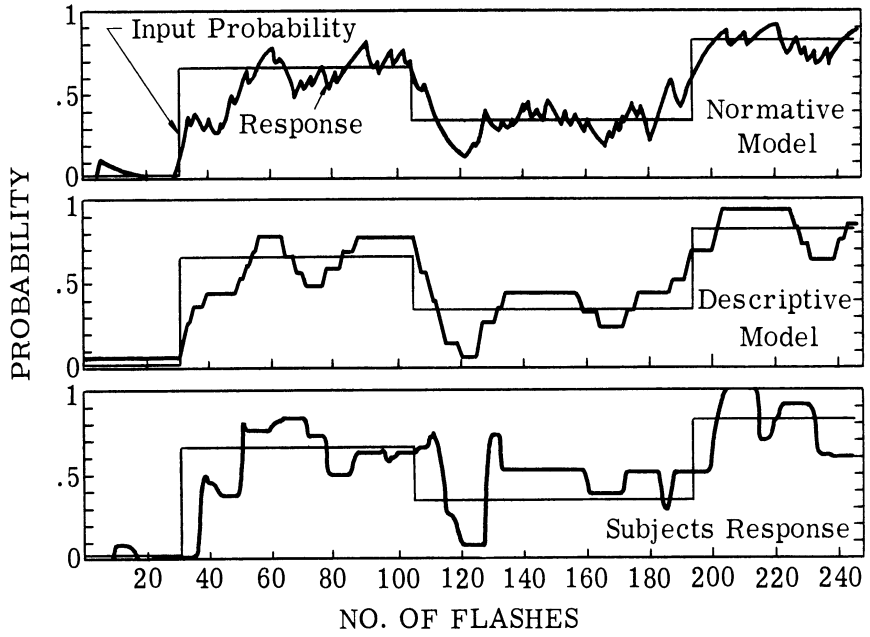


Figure 20. Responses of Two Mathematical Models and a Subject to a Portion of a Large Step Problem, Random Constraint, at 1 FPS. Normative Model $K_1 = 0.02$, $K_2 = 0.10$, $K_3 = 1$; Descriptive Model $K_1 = 0.12$, $K_2 = 0.60$, $K_3 = 12$, $K_4 = 0$.

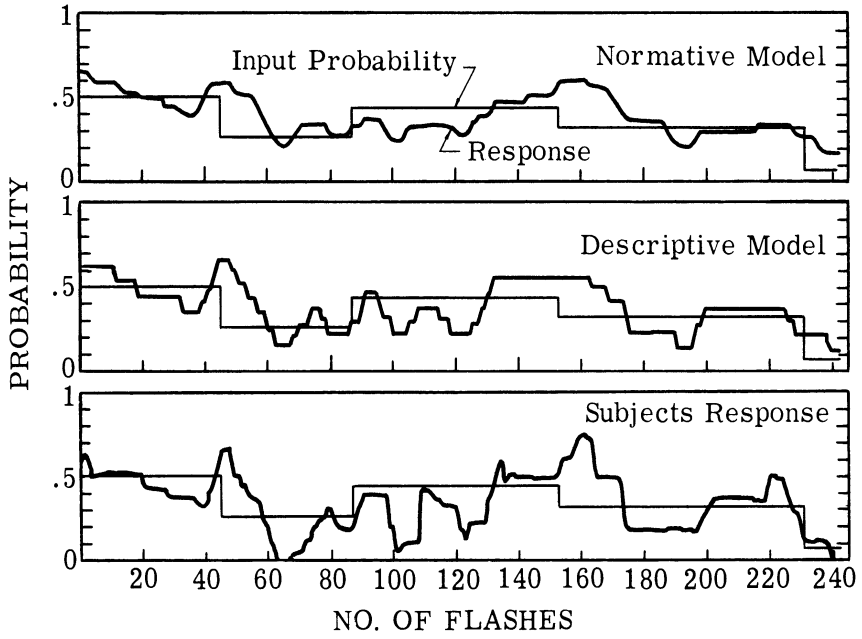


Figure 21. Responses of Two Mathematical Models and a Subject to a Portion of a Small Step Problem, Random Constraint, at 1 FPS. Normative Model $K_1 = 0.20$, $K_2 = 0.10$, $K_3 = 6$; Descriptive Model $K_1 = 0.15$, $K_2 = 0.40$, $K_3 = 10$, $K_4 = 0$.

V. DISCUSSION

The response measures presented in Chapter III do not provide a direct answer to the question of how good is the performance. Quantitative standards are necessary for the measures. Mean error is an exception in that a standard of zero is reasonable and was in fact achieved. The normative models derived in Chapter IV provide the standards for the other measures. They permit a comparison to be made between the subjects' performance and that of several simple machines.

Several important differences exist between the subjects' and the models' knowledge of the task. The subjects were not instructed on the step-function nature of the input. In fact they were specifically told to expect slow, continuous changes in the probability. The models, on the other hand, were optimized for step input functions. It is reasonable to assume, however, that the subjects' original ignorant and misinformed state did not persist for long after the tracking began. The rapid performance asymptote (less than 45 minutes) and the discrimination between the small and large step problems attest to this. The model does not have learning and adaptive abilities, of course, and it was therefore given the maximum knowledge that the subjects could theoretically derive from the task. The model-subject comparison thus includes the subjects' learning and adaptive abilities. This method of subject instruction will allow more valid generalization of the measured estimation ability to other input forms.

The same situation exists in the relative knowledge of the input statistics possessed by the subjects and the models. The models were completely informed of the distributions of step size, step direction,

sub-problem length, and probability. The subjects knew nothing of these initially. Again, however, it can be assumed that the subjects learned a considerable amount about these distributions while performing. The adaptation to the small and large step problems is an example of the subjects distinguishing between two distributions of step size.

The models were provided with a definite criterion for optimum performance; the minimum mean squared error. The subjects were instructed to use the same criterion. The actual criteria used by the subjects, however, corresponds to their conception of best performance and is a function of the instructions, of performing the task, and of personal abilities and sensitivities.

The comparison between the subjects and the models will be made using the following measures: detection, D; convergence, C; root mean squared error, RMSE; and root mean squared error after convergence, RMSE,C. These four measures, plus mean error, provide a fairly complete description of the performance.

Detection and convergence were calculated using Equations 4.10 and 4.33. These measures are for the expected values of the responses and are not the expected values of detection and convergence. The difference is not important of this comparison.

The root mean squared error was calculated using Equation 4.23 and 4.41.

The root mean squared error after convergence was calculated using equations 4.22 and 4.40 with the addition of a correction for the small error contributed by the remaining transient after convergence. This transient error was calculated using Equation 4.19 and 4.37.

Two values of RMSE,C were calculated, one with an infinite sample population as implied in Chapter IV and one with finite populations such as those used in the experiment. The correction factor for the finite populations is derived in Appendix C. It was calculated using an average number of flashes for the sub-problems and corresponds to the random problem type.

Two specific step sizes were selected for the comparison: 0.40 for the large step problem and 0.15 for the small step problem. These are approximately the average step sizes in their respective problems. These step changes are examined at a flash rate of 1 fps and with the random problem type.

Three forms of each normative model are used. Two of these correspond to the optimum models selected for the large and small step problems considered separately. They are called "discriminating" models. The third model is "nondiscriminating" in that it is required to be optimum over both the small and large step problems simultaneously. The parameters for these models were calculated in Chapter IV. The non-discriminating models represent the only case where the models are not provided the complete statistical information.

The models' and subjects' performance measures are shown in Table II. Also included are values of detection and convergence for the descriptive models used in Figures 20 and 21. These measures are averages over a set of sub-problems with average step sizes of approximately 0.40 and 0.15. RMSE and RMSE,C were not available from the descriptive model's data.

The response speed of the nondiscriminating model lies between the rapid response of the large step discriminating models and the smoothing

responses of the small step discriminating models. The discriminating models have lower values of RMSE, of course, since this was the optimization criterion. The nondiscriminating models are between the two discriminating models in RMSE,C, having lower values for the large step problem and higher values for the small step problem.

The subject-model comparison shows a striking difference in the detection values for the large step problem. The normative models have detection values considerably smaller than the subjects. This results, to a large extent, from the difference between the models' smooth response and the decision, criterion testing, nature of the subjects' response hypothesized in Chapter IV. The normative models begin to respond to the step change with the first flash of the new probability. The subjects require a number of flashes to perceive a significant probability change and the necessity of a response change. The large step descriptive model has a comparable detection value with the subjects'.

On the small step problem the subjects' detection values is higher than any of the models' although it is comparable to the discriminating model.

In convergence, however, the subjects performed comparably to the models. On the large step problem only the discriminating, constant model has a smaller value. On the small step problem the subjects' convergence value lies between those of the discriminating and nondiscriminating models.

The hypothesis that the subjects were adapting to the difference between the small and large step problems receives support from the convergence comparisons. The nondiscriminating models show a considerable decrease in convergence from the large to the small step problems. The

discriminating models show an increase in convergence fom the large to the small step problem as they change to a smoother response form. The subjects showed a similar slight increase in convergence from the large to the small step problems.

The subjects' delayed detection with comparable convergence illustrates the discontinuous nature of their behavior. Although unable, or unwilling, to indicate the presence of a change in the probability for the first seven to twelve flashes, they were then able, however, to converge on the new probability in five to seven more flashes.

The subjects were slightly higher than the models in RMSE. This is to a considerable extent the result of the subjects' pre-detection period where the errors were large.

The subjects compare favorably in RMSE,C with the infinite population values but are poorer than all but one of the finite population values. The introduction of the finite population correction caused an appreciable drop in RMSE,C, particularly for the models with long averages. It appears then, by comparison, that the subjects were not fully utilizing the series constraint. The subjects' RMSE,C dropped on the average only about 0.014 from the random problem type with an average population of close to 60 to the constrained problem type where the population was only 17.

In comparison with these models the subjects seem fairly adept at converging on a new probability after they decided a change had occured. This aspect of the task may well have received the most attention. Concentration on this would lead to increased RMSE,C due to false decisions during the static portion of the sub-problem. This represents a deviation from the explicit instructions.

Conclusion

The human performance on this task was considerably better than had been expected. Two features distinguish this task from other investigations of probability estimation. One is the dynamic set under which the subjects were performing. This expectancy for changing probabilities was probably induced primarily by the subjects' actual experience in estimating the dynamic probabilities. The change in behavior from the large to the small step problems could be viewed as a partial loss of this dynamic set.

The second distinguishing feature was the display and response mechanisms. The particular arrangement of lights, scale, and lever probably had a high stimulus-response compatability.

It seems unlikely that probability estimation is, or at least need be, the limiting factor in human binary decision making. Furthermore, it is reasonable to inquire into probability estimation as a possible useful function of man in future man-machine systems involving information from uncertain or probabilistic sources.

TABLE II

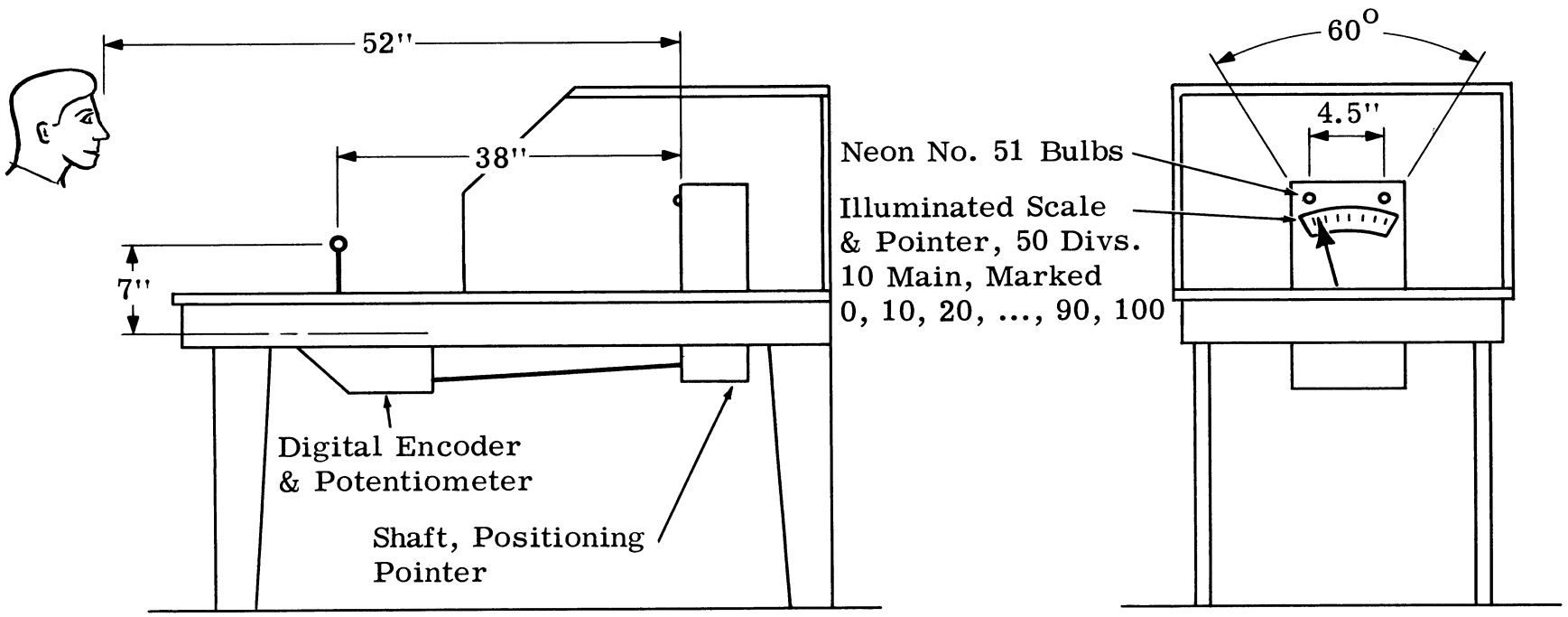
COMPARISON BETWEEN THE PERFORMANCES OF THE SUBJECTS AND THE MATHEMATICAL MODELS

	Detection (Flashes)	Convergence (Flashes)	RMSE	RMSE,C (Probability)	
				Infinite Population	Finite Population
Large Step Problem (step = 0.40)					
Geometric					
r = 0.883	1.1	16.7	0.143	0.101	0.087
* r = 0.915	1.5	23.4	0.148	0.086	0.067
Constant					
N = 14	1.8	12.2	0.155	0.108	0.093
* N = 20	2.5	17.5	0.161	0.090	0.072
Subjects (1fps)	7.5	15.0	0.170	0.091	
Descriptive Model	4.2	14.2			
Small Step Problem (step = 0.15)					
Geometric					
r = 0.953	8.4	22.8	0.094	0.073	0.042
* r = 0.915	4.6	12.4	0.103	0.094	0.073
Constant					
N = 35	11.7	23.3	0.100	0.076	0.048
* N = 20	6.7	13.3	0.110	0.099	0.079
Subjects (1fps)	12.5	17.5	0.112	0.095	
Descriptive Model	7.2	21.0			

* r = 0.915 and N = 20 are the nondiscriminating models.

APPENDIX A

PROBABILITY TRACKING CONSOLE



The duration of the flash was approximately 0.020 seconds.
The intensity was adjusted to provide a clear indication without glare.
The room illumination was low.

APPENDIX B

INPUT PROBABILITY GENERATION

The input step sequence was generated by exhausting the step changes systematically using the tables shown below. The generation procedure was as follows: All problems were started at P = 0.50, the row identified as "probability from," 0.50. The table entries are step sizes, those to the right of the diagonal being positive and to the left negative. One of the step changes in the 0.50 row was selected at random. This step selection led to a new probability, "probability to." This new probability was in turn selected in the "probability from" list and a step change from it selected randomly. The step selections were made without replacement. This procedure was continued until the entire table was exhausted. It was actually necessary to constrain the random selection at times in order to exhaust the table without repeating steps. This selection method gave a "problem" with exactly one step of each size and direction to each probability. The large step problem and the small step problem were produced by separate tables.

The number of flashes at each probability was selected randomly from the set; 42, 54, 66, and 78 for the small step problems, and 35, 51, 74, and 89 for the large step problems. For the constrained problems, both large and small step, the values were multiples of 17; 34, 51, 68, and 85.

Five problems were generated from each table, one for each rate. The same series of steps was used for the random and constrained problem types.

Large Step Problem

Probability From	Probability To .02	.18	.34	.50	.66	.82	.98
.02	-	.16	.32	.48	.64	-	-
.18	.16	-	.16	.32	.48	.64	-
.34	.32	.16	-	.16	.32	.48	.64
.50	.48	.32	.16	-	.16	.32	.48
.66	.64	.48	.32	.16	-	.16	.32
.82	-	.64	.48	.32	.16	-	.16
.98	-	-	.64	.48	.32	.16	-

Steps + and - .16, .32, .48, and .64

Small Step Problem

Probability From	Probability To .08	.14	.26	.32	.44	.50	.56	.68	.74	.86	.92
.08	-	.06	.18	.24	-	-	-	-	-	-	-
.14	.06	-	.12	.18	-	-	-	-	-	-	-
.26	.18	.12	-	.06	.18	.24	-	-	-	-	-
.32	.24	.18	.06	-	.12	.18	.24	-	-	-	-
.44	-	-	.18	.12	-	.06	.12	.24	-	-	-
.50	-	-	.24	.18	.06	-	.06	.18	.24	-	-
.56	-	-	-	.24	.12	.06	-	.12	.18	-	-
.68	-	-	-	-	.24	.18	.12	-	.06	.18	.24
.74	-	-	-	-	-	.24	.18	.06	-	.12	.18
.86	-	-	-	-	-	-	-	.18	.12	-	.06
.92	-	-	-	-	-	-	-	.24	.18	.06	-

Step + and - .06, .12, .18, .24

APPENDIX C

VARIANCES OF SAMPLE AVERAGES FROM FINITE POPULATIONS

Consider a population y_i with mean $\bar{Y}$ and with M members.

Let N samples x_i be drawn from y_i with

$$\bar{x} = \sum_{i=1}^{i=N} w_i x_i \qquad \text{where} \qquad \sum_{i=1}^{i=N} w_i = 1$$

The variance of $\bar{x}$ is

$$E[(\bar{x} - \bar{Y})^2] = E\left\{\left[\sum_{i=1}^{i=N} w_i x_i - \bar{Y}\right]^2\right\}$$

$$= E\left\{\left[\sum_{i=1}^{i=N} w_i (x_i - \bar{Y})\right]^2\right\}$$

$$= E\left\{\sum_{i=1}^{i=N} w_i (x_i - \bar{Y}) \sum_{j=1}^{i=N} w_j (x_j - \bar{Y})\right\}$$

$$= \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} w_i w_j \, E\,[(x_i - \bar{Y})(x_j - \bar{Y})]$$

This expression contains terms which can be written as

$$E(x_i - \bar{Y})^2 = \frac{1}{M} \sum_{k=1}^{k=M} (y_k - \bar{Y})^2$$

and

$$E(x_i - \bar{Y})(x_j - \bar{Y}) = \frac{1}{M(M-1)} \left[\sum_{k=1}^{k=M} \sum_{\ell=1}^{\ell=M} (y_k - \bar{Y})(y_\ell - \bar{Y}) - \sum_{k=1}^{k=M} (y_k - \bar{Y})^2\right]$$

$$= \frac{-1}{M(M-1)} \sum_{k=1}^{k=M} (y_k - \bar{Y})^2 , \qquad i \neq j$$

The variance then becomes

$$E[(\bar{x} - \bar{Y})^2] = \frac{1}{M} \sum_{i=1}^{i=N} w_i^2 \sum_{k=1}^{k=M} (y_k - \bar{Y})^2 +$$

$$\frac{1}{M(M-1)} \left[\sum_{i=1}^{i=N} \sum_{j=1}^{j=N} w_i w_j - \sum_{i=1}^{i=N} w_i^2 \right] \left[- \sum_{k=1}^{k=M} (y_k - \bar{Y})^2 \right]$$

$$= \frac{1}{M(M-1)} \left[M \sum_{i=1}^{i=N} w_i^2 - 1 \right] \sum_{k=1}^{k=M} (y_k - \bar{Y})^2$$

For a sub-problem with $P_i = \bar{Y}$ and of length M_i we have

$$\sigma_i^2 = \frac{1}{M_i} \sum_{k=1}^{k=M_i} (y_k - \bar{Y})^2$$

and we can then write

$$\sigma_r^2(n) = E[(\bar{x} - \bar{Y})] = \left[\frac{M_i}{M_i - 1} \sum_{i=1}^{i=N} w_i^2 - \frac{1}{M_i - 1} \right] \sigma_i^2$$

For the geometric weighted model with $w_i = ar^{i-1}$

$$\sigma_r^2(n) = \left[\left(\frac{M_i}{M_i - 1} \right) \left(\frac{1-r}{1+r} \right) - \frac{1}{M_i - 1} \right] \sigma_i^2$$

For the constant weighted model with $w_i = \frac{1}{N}$

$$\sigma_r^2(n) = \left[\left(\frac{M_i}{M_i - 1} \right) \left(\frac{1}{N} \right) - \frac{1}{M_i - 1} \right] \sigma_i^2$$

APPENDIX D

EXPERIMENTAL PRESENTATION ORDER

The problems were presented to the subjects in the following order.

Where L, S = large and small step problems

R, C = random and constrained problems

1, 2, ..., 5 after R or C = the particular problem

Part 1, 2, 3, 4 = divisions of a particular problem

Session	Problem	Part	Rate (fps)
1	LR1		2
	LC1	1	1
2	SC1		4
	SR1		8
	LC2	1	0.5
3	LR2	1	1
	SR2		2
4	LR3	1	0.5
	LR2	2	1
	LC5		8
5	SR3		4
	LC2	2	0.5
	LR5		4
6	SC2	1	1
	SR4	1	1
7	SC3	1	0.5
	SR5	1	0.5
8	LC3		2
	SR5	2	0.5
9	SC3	2	0.5
	LC4		4
	LR4		8
10	SR5	3	0.5
	SR4	2	1
11	LC2	3	0.5
	SC4		8
	SC3	3	0.5

Session	Problem	Part	Rate (fps)
12	SR5	4	0.5
	SC5		2
13	LC2	4	0.5
	LR3	2	0.5
	LR3	3	0.5
14	SC3	4	0.5
	LC1	2	1
15	LR3	4	0.5
	SC2	2	1

APPENDIX E

INSTRUCTIONS

The following formal instructions were used. The instruction method is discussed in Chapter II.

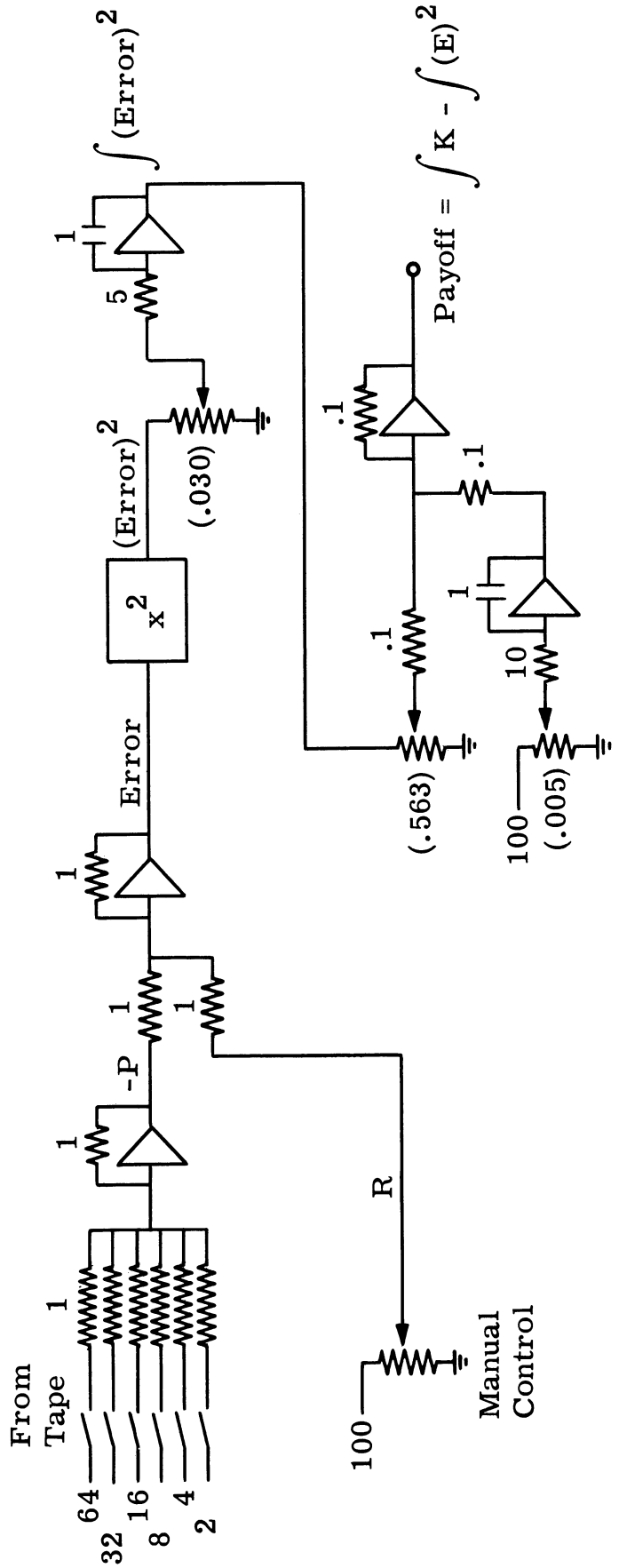
"This experiment is concerned with your ability to estimate probabilities and to follow changes that occur in them as time passes. You will see a display of two lights, a left and a right light. At each flash one or the other of the lights will light, indicating right or left. This is exactly analogous to the drawing at regular intervals of red and green balls from a jar. You will be asked to estimate, by setting a dial, your best guess as to the percentage of balls that are right. The dial is calibrated from 0 to 100 representing no right to all right flashes. For example, if you think that about 68% of the flashes are right then set the dial at 68. The actual percentages cover the entire range from 0 to 100 and have all values in between. The percentages do not necessarily fall on the dial markings.

"The important new work to come out of this experiment is your ability to notice changes in the percentages and to follow the changing percentage with the dial setting. The analogy with the balls in the jar is the case where one or the other color is being taken out of the jar by another person without your knowledge. At times the percentage will change slowly in a continuous fashion. At other times the percentage will change suddenly, as though a whole handful of one color had been removed. If you are uncertain as to the percentage set the dial at 50.

"You will be paid according to how well you do. At the end of each problem, 10 to 25 minutes, you will be able to read the amount of money off the meter on the computer. The computer calculates the difference between your estimate and the actual probability, the error, and accumulates this error over the problem. It also adds up a constant amount of money per minute. You are paid the difference. The computer is adjusted so that if you left the lever at 50 you would get no money.

"You will wear a pair of earphones and have a microphone. A low 'seashore' type noise will be fed into the earphones in order to mask out noises from the street and the laboratory. When I talk to you the noise will be removed. You can be heard at all times through your microphone. You are welcome to make verbal comments during the experiment. These are not being recorded and any sort of language is acceptable."

APPENDIX F



F

ANALOG COMPUTER CIRCUIT FOR PAYOFF

APPENDIX G

TWO QUALITATIVE RESPONSE EXCEPTIONS

On two occasions during approximately 70 hours of tracking the tracking response was qualitatively variant from the norm. These two situations lasted for a total of approximately 35 minutes.

The first occurred during the pilot experiment. During one particular problem in the third session a subject was accumulating error at a much higher rate than in any of the previous sessions or problems. Upon inspection of her records it was noted that detection was considerably higher than it had been before. The instructions concerning the error formation and the payoff were repeated with special emphasis on the rapid error build up with large discrepancies between probability and response. Her response returned to normal on the next problem.

The hypothesis here is that she was computing the new probability to a high degree of accuracy before she responded to the change. The "normal" response produces movement toward the new probability as soon as it is perceived, with further refinements as more data, flashes, are accumulated.

The second anomaly occurred in the response of a subject in his 12th session of the main experiment. He was tracking a large step problem at 2 fps. The experimenter noted that the payoff was going negative; the error accumulation was faster than the pay accumulation. Upon examining the records it was established that for about the first 3/4 of the problem, about 15 minutes, the response was the mirror image of the proper or normal response. The scale was reversed in relation to the light flashes.

A check on the equipment failed to reveal any malfunction. Upon questioning after the session the subject stated that he was a bit mixed up at times. He evidently had no idea that he was doing a fairly good job of mirror image tracking.

He was given this particular problem again in a special 16th session and this second run was used in the analysis.

APPENDIX H

DATA NOT AVERAGED OVER SUBJECTS

Figures 22 through 28 show some of the principal variable interactions for individual subjects. The data for detection and convergence show appreciable magnitude variations between subjects but maintain the same qualitative relationships as regards direction of change and the distinction between the small and the large step problems. Subject S-1 was quite consistently slower in his response than the other three. All subjects show a similar increase in RMSE with rate from 1 to 8 fps. Subject S-3 is consistently higher. All subjects show a decrease in RMSE,C from 0.5 to 2 fps. Subjects S-2 and S-4 show a continued decrease at 4 and 8 fps whereas S-3 and S-1 show an increase. FAR decreases with rate for three subjects; subject S-1 having little variation by comparison.

Mean error did not vary significantly among subjects.

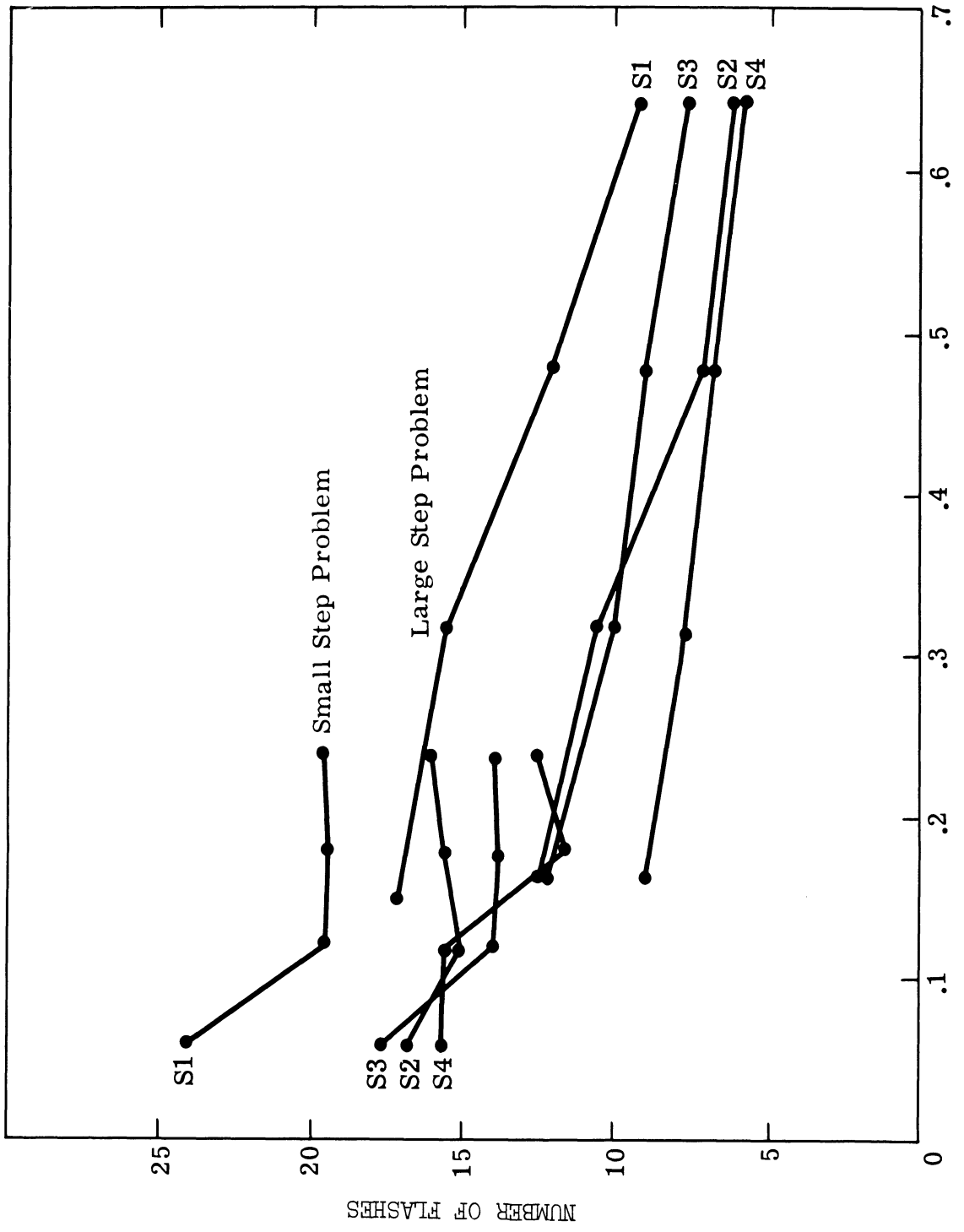


Figure 22. Detection as a Function of Step Size for Four Subjects. Detection is measured in flashes.

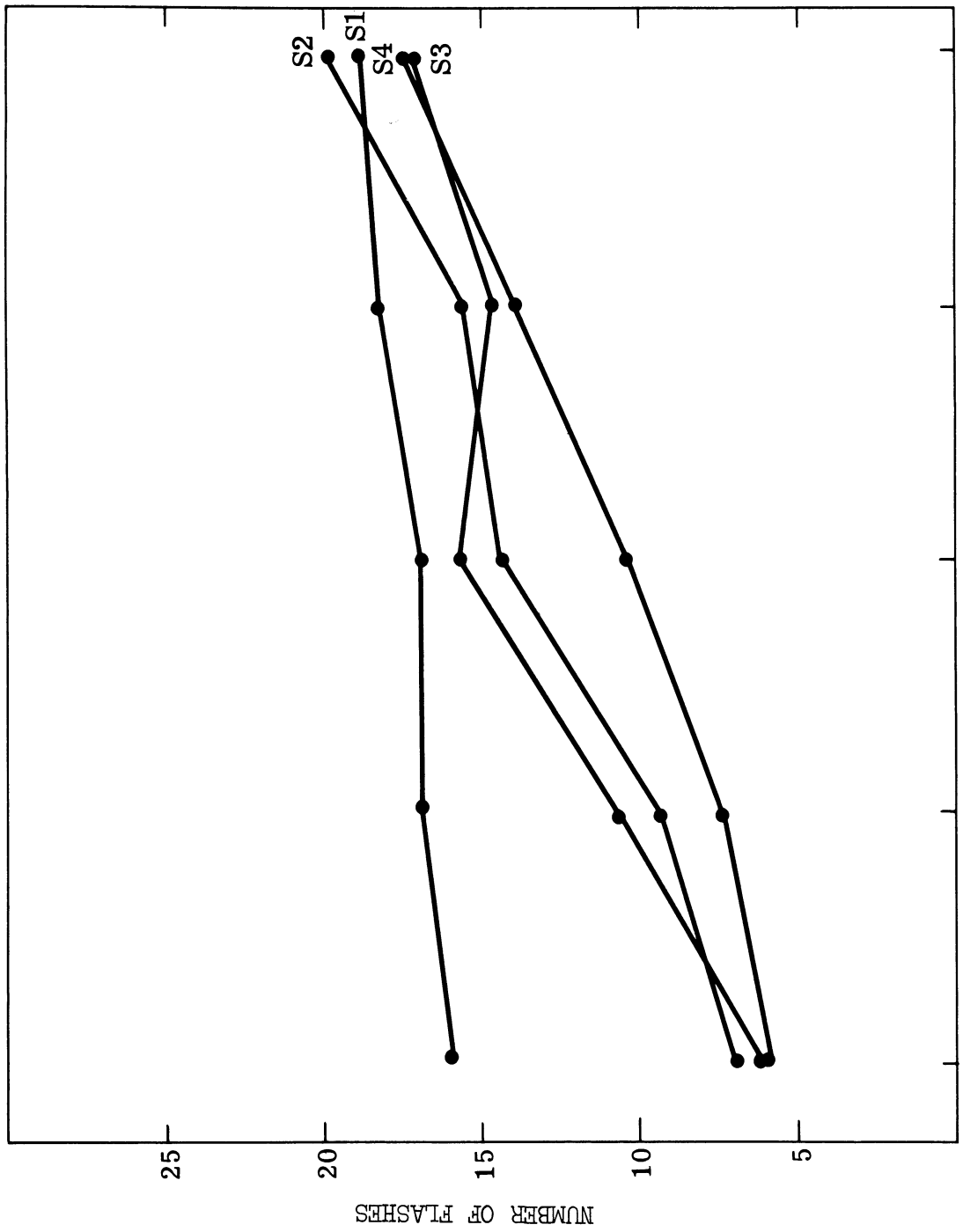


Figure 23. Detection as a Function of Flash Rate for Four Subjects. Detection is measured in flashes.

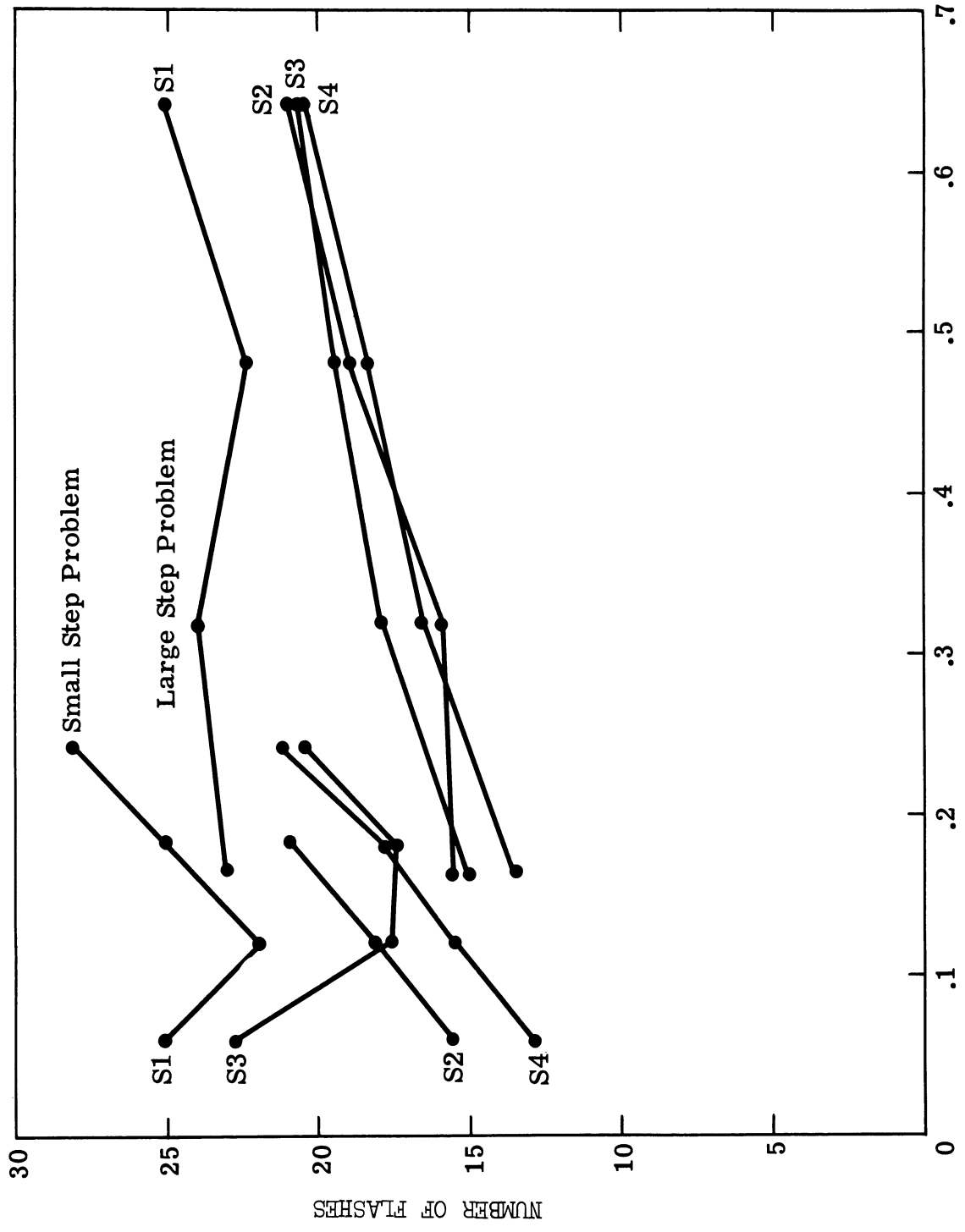


Figure 24. Convergence as a Function of Step Size for Four Subjects. Convergence is measured in flashes.

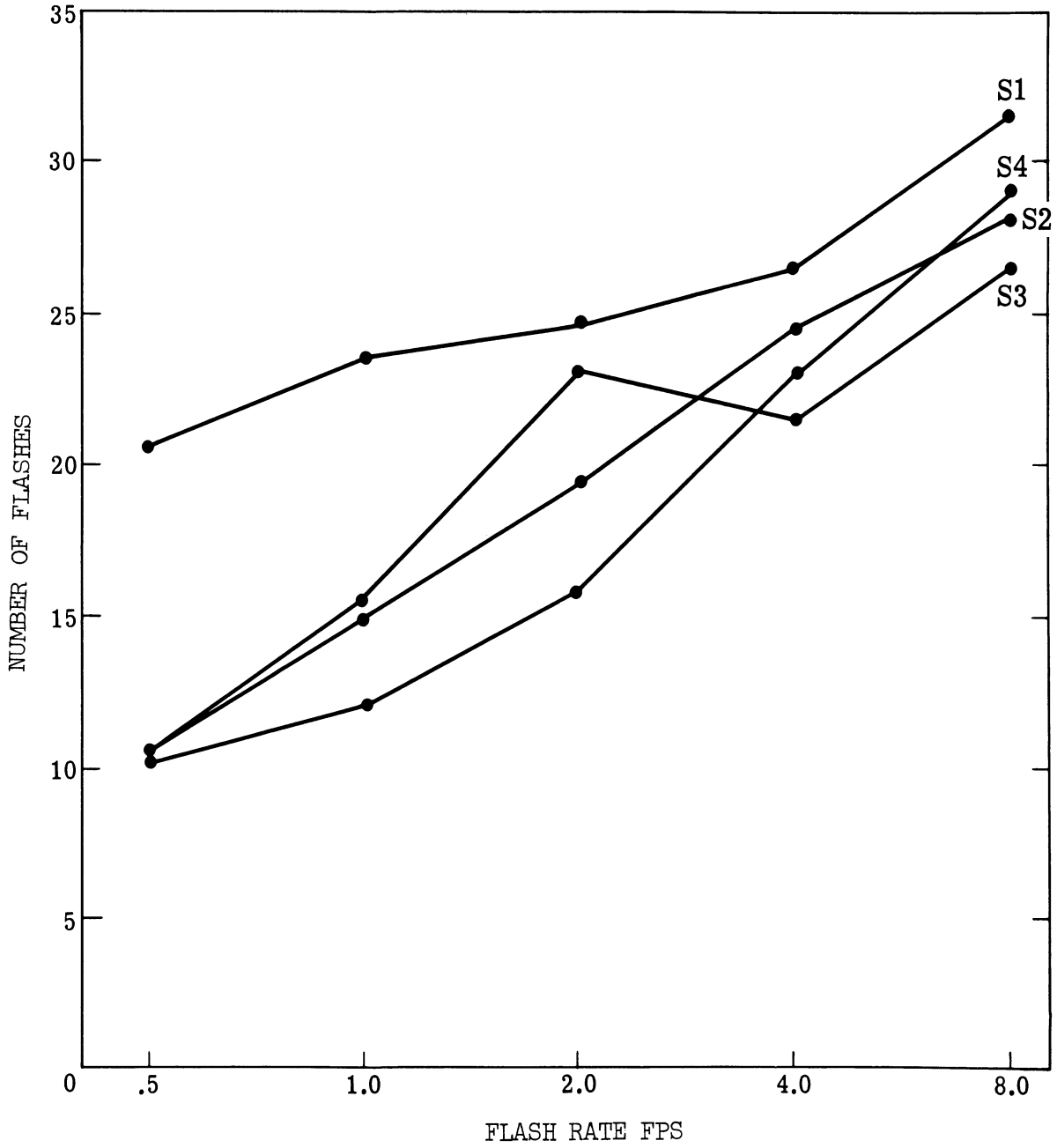


Figure 25. Convergence as a Function of Flash Rate for Four Subjects. Convergence is measured in flashes.

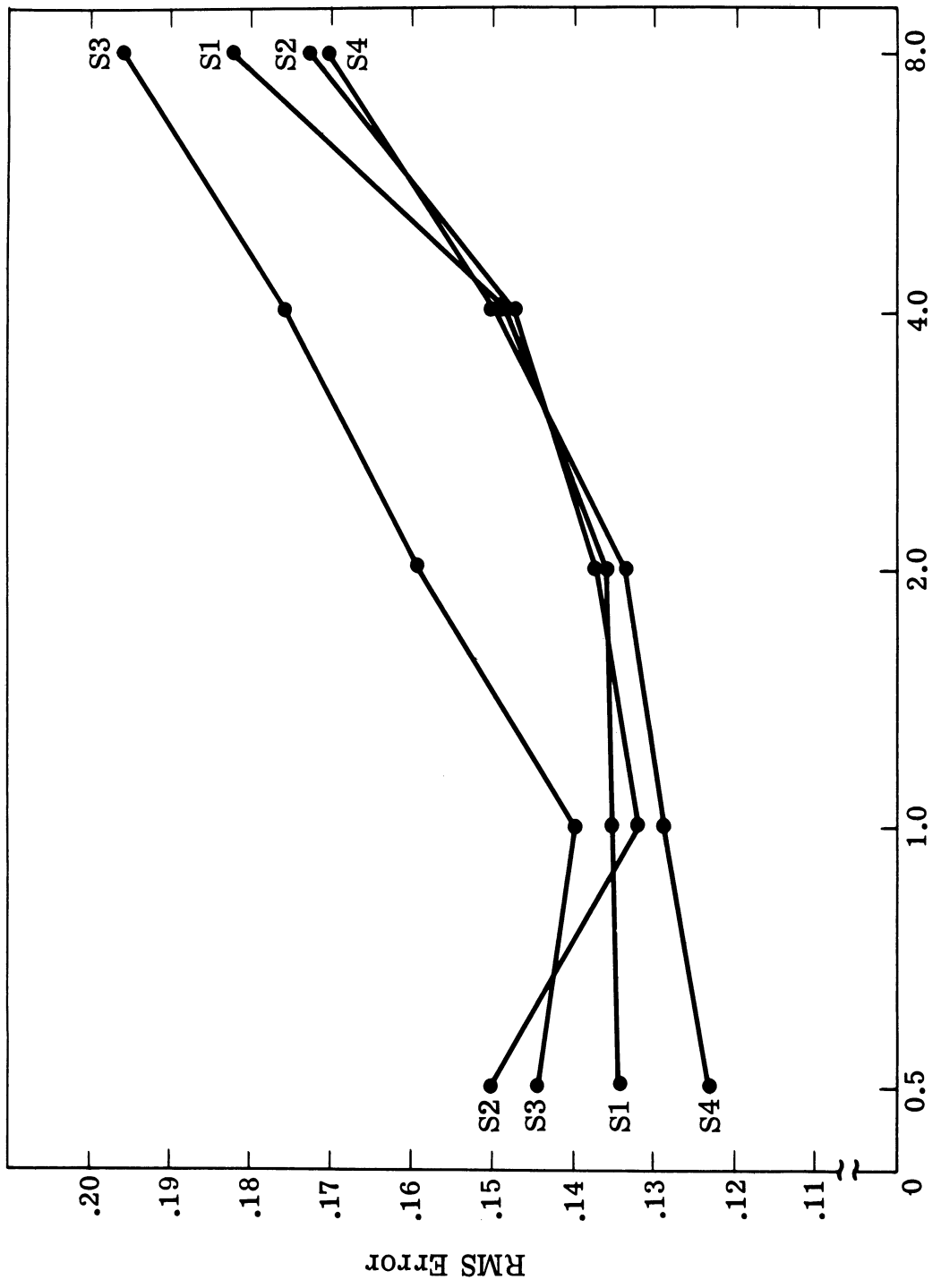


Figure 26. Root Mean Square Error as a Function of Flash Rate for Four Subjects.

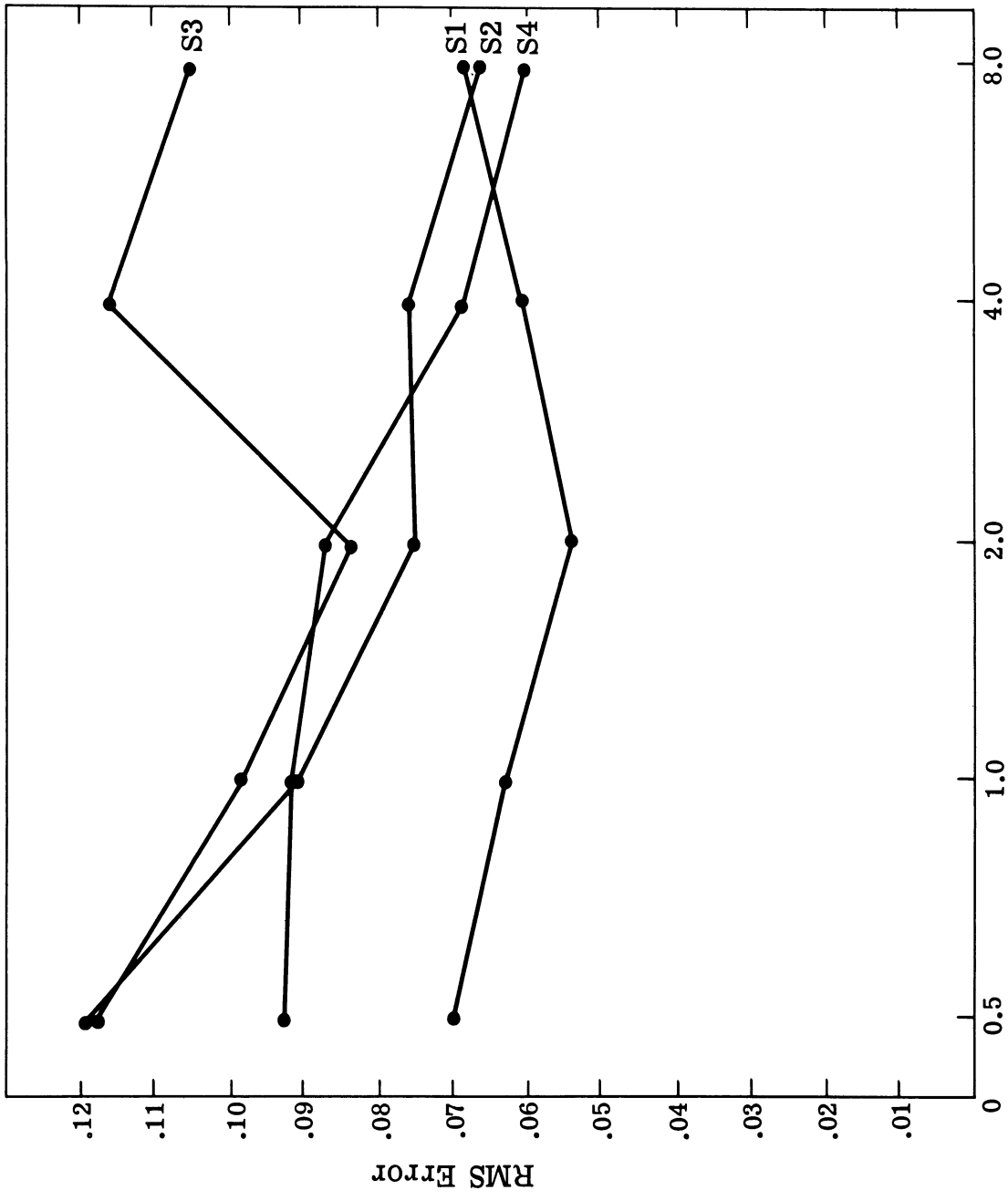


Figure 27. Root Mean Square Error After Convergence as a Function of Flash Rate for Four Subjects.

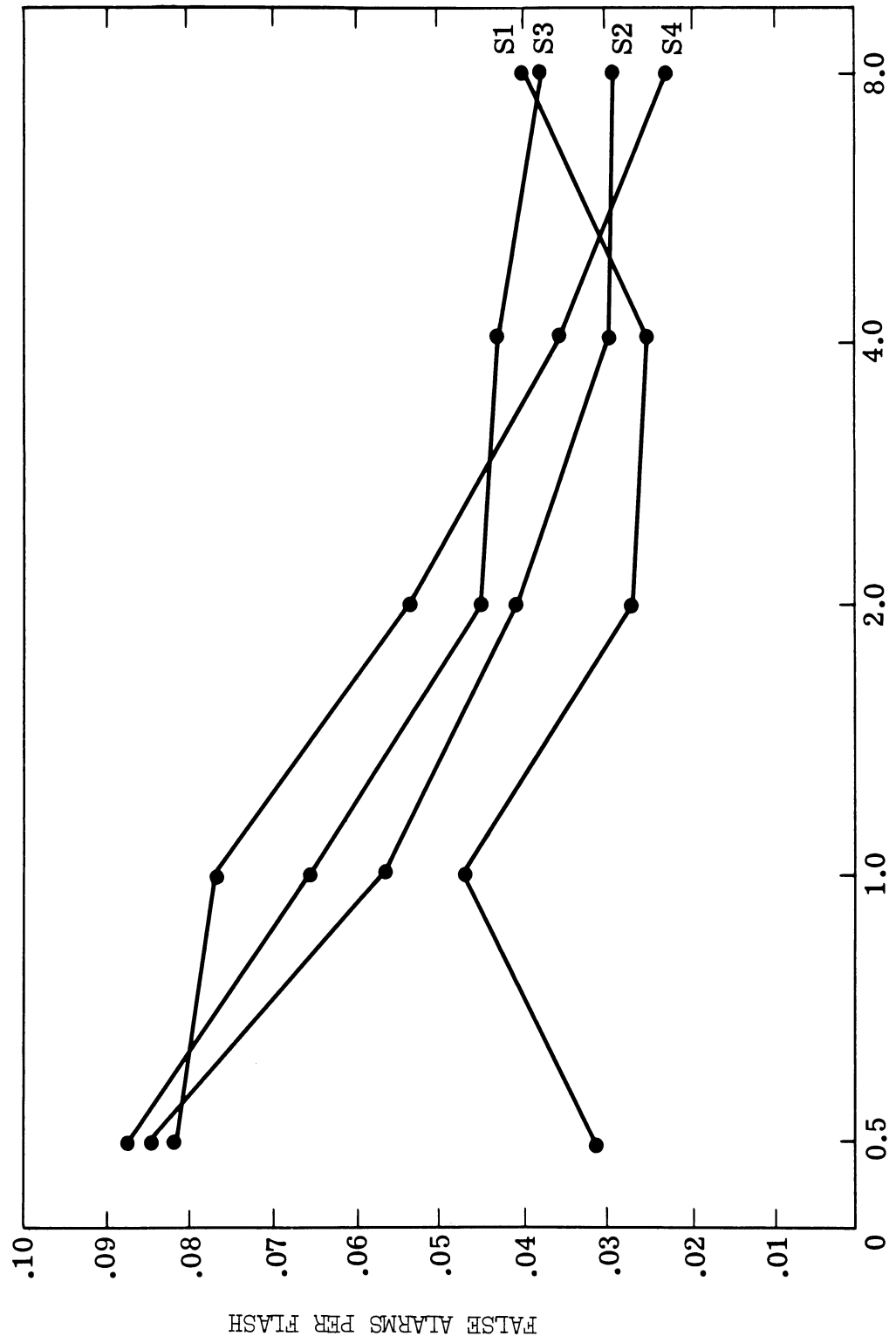


Figure 28. False Alarm Rate, in False Alarms Per Flash, as a Function of Flash Rate for Four Subjects

REFERENCES

1. Cohen, J. and Dinnerstein, A. J. Flash Rate as a Visual Coding Dimension for Information. AF WADC TR 57-64, 1958.

2. Conrad, R. and Hille, B. A. "The Decay Theory of Immediate Memory and Paced Recall." Canadian J. Psychol., 12 (1958) 1-6.

3. Dember, W. N. "The Relation of Decision-Time to Stimulus Similarity." J. Exp. Psychol., 53 (1957) 68-72.

4. Edwards, W. "The Theory of Decision Making." Psychol. Rev., 51 (1954) 380-417.

5. Edwards, W. "Behavioral Decision Theory." Annu. Rev. Psychol., 12 (1961) 473-498.

6. Edwards, W. "Probability Learning in 1000 Trials." J. Exp. Psychol., 62 (1961) 385-394.

7. Erlick, D. E. Judgments of the Relative Frequency of Sequential Binary Events. Aerospace Medical Labs. WADC TR 59-580, 1959.

8. Flood, M. M. "Environmental Non-Stationarity in a Sequential Decision Making Experiment." In R. M. Thrall, C. H. Coombs, and R. L. Davis (Eds.), Decision Processes. New York: Wiley, 1954.

9. Forsyth, D. M. and Chapanis, A. "Counting Repeated Light Flashes as a Function of Their Number, Rate of Presentation, and Retinal Location Stimulated." J. Exp. Psychol., 56 (1958) 385-391.

10. Gardner, R. A. Perception of Relative Frequency as a Function of the Number of Response Categories. AMRL (1959) 408.

11. Goodnow, J. J. and Pettigrew, T. F. "Effect of Prior Patterns of Experience Upon Strategies and Learning Sets." J. Exp. Psychol., 49 (1955) 381-389.

12. Grant, D. A. "Information Theory and the Discrimination of Sequences in Stimulus Events." In B. McMillan (Ed.), Current Trends in Information Theory. Pittsburgh: University of Pittsburgh Press, 1954.

13. Hake, H. W. "The Perception of Frequency of Occurrence and the Development of 'Expectancy' in Human Experimental Subjects." In H. Quastler (Ed.), Information Theory in Psychology. Glencoe, Illinois: The Free Press, 1954.

14. Hake, H. W. and Hyman, R. "Perception of the Statistical Structure of a Random Series of Binary Symbols." J. Exp. Psychol., 45, (1953) 64-74.

15. Hornseth, J. P. and Grant, D. A. The Discrimination of Random Series of Stimulus Frequencies as a Function of Their Relative and Absolute Values. AFPTRC Research Bulletin TR-54-76, 1954.

16. Hyman, R. "Stimulus Information as a Determinant of Reaction Time." J. Exp. Psychol., 45 (1953) 188-196.

17. Jarvik, M. E. "Probability Estimates and Gambling." In Mathematical Models of Human Behavior. Stamford, Connecticut: Dunlap and Associates, Inc., 1955.

18. Mackworth, J. F. "Paced Memorizing in a Continuous Task." J. Exp. Psychol., 58 (1959) 206-211.

19. Mackworth, N. H. and Mackworth, J. F. "Visual Search for Successive Decisions." British J. Psychol., 49 (1958) 210-221.

20. Neimark, E. D. and Shuford, E. H. "Comparisons of Predictions and Estimates in a Probability Learning Situation." J. Exp. Psychol., 57 (1959) 294-298.

21. Pollack, L., Johnson, L. B. and Knaff, P. R. "Running Memory Span." J. Exp. Psychol., 57 (1959) 137-146.

22. Reese, R. W. "The Application of the Theory of Physical Measurement to the Measurement of Psychological Magnitudes, with Three Experimental Examples." Psychological Monographs, 55 (1943) 1-88.

23. Schreiber, R. J. "Estimates of Expected Value as a Function of Distribution Parameters." J. Exp. Psychol., 53 (1957) 218-220.

24. Stevens, J. C. and Shickman, G. M. "The Perception of Repetition Rate." J. Exp. Psychol., 58 (1959) 433-440.

25. Taves, E. H. "Two Mechanisms for the Perception of Visual Numerousness." Archives of Psychol., (1941) 265.