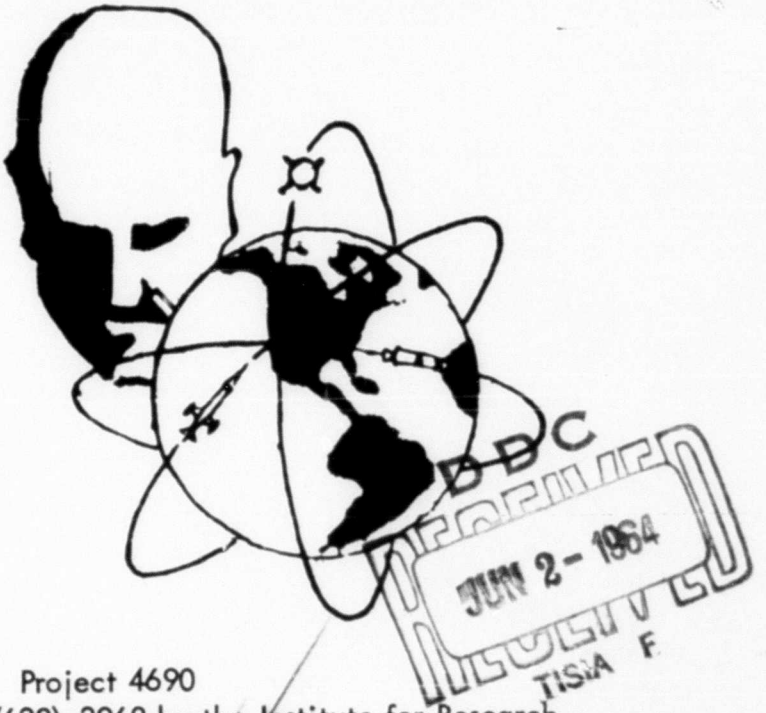
MICROSTRUCTURE OF GUESS PROCESSES: PART C

TECHNICAL DOCUMENTARY REPORT No. ESD-TDR-63-548
SEPTEMBER 1963

Masanao Toda

DECISION SCIENCES LABORATORY
DEPUTY FOR ENGINEERING AND TECHNOLOGY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
L. G. Hanscom Fld, Bedford, Massachusetts



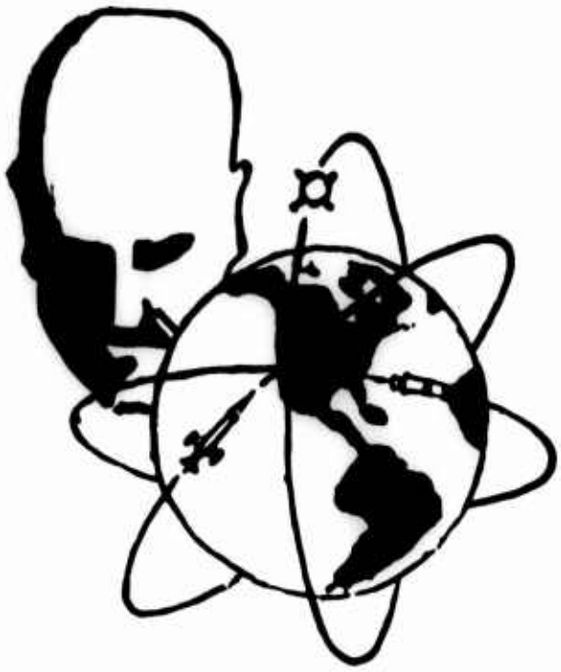
Project 4690
(prepared under contract AF19(628)-2968 by the Institute for Research,
257 South Pugh Street, State College, Pa.)

MICROSTRUCTURE OF GUESS PROCESSES: PART C

TECHNICAL DOCUMENTARY REPORT No. ESD-TDR-63-548
SEPTEMBER 1963

Masanao Toda

DECISION SCIENCES LABORATORY
DEPUTY FOR ENGINEERING AND TECHNOLOGY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
L. G. Hanscom Fld, Bedford, Massachusetts



Project 4690
(prepared under contract AF19(628)-2968 by the Institute for Research,
257 South Pugh Street, State College, Pa.)

FOREWARD

This report is adapted from the paper read at the Harvard Statistics Colloquium February 1962. The author acknowledges his thankfulness to Drs. George A. Miller and Paul Kolers and to Mr. Shiro Imai.

This paper is the third in a series reporting work performed for the Decision Sciences Laboratory, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts.

This is Report Number 7 from the Division of Mathematical Psychology, Institute for Research, 257 South Pugh Street, State College, Pennsylvania.

MICROSTRUCTURE OF GUESS PROCESSES: PART C

ABSTRACT

Trial-to-trial changes in the proportion of human subjects predicting the occurrence of one of two events in a complex sequence of binary events (probability learning) are analyzed in terms of several simple models. The direction of change predicted by linear-operator reinforcement models (Estes, Bush and Mosteller) is wrong on about 75% of the trials. A no-learning model, a time-dependent decay model, and a cycle-dependent decay model are used to provide some insight into the nature of probability learning.

Some suboptimal procedures for estimating parameters of stochastic processes are compared. The method of minimum absolute error is recommended as being very useful.

PUBLICATION REVIEW AND APPROVAL

This Technical Documentary Report has been reviewed and is approved.

WALTER E. ORGANIST
Chief, Operator Performance Division
Decision Sciences Laboratory

ANTHONY DEBONS, Col, USAF
Director
Decision Sciences Laboratory

LIST OF FIGURES

LIST OF FIGURES (Continued)

BLANK PAGE

MICROSTRUCTURE OF GUESS PROCESSES

Masanao Toda

A couple of years ago, Professor Mosteller gave a presidential address to the Psychometric Society entitled "The mystery of the missing corpus" (Mosteller, 1958). In a slightly different context, I sometimes feel about my own research on guess process that I am trying to solve a like mystery called "The Case of a Deceptive Beauty." But unlike Sherlock Holmes or Perry Mason, I am no genius as a detective. I am just a plain, ordinary man with dogged perseverence, and I have just succeeded in getting a confession from my suspect, that deceptive beauty, known as guess process, also as probability learning. And I am still wondering, whether this confession might be another deception, and I am just making a fool of myself by triumphantly talking about this confession. Anyway, the confession is not yet consistent, and I am not yet at the stage of getting a successful trial.

However, here is one thing about which I can talk with complete confidence; this deceptive beauty, probability learning, has a very complicated character, no matter how plainly simple she may appear, and no matter how many psychologists are honoring her simplicity by sonnets in the form of simple stochastic learning theories.

My plan for today's talk on my unfinished detective story is like this. First, I will introduce her to you formally with appropriate courtesy;

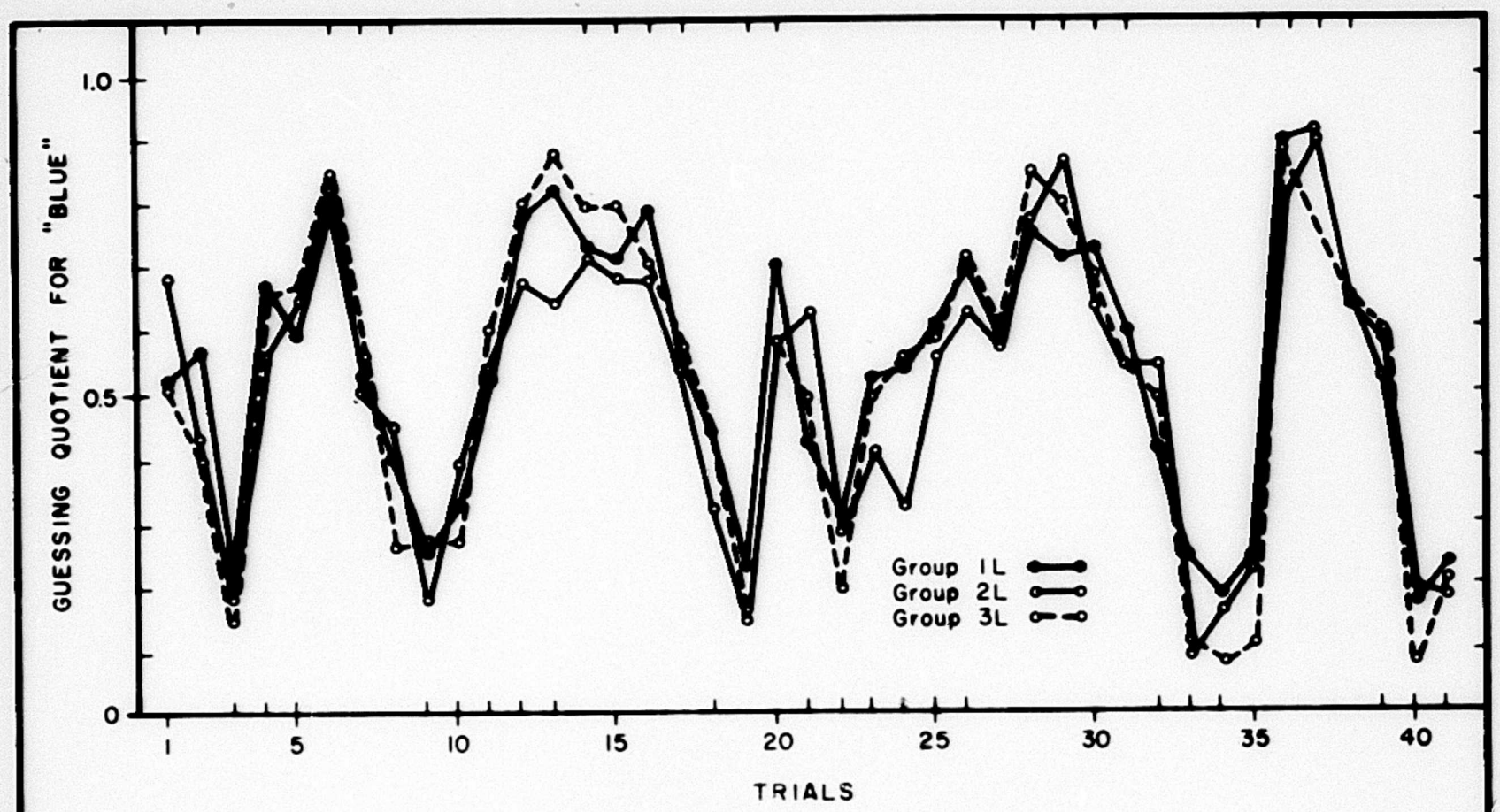


Fig. 1 Guessing curves for the long-run sequence. The short line attached either to the top or to the bottom of the graph corresponding to each trial indicates the event occurs on the trial, "blue" if top, and "red" if bottom.

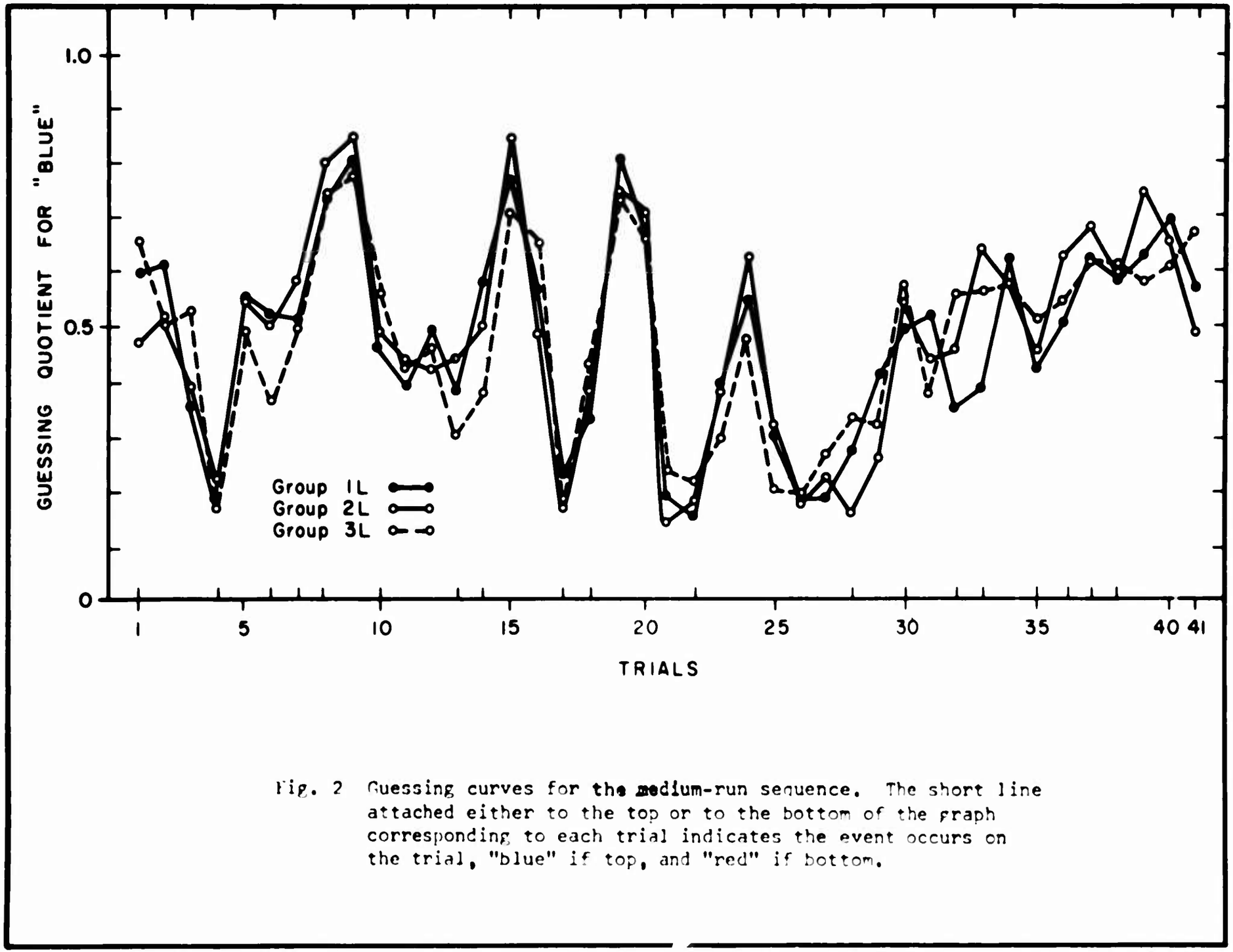


Fig. 2 Guessing curves for the medium-run sequence. The short line attached either to the top or to the bottom of the graph corresponding to each trial indicates the event occurs on the trial, "blue" if top, and "red" if bottom.

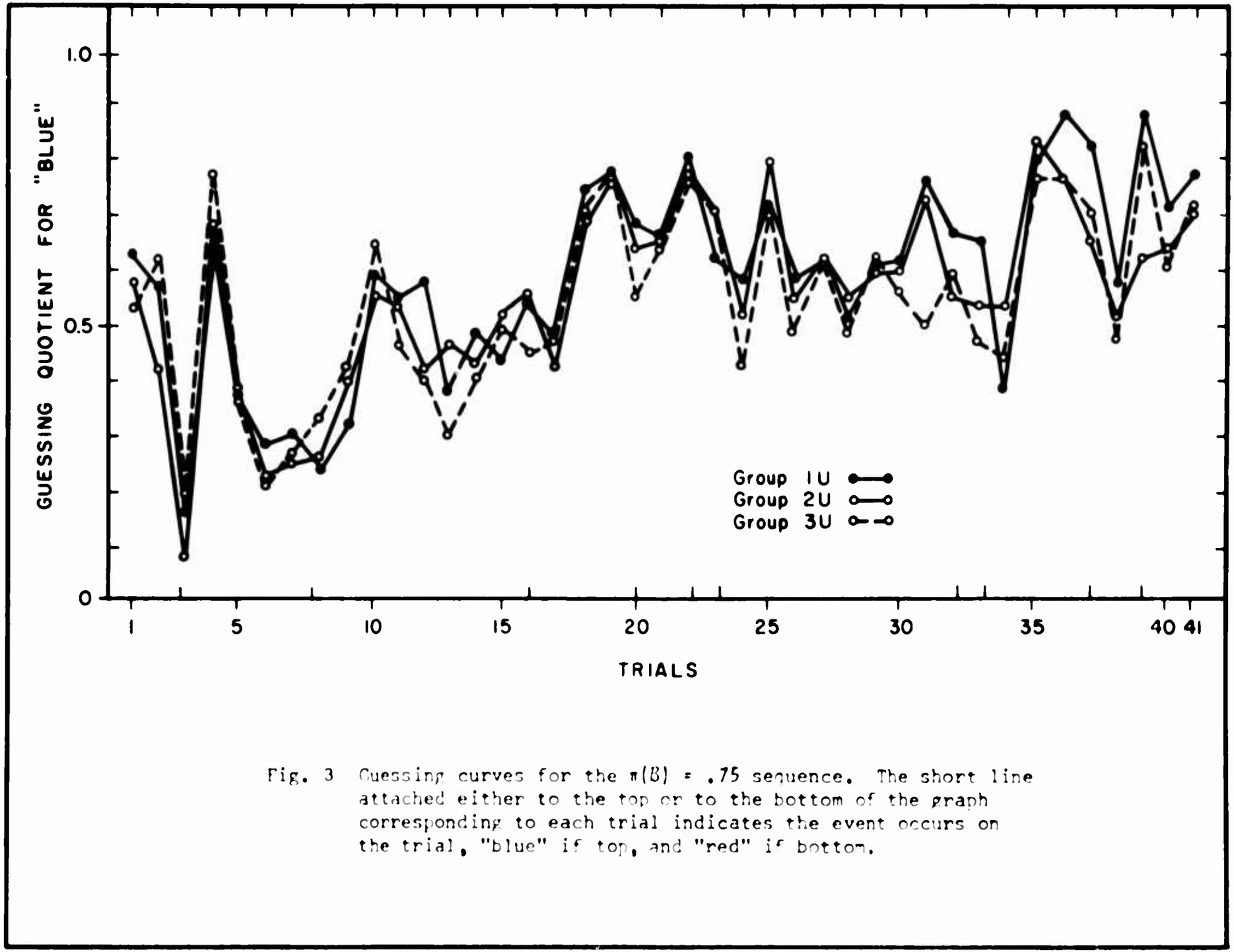


Fig. 3 Guessing curves for the $\pi(B) = .75$ sequence. The short line attached either to the top or to the bottom of the graph corresponding to each trial indicates the event occurs on the trial, "blue" if top, and "red" if bottom.

BLANK PAGE

second, I will tell you something about her shadowy inside life when she is out of sight, and then, finally, I will talk something about police science, or parameter estimation.

I think most of you are familiar with guessing experiment, or two-armed bandit experiments. So, I will show you just an example of the experimental procedure. Imagine a deck of, say, playing cards. The experimenter shows these cards one by one from the top of the deck. Now, the experimental Ss' task is to predict the color of each card each time before the card is shown. This is just a kind of game, and Ss are encouraged to maximize the number of correct predictions. That's all. Suppose the total number of Ss is N. Suppose the number of Ss who predicted "black" on trial i is n_i. Then I call n_i/N the guessing quotient with respect to "black" response on trial i. By plotting these guessing quotients on all the trials, we obtain the guessing curve. You will see examples of guessing curves in Figures 1, 2, and 3. Please look at the Figure 3 first. The short lines attached to the top and the bottom lines of the graph represent the arrangement of the cards used in the experiment. There are two short lines attached to the leftmost part of the top line, which are then followed by a blank. And wherever there is a blank on the top line you will find a short line on the bottom line. These three short lines then mean that the first two cards were blue, and the third was red. So you will see that the arrangement of the cards, or the sequence of events, used here is a random sequence with probability $\pi(B) = .75$ for obtaining blue. Now the ordinate of this graph shows the values of guessing quotients with respect to prediction of blue. Three groups of Ss were given this same sequence. So the three points corresponding to the first trial indicate that in each group about 50 or 60% of the Ss predicted blue as the color of the first card.

After Ss made their responses, they were shown the first card, which was blue, and each of them recorded his prediction and the color of the card in the answer sheet, and then they proceeded to predicting the color of the next card.

Now, I think some of you who are familiar with guessing experiments might be puzzled by this figure. Guessing curves you find in psychological journals do not usually look like this. Usually, they start off at about 0.5, and smoothly and monotonically approach a certain asymptote. But here in Figure 3, there is nothing smooth and nothing monotonic. This reminds me of a joke. Mona Liza had a toothache, and Leonardo had an ideal model. Now you can pull out all the teeth of the original guessing curves like those presented in Figure 3, by averaging guessing quotients over each block of ten or more trials. And this is what usually is done by psychological artists, or dentists, and as a result we get mysteriously simple and smooth guessing curves. Averaging across blocks of trials is, of course, a completely legitimate procedure if these structures of the curves are just outcomes of random fluctuations. But random fluctuations cannot be reproduced so regularly as occurs in Figure 3 as well as in Figures 1 and 2.

However, I could hardly do justice to the beauty of the averaged guessing curves if I said it is all due to the plastic surgery of averaging. There is a mystery something beyond that, and it is the value of the asymptote to which smooth averaged guessing curves approach. As far as experiment is conducted under ordinary conditions, i.e., when Ss are just guessing and not making money in proportion to the number of correct predictions, the asymptotic value of $P'(B)$, the guessing quotient with respect to the event B, is almost always approximately equal to $\pi(B)$ no matter what the value $\pi(B)$. This effect is called probability matching.

This result has puzzled many people, since to match response probability with event probability is so obviously non-optimal if Ss are maximizing the number of correct predictions and if they know the event sequence is random. Suppose $\pi(B) > 1 - \pi(B)$. Then, S can get $\pi(B)$ as the mean number of correct predictions if he always predicted blue. But if he matches his prediction probability with $\pi(B)$ then his mean number of hits reduces to $\pi^2 + (1 - \pi)^2 < \pi$. Equality holds only for $\pi = 1$.*

The reactions to this effect among psychologists who were interested in this process were not unanimous. A group of people, including myself, were rather deeply annoyed by this apparent irrationality, and attempted to prove either or both of the following two hypotheses: (1) Ss were not simply maximizing the plain, unweighted total number of correct predictions, (2) Ss were not perceiving the event sequence as random. One of my Ss told me that he could not resist the temptation of trying to hit the jackpot by predicting a very infrequent event. If this kind of uneven utilities for more frequent and less frequent events is responsible for probability matching, we should be able to get rid of probability matching by inducing an even utility distribution by means of paying Ss money in proportion to the correct predictions. This hypothesis has been very well confirmed by a couple of experiments done by different people. Obviously Ss preferred real pennies to imaginary jackpots.

---

*PROOF. $\pi > 1 - \pi$. Put $1/2 + \varepsilon = \pi$, $0 < \varepsilon < 1/2$. Then we have $\pi^2 + (1 - \pi)^2 = (1/2 + \varepsilon)^2 + (1/2 - \varepsilon)^2 = 1/2 + 2\varepsilon^2$ for the mean number of hits per trial under the probability matching strategy. On the other hand, the mean number of hits per trial under the pure strategy of predicting B all the time is $\pi = 1/2 + \varepsilon$. The latter is greater than the former since $2(\pi - (\pi^2 + (1 - \pi)^2)) = 1 + 2\varepsilon - 4\varepsilon^2 = 2\varepsilon(1 - 2\varepsilon) > 0$.

The reaction of another group of people was a kind of artistic inspiration. As a consequence, we are now able to appreciate a couple of masterpieces of mathematical art.

The greatest of all is, according to my opinion, Estes' model (Estes, 1959), since he uses only two parameters to describe guessing curves. I am using wrong words. He calls his model a theory, and he is not describing, but predicting, since a theory should predict, not describe. You may wonder if it is possible to predict without describing. But Estes did it, and I will show you how this stunt was done. His basic assumption will be stated like this:

$$p_{i+1} = \lambda p_i + \alpha_i(1 - \lambda), \quad 0 < \lambda < 1.$$

His original expression (Estes, 1950) is different from this, but these are equivalent. Now, p_i is the probability of predicting a specified event on trial i, α_i is a function taking the value 1 or 0 according as the specified event has occurred or not on trial i, respectively. In psychology, this type of theory belongs to a class of reinforcement theories, since the event obtained on trial i reinforces the response oriented to that particular event. λ is a parameter, and another parameter in this model is obviously p_1.

Now this equation has a form directly applicable to individual guessing quotients and it should be true, if a reinforcement theory of this type is to have any validity at all, that individual guessing quotient in general increases when the specified response is reinforced and decreases when the alternative response is reinforced. I tested this assumption with my data and the assumption was confirmed only in about 75% of the whole set of 200 trials. Now Estes did not use individual guessing quotients but only their

averages. So, in this equation, α_i is also replaced by its average, or its expectation π. Once this was done, it is really easy to obtain an explicit form for p_{i+1}.

$$p_{i+1} = \pi - \lambda^i(\pi - p_1)$$

Since $0 < \lambda < 1$, now we have

$$\lim_{i \to \infty} p_i = \pi.$$

Thus the probability matching effect is predicted, even though its prediction concerning the direction of change of guessing quotient is wrong 75% of the time.

Parsimony in the number of necessary parameters is certainly a virtue in a good theory. But, according to an oriental belief, a virtue is something hard to obtain. So, I like to take a hard way, starting with a purely descriptive model which has as many parameters as possible to take care of various information involved in the data. The number of parameters may then be reduced if one is lucky enough to find that some of the parameters are redundant.

I should say that this had been my belief before I got into the present problem. Then, I found out that I was too optimistic. If I use a model with too many parameters, I would simply be stuck with the impossibility of parameter estimation, and furthermore, there is no purely descriptive model. A model becomes a theory once the model is applied to real data. So, there is always a danger in using a single model for the purpose of analysis of data, even if the model is primarily oriented toward a description. That

much was the lesson I obtained from my frustrating experience of trial-and-errors and I am now just hoping that, after hearing my experience, some of you could tell me if there is a better strategy.

Now let me get back to the data. The three sequences used in my first experiment, the results of which are shown in Figures 1, 2, and 3, are named "long-run sequence," "medium-run sequence" and "uneven probability sequence." The long-run sequence is characterized by $\pi(X) = 0.50$ and also by the conditional probability $\pi_X(X) = .70$ (X = "Blue" and "Red"). The medium-run sequence is characterized by $\pi(X) = 0.50$ and $\pi_X(X) = .47$. The third sequence is characterized by $\pi(B) = \pi_B(B) = .75$ and $\pi(R) = \pi_R(R) = .25$.

Now all the three sets of guessing curves shown in Figures 1, 2, and 3, have definite but different structures. If I want to say anything more specific, however, I need a descriptive model. And at that stage of my research, there was none. Even the most well-formed descriptive model of learning, the Bush-Mosteller model (Bush & Mosteller, 1955), has too strong a set of assumptions to be applicable to these structures.

This much seemed to be obvious: If a sequence of responses had a structure, and if the structure was different for different event sequences, then the structure of response sequences should somehow correspond to the structure of event sequences. This hypothesis was easy to check, particularly as I had an impression that Ss were responding principally to run length. Although this could not be the only factor responsible for the response structure, I decided to employ a simple pilot model which, while it was very poor in its descriptive capacity and just absurd as a theory, had the virtue of giving no trouble in estimating its parameters and could serve to test my hypothesis about run length.

BLANK PAGE

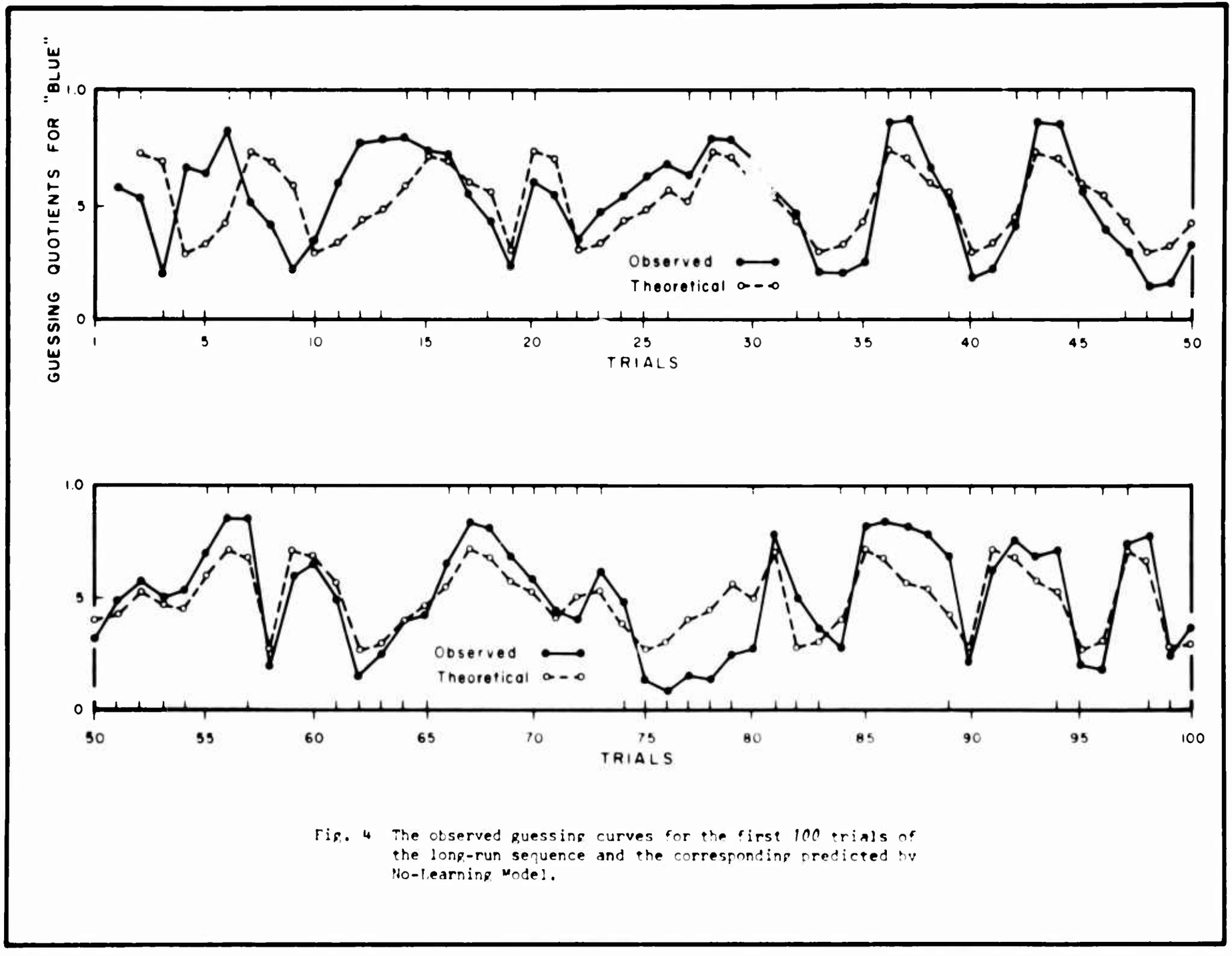


Fig. 4 The observed guessing curves for the first 100 trials of the long-run sequence and the corresponding predicted by No-Learning Model.

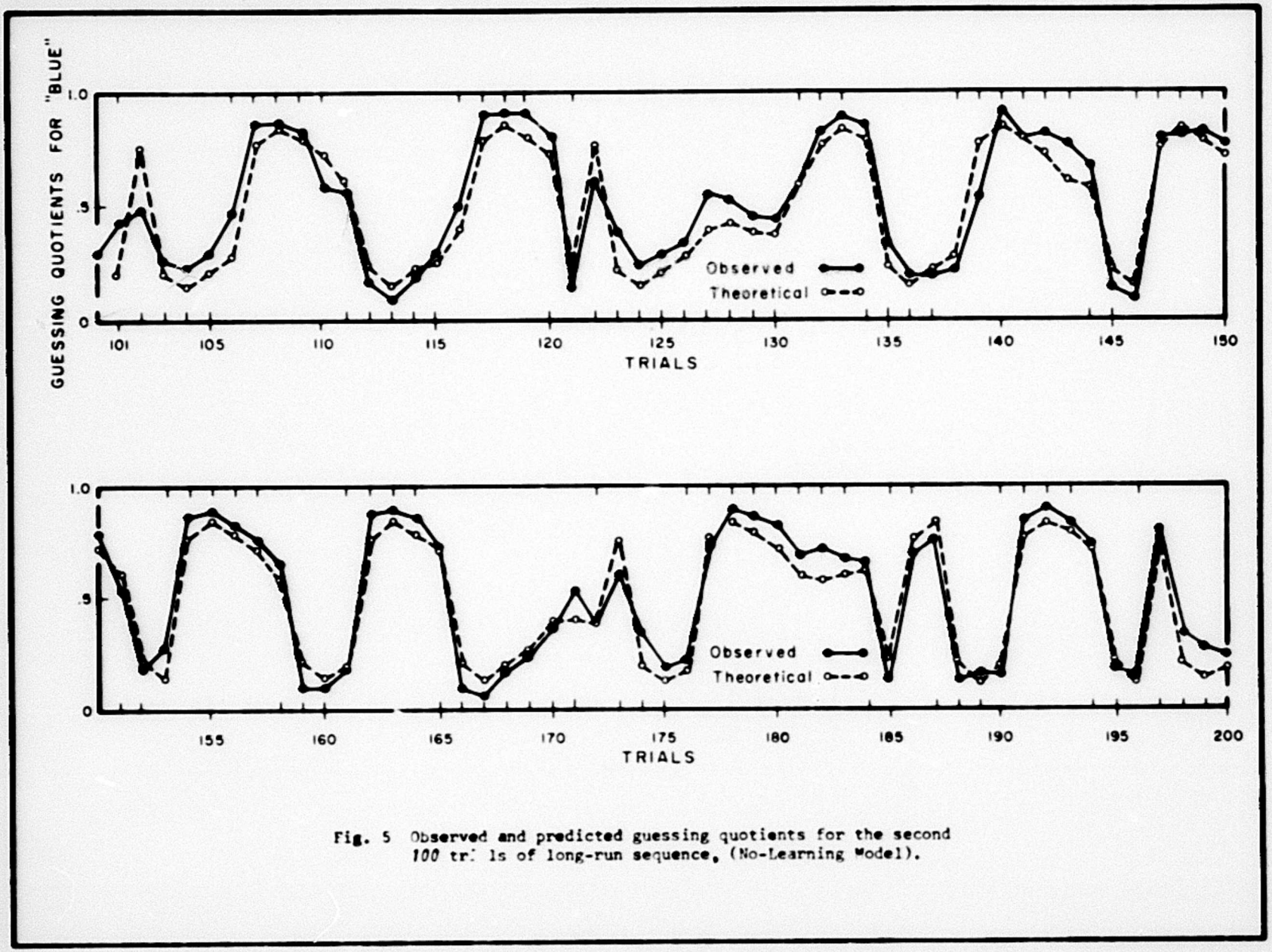


Fig. 5 Observed and predicted guessing quotients for the second *100* tr: ls of long-run sequence, (No-Learning Model).

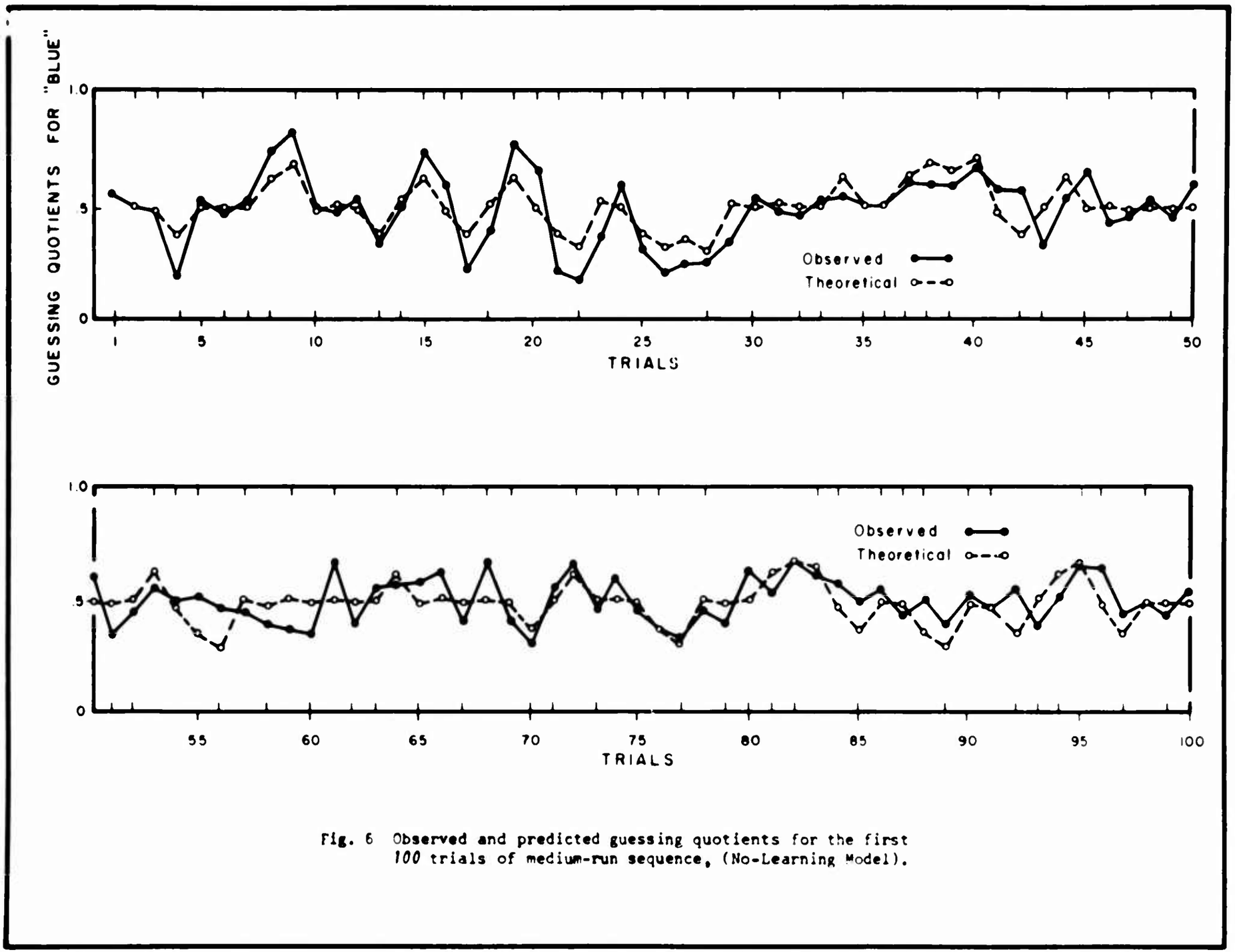


Fig. 6 Observed and predicted guessing quotients for the first *100* trials of medium-run sequence, (No-Learning Model).

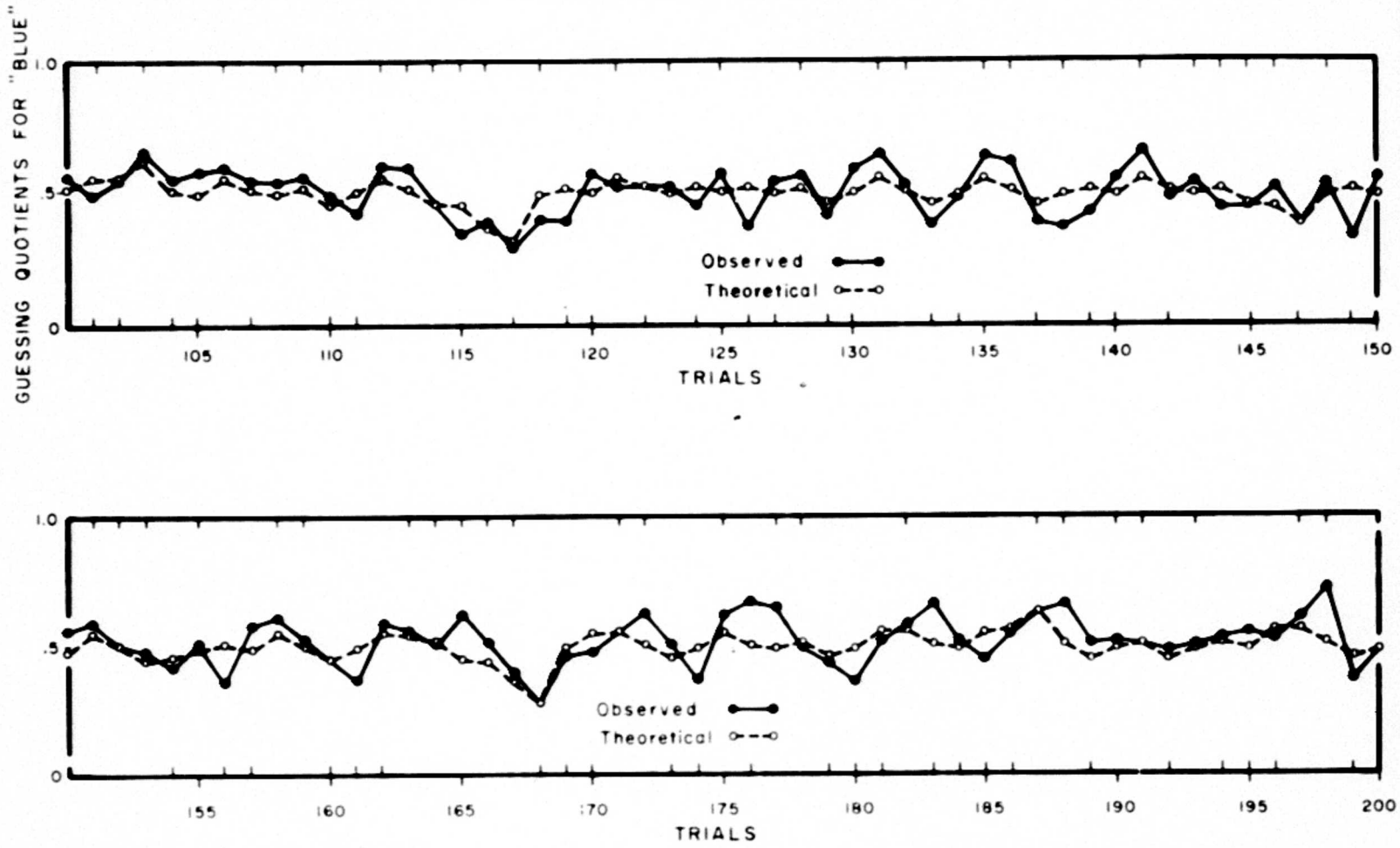


Fig. 7 Observed and predicted guessing quotients for the second 100 trials of medium run sequence, (No-Learning Model).

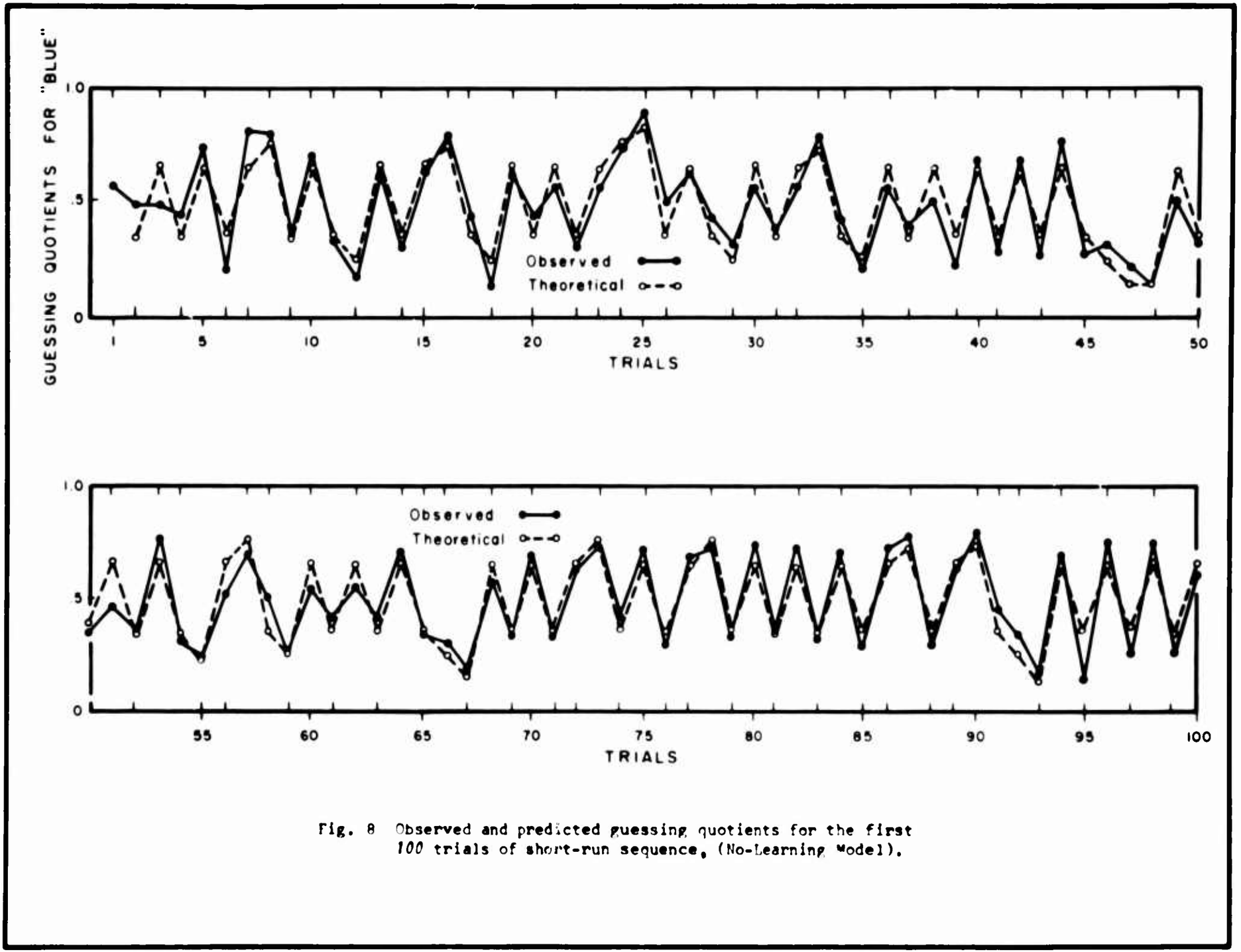


Fig. 8 Observed and predicted guessing quotients for the first 100 trials of short-run sequence, (No-Learning Model).

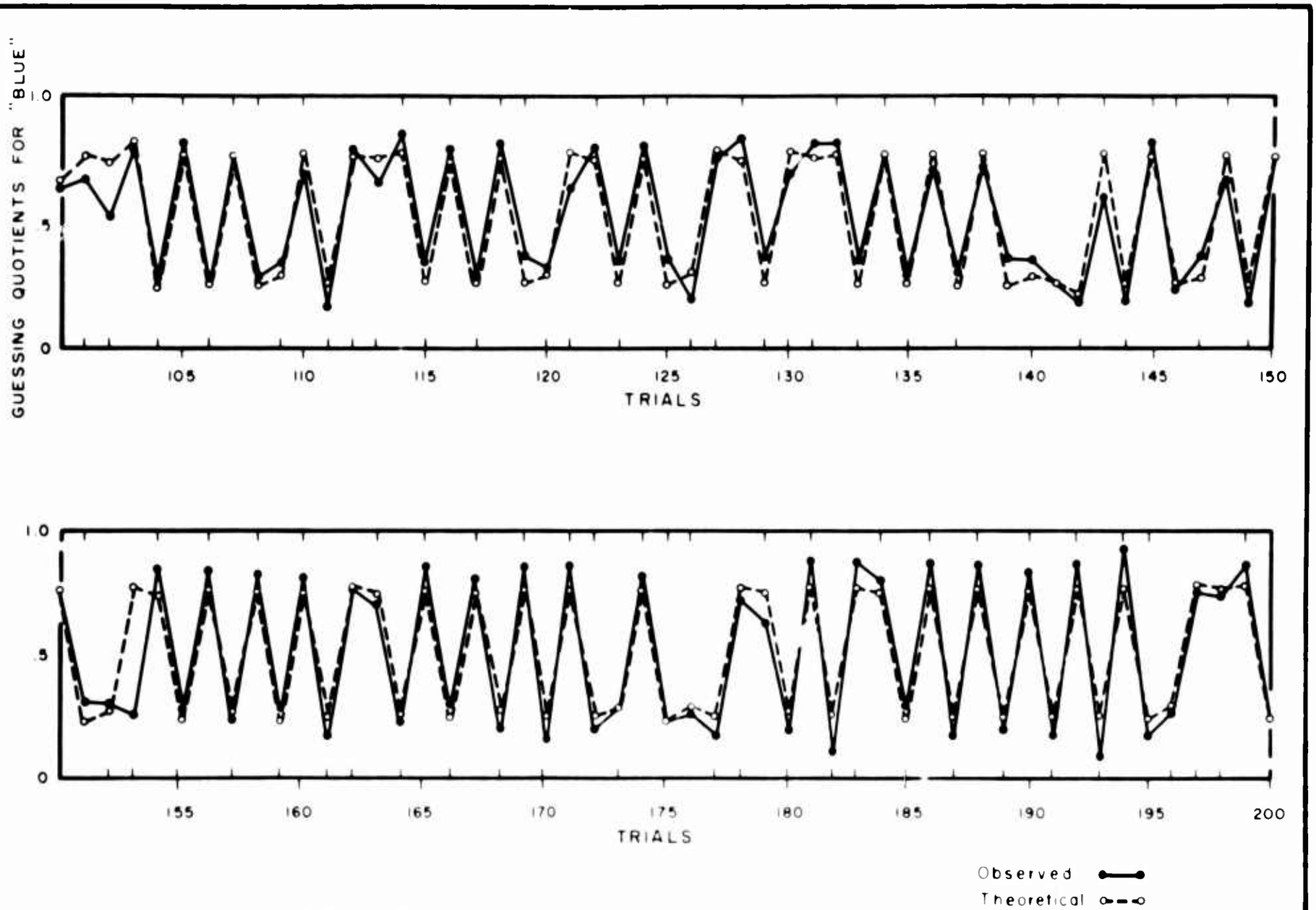
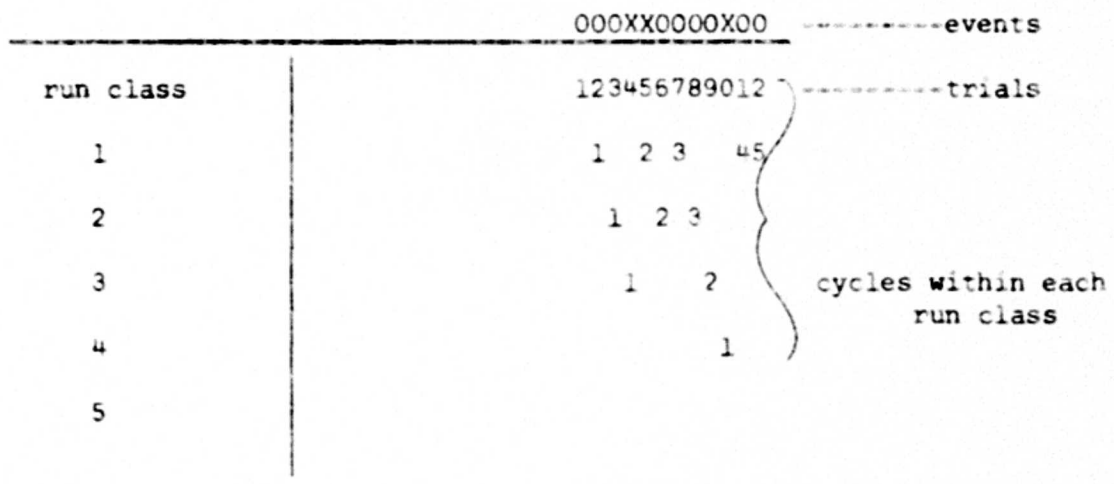


Fig. 9 Observed and predicted guessing quotients for the second 100 trials of short-run sequence, (No-Learning Model).

BLANK PAGE

Now, my first pilot model may be called *no-learning, run-dependent model*. Its basic set of assumptions is as follows: The length of run of the same event which Ss have just observed is the only factor that determines their response probability. This run-length dependent response probability is assumed constant throughout the course of the experiment. That is, there is absolutely no learning. Let me give you an example and define the notation I am going to use. Let the sequence of events be like this:



Any *trial* on which S is in the state of just having observed a run of length n is said to belong to the *run class* n. A serial number is attached to each trial belonging to the same run class in the order of its appearance in the whole sequence, and is called the *cycle number* of the trial within the run class. $P_i(n)$ denotes the proportion of Ss (guessing quotient) who predicted on trial i the *same* event obtained on trial $i-1$, and n indicates that trial i belongs to run class n. In general, $p_i(n)$ is used to denote the theoretical prediction for $P_i(n)$. Now, what the no-learning model amounts to is that $p_i(n) = c_n$ where c_n is a constant for each n independent of trial number i.

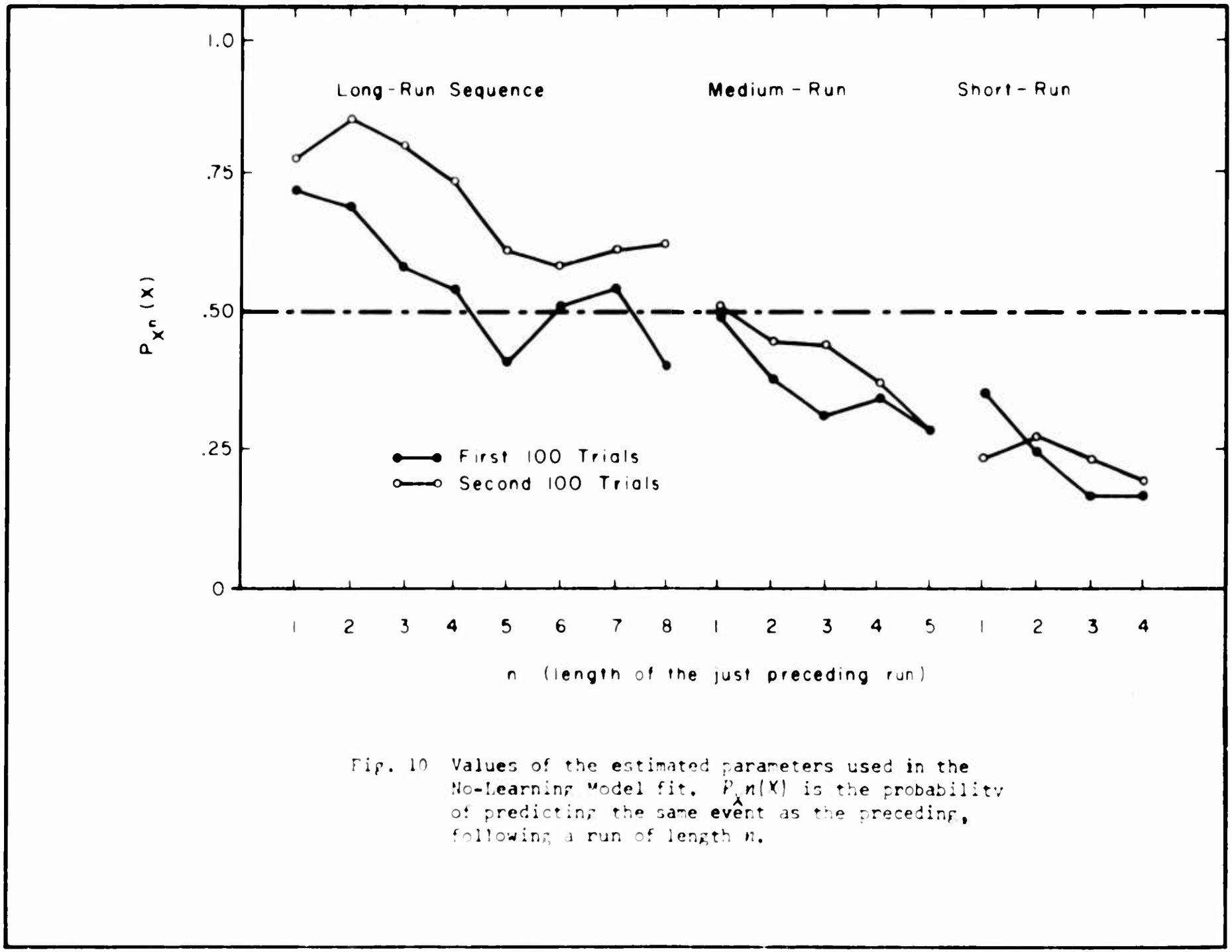


Fig. 10 Values of the estimated parameters used in the No-Learning Model fit. $\hat{P}_{X^n}(X)$ is the probability of predicting the same event as the preceding, following a run of length n.

BLANK PAGE

For the purpose of testing this model I conducted the second experiment in which the long-run sequence and the medium-run sequence are extended to 200 trials, and a short-run sequence characterized by $r(X) = .50$ and $r_X(X) = .25$ is added to them. The results are shown in Figures 4 through 9 along with the corresponding no-learning model predictions.

Obviously, this assumption of no-learning is absurd. So, to keep an eye on its absurdity, I estimated the parameters c_n separately for the first 100 trials and the second 100 trials of each sequence. The estimated parameter values are plotted in Figure 10. From these figures it is all too clear that this absurd model worked very well.

To sum up the conclusions drawn from this pilot analysis: First, Ss differentially respond to different lengths of preceding run. Secondly, there is a learning effect as seen in Figure 10, and this effect is most pronounced in long-run sequence and very little in short-run sequence.

Now let me proceed to my next pilot model, which now contains an element of learning, so that $p_i(n)$ is no longer constant. And this model will be called *Decay Model I* or *time-dependent decay model*. The exact description of this model is given as follows:

DECAY MODEL I: TIME-DEPENDENT DECAY MODEL

$p_i(n)$: probability of predicting the same event as occurred on the just preceding trial. This probability depends upon which run class n the trial i belongs to.

$u_i(n)$: response weight for predicting the same event as occurred on the just preceding trial.

$v_i(n)$: response weight for predicting the opposite event to the one occurred on the just preceding trial.

λ a parameter, $0 < \lambda < 1$

μ a parameter, $0 < \mu$

ν a parameter, $0 < \nu$

$\alpha_i(n) = 1$ if i belongs to run class n and the same event occurs on trial i,

$= 0$ otherwise

$\beta_i(n) = 1$ if i belongs to run class n and the opposite event occurs on trial i,

$= 0$ otherwise

The following system of equations holds for each value of

(1) $$p_i(n) = u_i(n)/w_i(n)$$

(2) $$w_i(n) = u_i(n) + v_i(n)$$

(3) $$u_{i+1}(n) = \lambda u_i(n) + \alpha_i(n)\mu$$

(4) $$v_{i+1}(n) = \lambda v_i(n) + \beta_i(n)\nu$$

As obvious from equations (1) through (4), the trial-by-trial change in response tendency is not directly described in terms of p as it is in the Bush-Mosteller or Estes models, but it is described in terms of response weights u and v, and response probability p is given by normalizing u with respect to the total weight w.

This type of model is often called a non-linear model, but I would rather like to call it a quasi-linear model, since it has many characteristics in common with linear models.

Now let me explain about this Decay Model I characterized by Eqs. (1) through (4). Take a trial i for example. The trial may be preceded by a run of length n so that it will belong to run class n. Suppose that the maximum run length appearing in the total event sequence is m. Then the model assumes that at least m pairs of response weights, u and v, potentially exist, among which only such pair which corresponds to

n, $u(n)$ and $v(n)$, determines p_i, the probability of predicting on trial the same event as occurred on trial $i-1$. Now, suppose that the same event obtained again on the trial i. Then these m pairs of response weights changes on the next trial in such a way that

$$\begin{cases} u(1)_{i+1} = \lambda u(1)_i \\ v(1)_{i+1} = \lambda v(1)_i \end{cases}$$

$$\vdots$$

$$\begin{cases} u(n)_{i+1} = \lambda u(n)_i + \mu \\ v(n)_{i+1} = \lambda v(n)_i \end{cases}$$

$$\begin{cases} u(n+1)_{i+1} = u(n)_i \\ v(n+1)_{i+1} = v(n)_i \end{cases}$$

$$\vdots$$

That is, all the weights except $u(n)$ decrease by constant fraction λ, and $u(n)$ ordinarily increases. This change in response weights is reflected in $p(n)$ in such a way that

$$p_{i+1}(1) = p_i(1)$$
$$p_{i+1}(2) = p_i(2)$$

$$\vdots$$

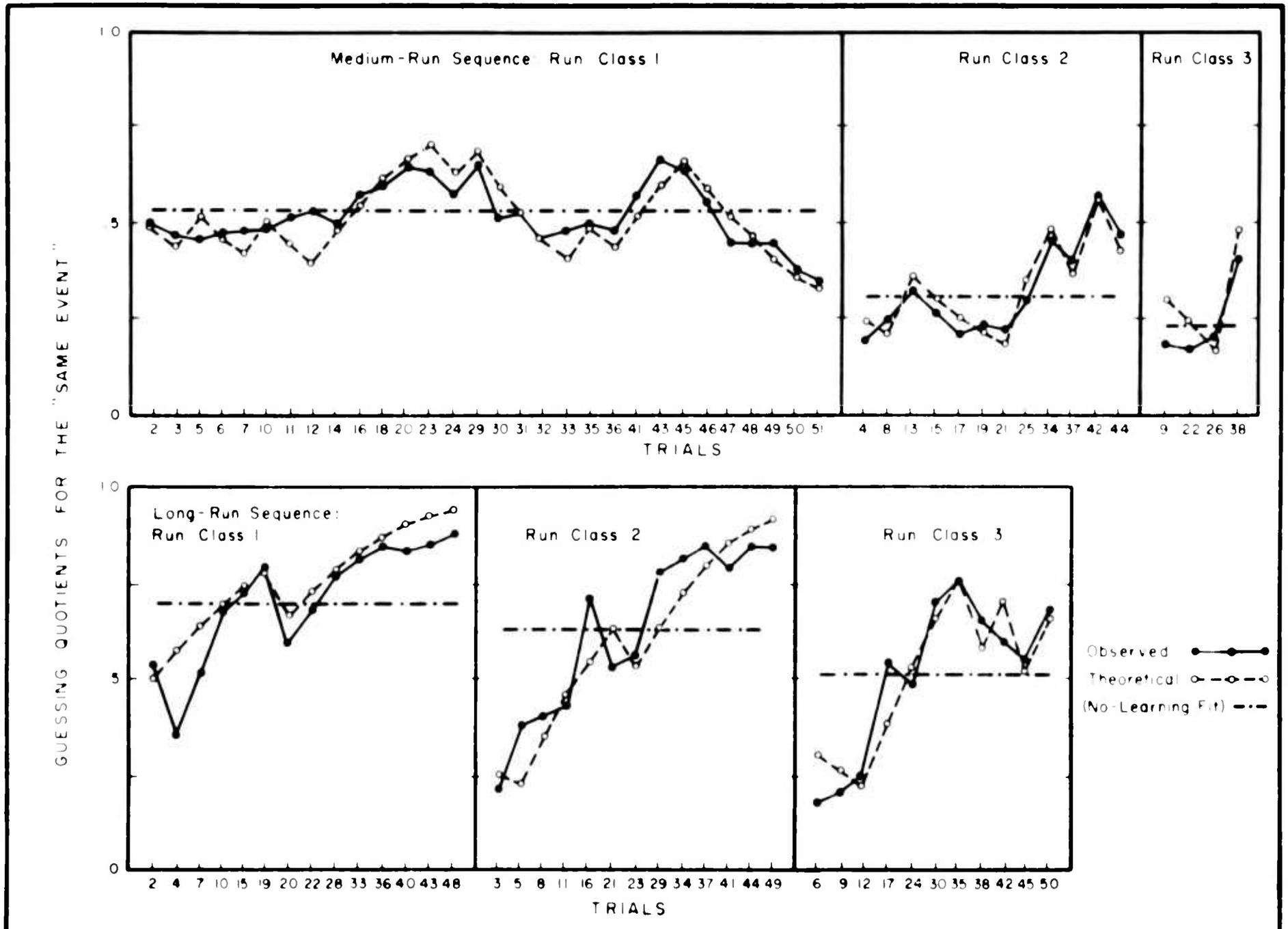


Fig. 11 Approximate fit of Decay Model I to the first 51 trials of the long-run and medium-run sequences.

BLANK PAGE

$$p_{i+1}(n) \neq p_i(n)$$

$$p_{i+1}(n+1) = p_i(n+1)$$

So, all the $p(n)$ except one remain unchanged on each trial. The covert process of constant decay of response weights, however, has the following effect. As the interval between two successive cycles within the same run class tends longer, the impact of additional constant u or v upon the resultant weight becomes greater, and therefore, the accompanying change in $p(n)$ becomes also greater.

Now let me give you just an intuitive interpretation of this model. Suppose that S is classifying information given by each observation of event according to the length of the just preceding run. Suppose that $u_i(n)$ and $v_i(n)$ can be interpreted as the subjectively evaluated amounts of evidences respectively supporting the predictions "same" and "opposite" corresponding to the category n. Then this model means that S is employing a strategy for information book-keeping such that the whole stock of evidences is depreciated by a constant fraction λ each time he proceeds one trial forward. Obviously, this strategy has a certain sense in view of adaptation.

I did not use all the data for the purpose of testing this Decay Model I, but used the first 50 trials of the long run sequence and the first 50 trials of the medium-run sequence. These were the trials on which the data of the first experiment and the second experiment could be pooled. Since this Decay Model I was another pilot model I wanted first to try out the model with the most precise part of the data. For the same reason I did not use precise, but time consuming, method for parameter estimation, but

attempted to find plausible looking parameter values through trial-and-error. The result of the fitting is shown in Figure 11. The parameter values used here as follows:

$\lambda = .926$, $w_i(n) = 10.0$ (for $n = 1, 2$ and 3)

$u_i(1) = 5.0$ $(v_1(1) = 5.0)$, $u_1(2) = 2.5$, $(v_1(2) = 7.5)$

$u_i(3) = 3.0$ $(v_1(3) = 7.0)$, $\mu = 1.4$ and $(v = 1)$

(Those parameter values given in parentheses are derived from others. $v = 1$ is chosen since we can choose the unit of weight arbitrarily.) Now, in Figure 11, horizontal broken lines represent corresponding no-learning model predictions. This no-learning model fit uses six parameters and Decay Model I also uses six independent parameters. As you see, a considerable improvement of data description has been made by moving from the first pilot model to the second.

Having been encouraged by this success, I tried to fit this model to the remaining part of the data. The result was that the fit was by and large worse than in the no-learning model. There may be two possible alternative interpretations of this result. One is that the success of Decay Model I on the first 50 trials is an artifact, and the other is that some change in S's response structure takes place at about 50th trial. I am now inclined to believe the second possibility for a couple of reasons, but I will not go into that issue now.

Because of this partial success of the time-dependent decay model, I wanted to try another type of decay model which may be called "*Cycle dependent Decay Model*," or simply *Decay Model II*. The model is my third pilot model and its formal description is as follows

DECAY MODEL II: CYCLE-DEPENDENT DECAY MODEL

In the Decay Model II, the meaning of subscript i of variables is so changed that it now indicates the cycle number of the run class n instead of serial trial number as it is in Decay Model I.

λ: a parameter, $0 < \lambda$

ε: a parameter, $-1 < \varepsilon < 1$

$\alpha_i(n) = 1$ if the same event as that on the just preceding trial occurs on cycle i of run class n.

$= 0$ otherwise

$\beta_i(n) = 1$ if the opposite event to that on the just preceding trial occurs on cycle i of run class n.

$= 0$ otherwise

The following system of equations holds for each value of n.

$$p_i(n) = u_i(n)/w_i(n) \tag{5}$$

$$w_i(n) = u_i(n) + v_i(n) \tag{6}$$

$$u_{i+1}(n) = \lambda u_i(n) + (1+\varepsilon)\alpha_i(n) \tag{7}$$

$$v_{i+1}(n) = \lambda v_i(n) + (1+\varepsilon)\beta_i(n) \tag{8}$$

Now, the major difference of this new decay model from the first one is that response weights do not decay on each trial, but decay only on each cycle belonging to the same run class. Then this is certainly a simpler model than the first. Another difference is a minor modification of notation

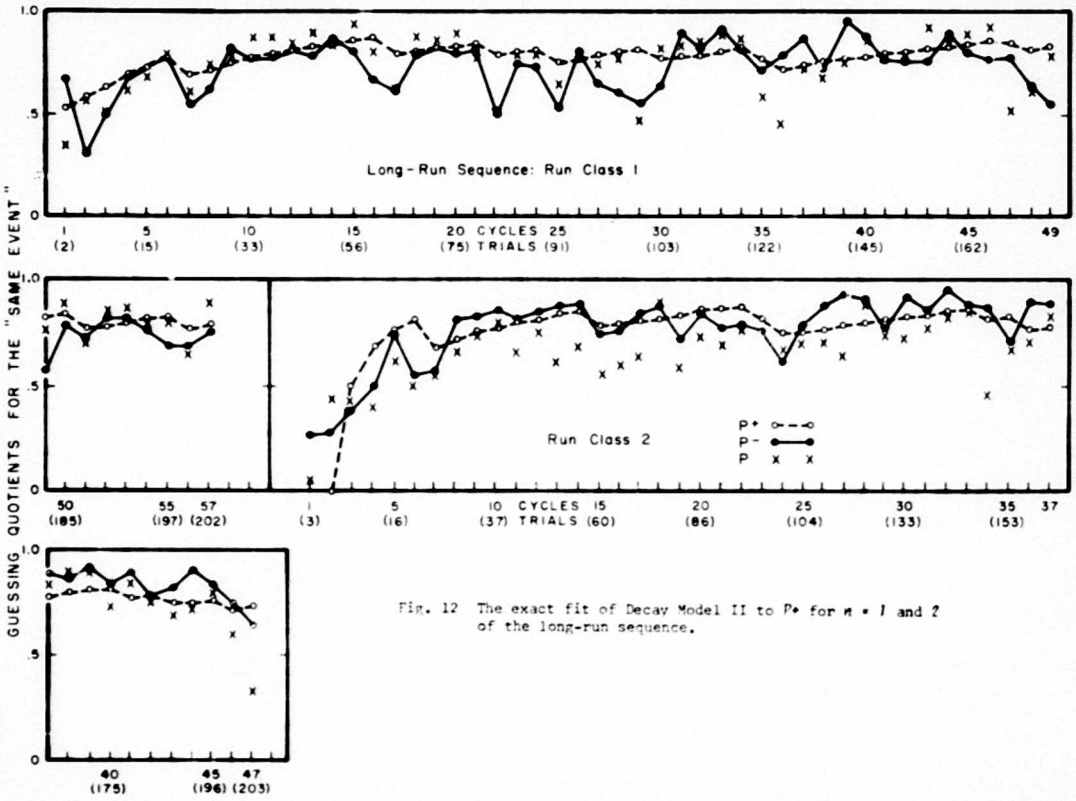


Fig. 12 The exact fit of Decay Model II to P_+ for $n = 1$ and 2 of the long-run sequence.

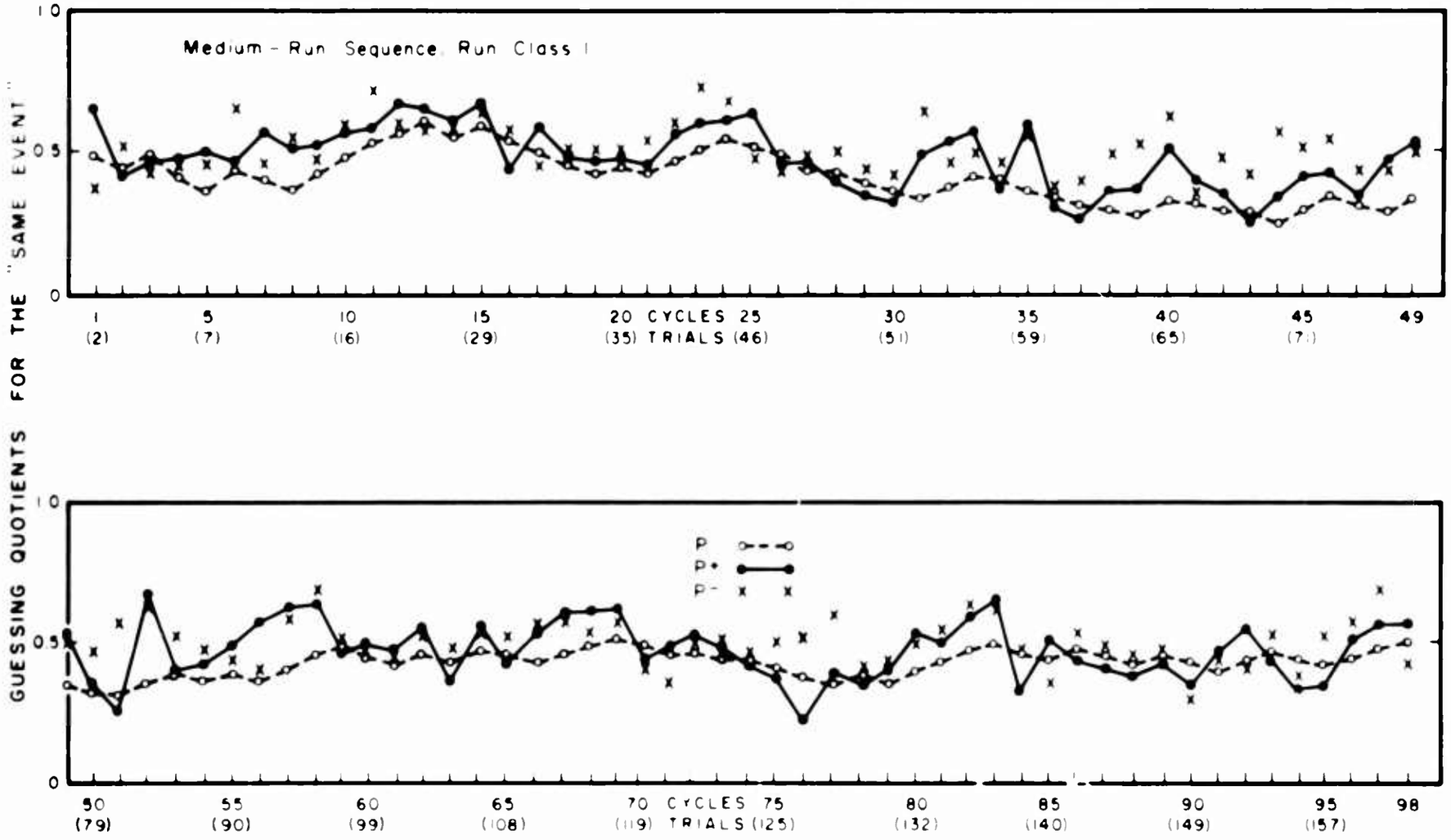


Fig. 13 Exact fit of Decay Model II to P_+ for $n = 1$, medium-run sequence.

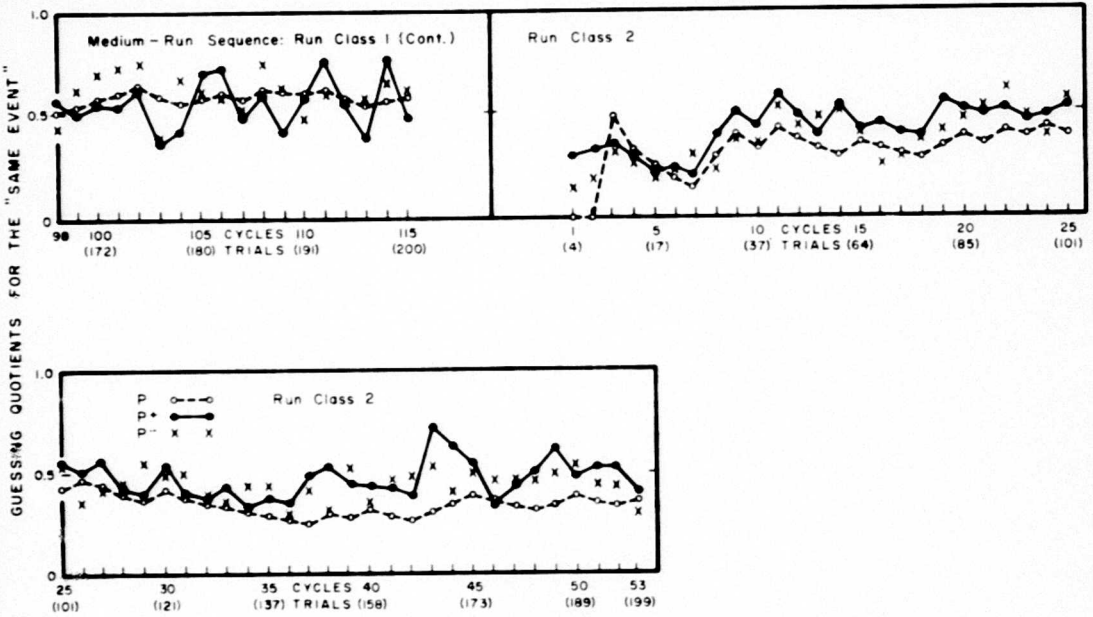


Fig. 14 Exact fit of Decay Model II to $P+$ for $n = 1$ (continued) and 2, medium-run sequence.

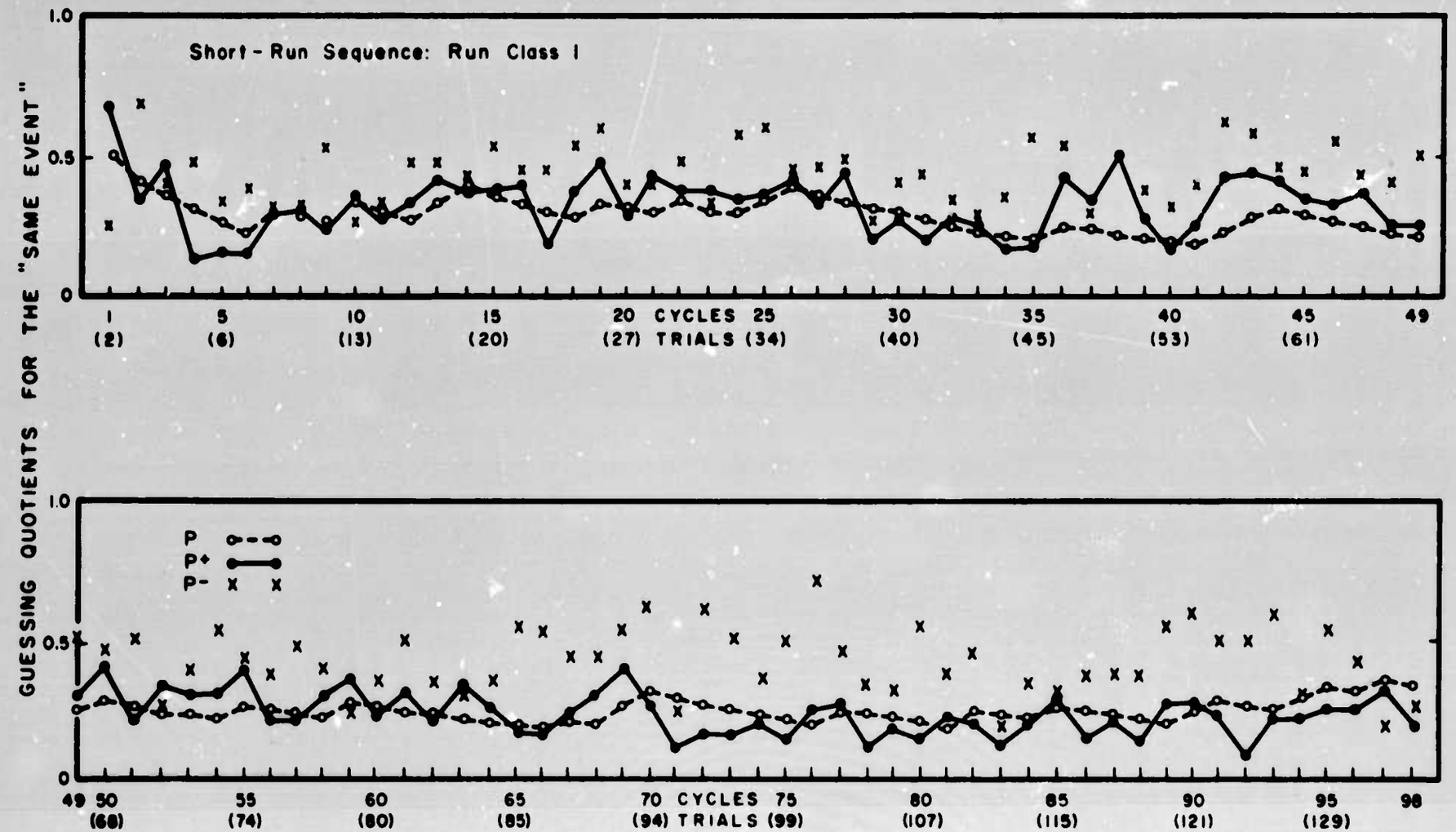


Fig. 15 Exact fit of Decay Model II to P_+ for $n = 1$, short-run sequence.

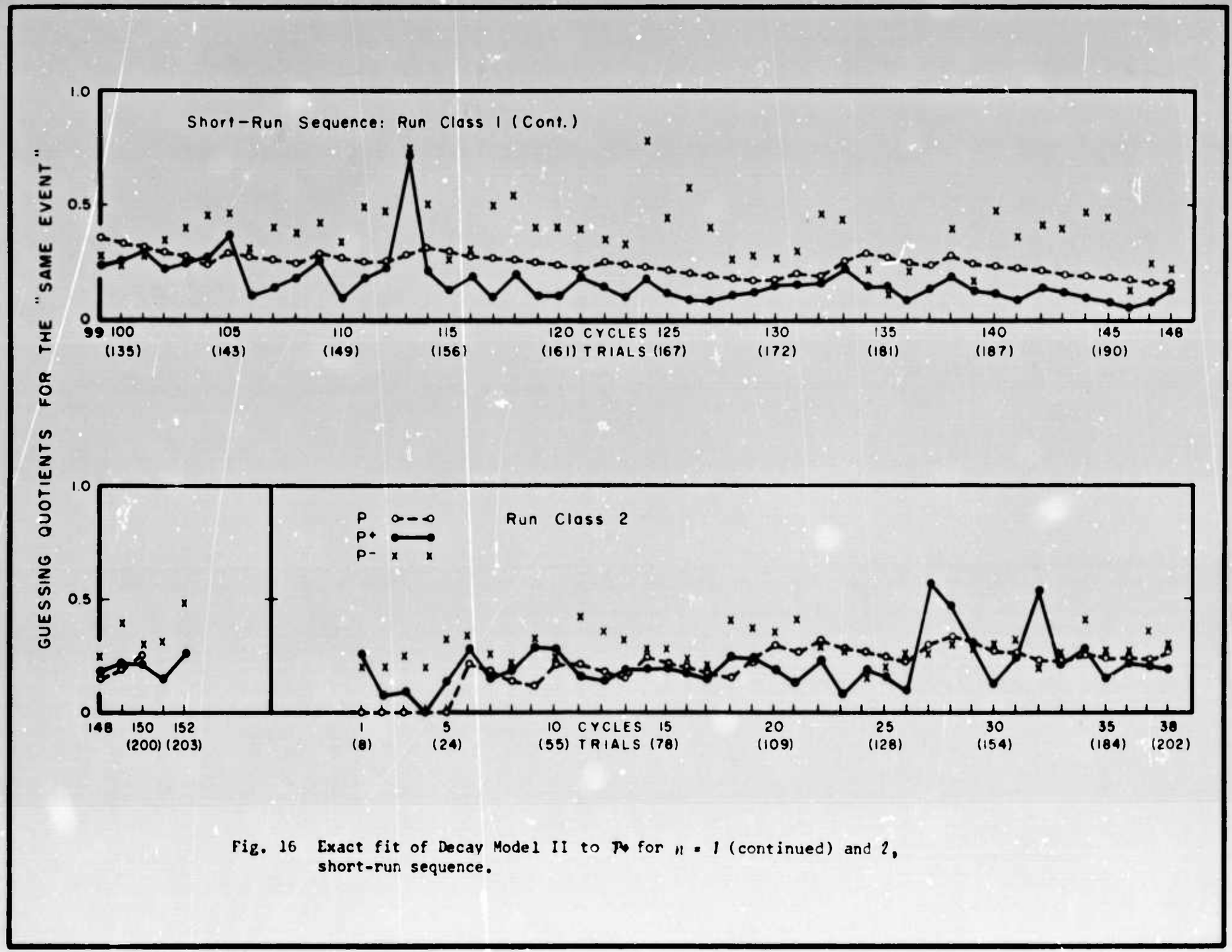


Fig. 16 Exact fit of Decay Model II to $P_\bullet$ for $n = 1$ (continued) and 2, short-run sequence.

BLANK PAGE

$\mu = 1 + \varepsilon$ and $\nu = 1 - \varepsilon$

Before entering into the application of this model, I should mention another type of analysis I did. All the mathematical models so far applied to guess process including my own assume that response probabilities are affected only by physical events and not by Ss' own responses. But it is psychologists' common sense that responses are affected by previous responses too. The existence of this effect in guess processes was first demonstrated by Hake and Hyman (1953).

As a matter of fact, the effect of success and failure on preceding trials upon the response found in my data is really complicated. The nature and amount of effect differs from sequence to sequence and from run class to run class. Furthermore, the effect does not disappear even at the end of 200 trials. And a trouble with linear and quasi-linear models is that they are very rigid about their probability matching property and their descriptive capacity tends to be poorer and poorer as trial proceeds. So after finding this effect, I was again forced to use partial data. I recalculated guessing quotients on each trial for only those Ss whose prediction on the just preceding trial was success, and denoted them as P_+. Analogously, I calculated P_- for those Ss who failed on the just preceding trial. These P_+ and P_- are plotted in Figures 12 through 16 for the three sequences and the corresponding run classes 1 and 2. And I applied the Decay Model II only to P_+, since by and large P_+ is more reliable than P_-. (I am making full use of the excuse that I am dealing with pilot models.) The results of the fit of Decay Model II to P_+ are also

plotted in the same figures. The parameter values used are as follows: $\lambda = 0.942$, $w_1(1) = 5$, $w_1(2) = 0$, $u_1(1) = 2.5$ and $\varepsilon = 0$. ($u_1(2)$ is automatically 0 since $w_1(2) = 0$.) Here $\varepsilon = 0$ actually means that I gave up using ε, and therefore, that ε is dropped from the model. So, the real number of parameters I used to fit the model to the partial data used if four. Taking into account this small number of parameters used I could say that the fit of the model to the data is moderately good, although the fit to the first 50 trials is worse than that of Decay Model I.

Anyway, I think I have definitely demonstrated one thing through the applications of these pilot models; guessing processes are by no means simple. The apparent simplicity of averaged guessing curves is a complete deception. Meanwhile, I still have a hope that someday I will be able to solve this mysterious case.

Now, let me shift to my second theme of the present talk. So far, it has been a detective story. From now on, it will be a speech on police science. That is, I want to talk about the parameter estimation of Decay Model II.

I think the most valuable information I obtained through the course of parameter estimation are not the final outcomes of the parameter estimation but the things I learned through the course of estimation. In books on statistics we find how optimal procedures of parameter estimation are to be carried out, e.g., how one can use the maximum likelihood method or the method of least square. But I can find nowhere what are the next bests when the bests are not practicable. In Bush and Mosteller's book, the authors point out that these best methods can be applied to linear models only in very special cases. If they are impracticable for linear models, how much worse for quasi-linear models. So what I have done first was to

BLANK PAGE

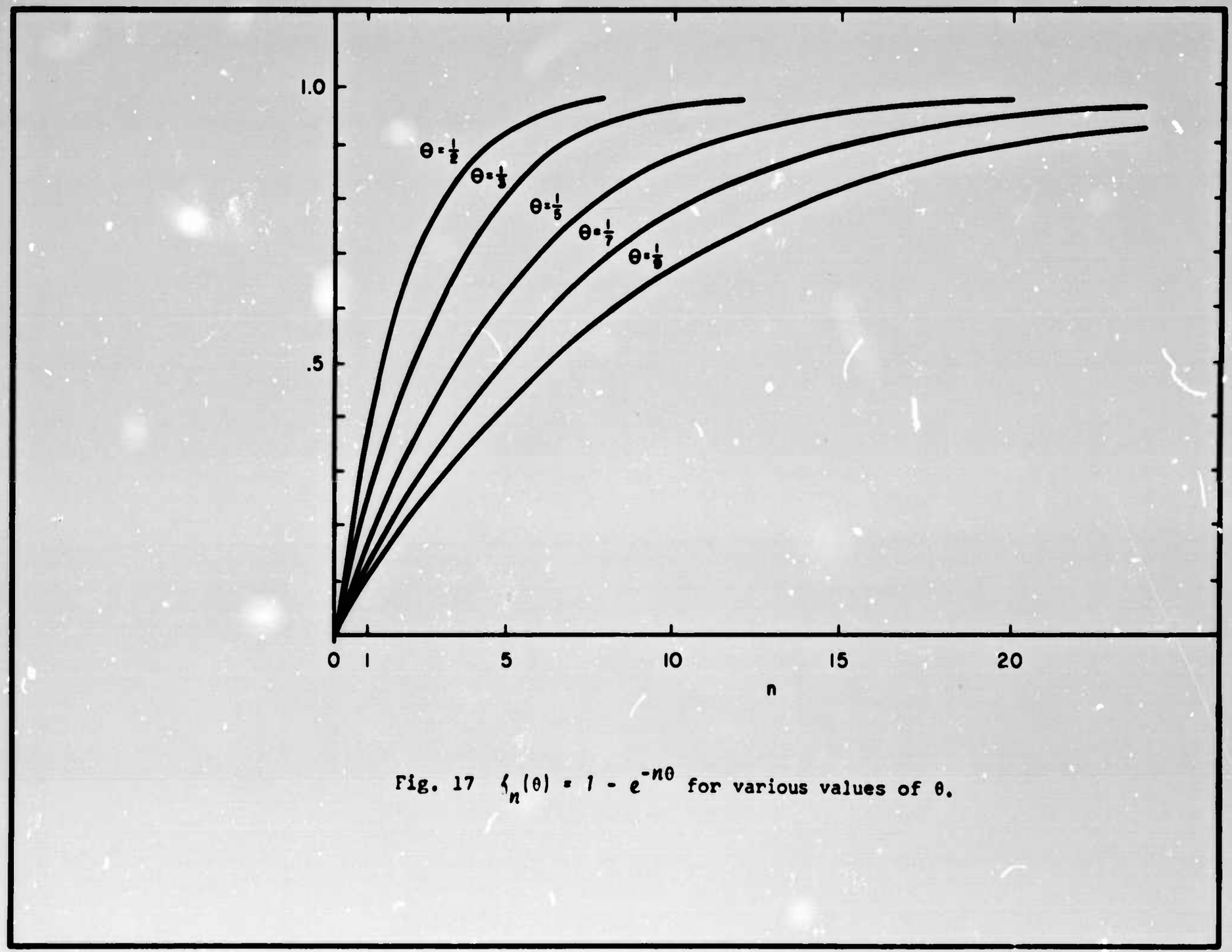


Fig. 17 $\xi_n(\theta) = 1 - e^{-n\theta}$ for various values of θ.

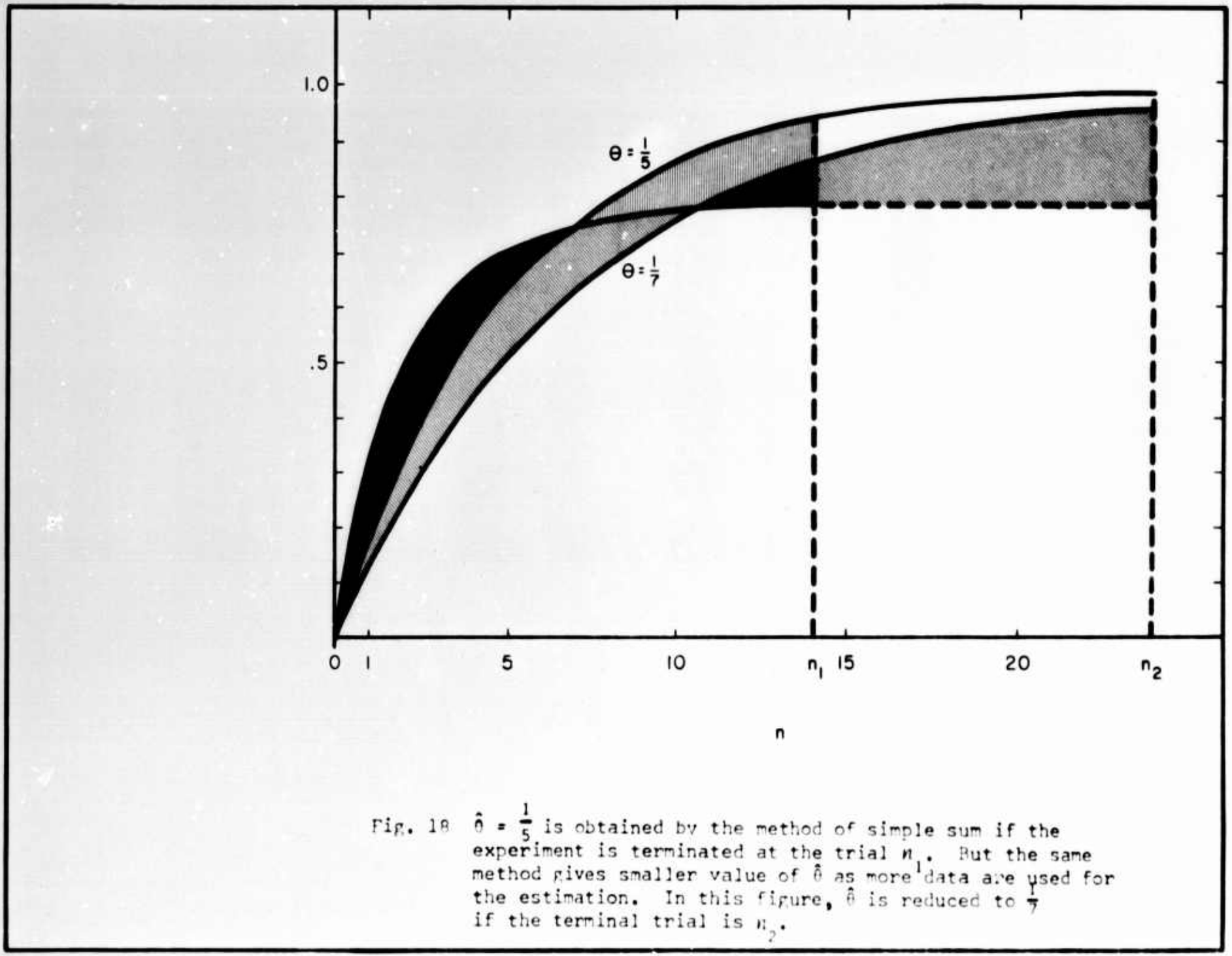


Fig. 18 $\hat{\theta} = \frac{1}{5}$ is obtained by the method of simple sum if the experiment is terminated at the trial n_1. But the same method gives smaller value of $\hat{\theta}$ as more data are used for the estimation. In this figure, $\hat{\theta}$ is reduced to $\frac{1}{7}$ if the terminal trial is n_2.

BLANK PAGE

learn how mathematical psychologists estimate their parameters. What I discovered was awful. I tentatively named one of the most popular methods they used "the method of simple sum," which may be described as follows:

For the sake of simplicity, let us consider a single parameter model for a psychological process. The model gives a sequence of functions $\delta_1(\theta), \delta_2(\theta), \ldots, \delta_i(\theta), \ldots$ where i is the trial number and θ is the parameter. On the other hand, there is a sequence of data values $x_1, x_2, \ldots x_i, \ldots$ Since nobody can hope that the model completely fits the data, we should expect a deviation between model and the data on each trial. Let me denote the deviation δ_i and call it the error on trial i. Then we have a system of equations,

$$x_i - \delta_i(\theta) = \delta_i(\theta), \quad i = 1, 2, \ldots$$

Now if we sum each side of the equations, over all the trial number, we obtain

$$\sum_i x_i - \sum_i \delta_i(\theta) = \sum_i \delta_i(\theta)$$

If we estimate θ by putting the right hand side of this equation zero, we have the method of simple sum. Now, what this method of simple sum really amounts to is to make total positive errors and total negative errors be balanced. And some examples will easily show you how wrong a conclusion one might be led to under certain rather common circumstances. Take, for example such a single parameter model as

$$\delta_n(\theta) = 1 - e^{-n\theta}$$

For various values of θ, we obtain a family of theoretical curves as shown in Figure 17. Now suppose that we obtained data which, although increasing monotonically, has an asymptote less then 1. Then the absurd result we obtain is that, the more the number of trials the experimenter runs the less the estimated θ obtained by the method of simple sum as illustrated in Figure 18. One may easily find this kind of absurd theoretical curve in psychological journals. A more dramatic but more artificial example will be given as follows:

Consider the following model:

$$\delta_n(\theta) = 1 - e^{-n\theta}, \quad n = 1, 2, \ldots, m$$

$$= e^{-(n-m)\theta}, \quad n = m+1, m+2, \ldots, 2m$$

Consider an extreme case such that the obtained data exactly follow the model:

$$x_n = 1 - e^{-n\theta_0}, \quad n = 1, 2, \ldots, m$$

$$= e^{-(n-m)\theta_0}, \quad n = m+1, \ldots, 2m.$$

Any reasonable parameter estimation procedure should give $\hat{\theta} = \theta_0$, where $\hat{\theta}$ is the estimated θ. But the method of simple sum can give no estimate of θ, since any value of θ satisfies the equation $\sum_i \delta_i(\theta) = 0$. Suppose now that we had one more value on trial $2m+1$ which is not equal to

$e^{-(m+1)\theta_0}$ but equal to $e^{-(m+1)\theta'}$, where $\theta_0 \neq \theta'$. Then the method of simple sum gives the estimate $\hat{\theta} = \theta'$. That is, the estimate is determined just by a single irregular value. One may wonder who would use such an obviously absurd method. The fact is that this is one of the most popular parameter estimation methods when the best methods are impracticable.

This method of simple sum appears under various disguises when the number of parameters is more than one. Whether it is used or not can easily be checked, however, by seeing if the equation used to estimate a parameter is equivalent to putting unweighted sum of errors equal to 0.

Now, after making this awful discovery, I attempted to obtain a set of criteria for the admissibility of suboptimal methods of parameters estimation. To do this, I chose the method of least square as the ideal, the closer to it is the better, since the maximum likelihood method is usually further away in its practicability.

As you know the method of least square minimizes the sum of square deviations. Therefore, in our notation, the parameter estimation equation is expressed as

$$\frac{d}{d\theta} \sum_i [x_i - \delta_i(\theta)]^2 = 0$$

Or,

$$\sum_i \delta_i'(\theta)[x_i - \delta_i(\theta)] = \sum_i \delta_i'(\theta)\,\epsilon_i(\theta) = 0$$

Let me call $\delta_i'(\theta)\,\epsilon_i(\theta)$ the *error function* of the method of least square. It seems to me that most methods of parameter estimation have their

characteristic error functions of the form:

$$w_i(\theta)\ \delta_i(\theta)$$

and the estimation of parameter θ is made by putting the function equal to zero. And I think the inherent nature of a method of parameter estimation is best demonstrated by the sequence of weights $w_i(\theta)$. For the method of least square, $w_i(\theta) = \frac{d}{d\theta}\delta_i(\theta)$, and for the method of simple sum, $w_i(\theta) = 1$ for all i.

Now, from the nature of the least square method error function, I derived two criteria for the admissibility of suboptimal methods. The first one corresponds to the absolute value of $\delta_i'(\theta)$; The absolute value of weight of an admissible method should be great for such i where the prediction $\delta_i(\theta)$ is relatively sensitive to the variation of θ, and it should be small where $\delta_i(\theta)$ is relatively insensitive. Now you see that the absurdity of the method of simple sum demonstrated by my first example is due to its failure to comply with this criterion.

The next criterion for the admissibility is concerned with the sign of $\delta_i'(\theta)$; the sign of admissible weight $w_i(\theta)$ should be different when $\delta_i'(\theta)$ is positive and when $\delta_i'(\theta)$ is negative. If a method satisfies this criterion, the kind of absurdity I have shown in the second example could never happen.

With these two criteria, I attempted to obtain admissible method for estimating parameters of my Decay Model II. I picked up λ for the parameter to be estimated first, since $p_i(\lambda, w_i, u_i, \sigma^2)$ becomes almost independent of all the parameters other than λ for large i. Then immediately I found a difficulty. There was no practicable *and admissible*

method for estimating λ. The only practicable method I found may be called the *method of simple ratio*. (You see I hate everything that is simple.) This method is almost as popular as the method of simple sum, and is slightly better than the latter since the former satisfies the second criterion. But it does not satisfy the first criterion of admissibility.

Let me briefly describe this method since I was bound to use it. Error δ_i is defined as before:

$$\delta_i = x_i - p_i$$

According to the definition of Decay Model II,

$$p_{i+1} = u_{i+1}/w_{i+1} = (\lambda u_i + \alpha_i)/(\lambda w_i + 1)$$

Define *simple ratio* k_i in such a way as

$$\begin{cases} k_i = (1 - p_{i+1})/(1 - p_i) = (1 - x_{i+1} + \delta_{i+1})/(1 - x_i + \delta_i) \\ \qquad\qquad \text{if } \alpha_i = 1 \\ k_i = p_{i+1}/p_i = (x_{i+1} - \delta_{i+1})/(x_i - \delta_i), \\ \qquad\qquad \text{if } \alpha_i = 0. \end{cases}$$

Now, by substituting p_i and p_{i+1} in the definition of k_i by the above expression of p_{i+1} and the analogous expression of p_i, we obtain

$$k_i = \lambda w_i/(\lambda w_i + 1)$$

irrespective of the value of α_i.

Now from the definitions of Decay Model II we can easily derive

$$w_i = \lambda^{i+1} w_1 + (1 - \lambda^i)/(1 - \lambda) = 1/(1 - \lambda) + O(\lambda^{i-1})$$

So, by substituting this into k_i, we obtain

$$k_i = \lambda + O(\lambda^i)$$

So, by assuming $k_i = \lambda$ for large i, we obtain

$$(1 - x_{i+1}) - \lambda(1 - x_i) = -\delta_{i+1} + \lambda\delta_i \quad \text{if } \alpha_i = 1$$

$$x_{i+1} - \lambda x_i = \delta_{i+1} - \lambda\delta_i \quad \text{if } \alpha_i = 0$$

We obtain either one of the two types of equations for each large i so that by taking sum of each side of these equations across large i and putting the right hand side, which is the error function of this method, equal to zero, we can estimate λ.

Now, let us take a look at the error function of this method. From the above equations you can easily see that the weights of the error function have the following form:

$$w_i = \pm (1 - \lambda), \quad \text{if } \alpha_{i+1} = \alpha_i$$

$$= \pm (1 + \lambda), \quad \text{if } \alpha_{i+1} \neq \alpha_i$$

So each weight can take only one of the two values, and the difference between these two values is great since the estimated λ is fairly close to 1. So in estimating λ, what this method is actually doing is taking into account only those data on the trials where the theoretical curve inflects. And the worst aspect of this is that those are the trials where, by and large, p_i is most insensitive to the variations of λ.

Anyway, this is what I have done for estimating λ, and the estimated λ's are fairly similar for u and v, for different sequences and for different run classes. However, a slight increasing tendency with trial is observed in the estimated value of λ.

Once λ is estimated, the nest problem is the simultaneous estimation of the remaining three parameters, u_1, w_1 and ε. But since this appeared to me technically impossible, I first dropped ε from the model by assuming $\varepsilon = 0$. Then it is possible, at least in principle, to estimate w_1 and w_1 simultaneously, since they can be separated by utilizing k_i again this time for small values of i. From now on, I will not go into technical details, except a few points of major interest.

By utilizing k_i and again applying the method of simple ratio, we obtain an equation for estimating w_1, of which the error function is characterized by weights of the following form

$$w_i = z(1+\lambda)(1+\lambda^{i+1}z)$$

where

$$z = (1+\lambda^1 w_1)^{-1},$$

Now the point here is that this time we can improve this method to some extent by taking a look at the error function of the least square method. That is, even though the lest square method itself is not applicable, we can modify the error function of a suboptimal method so as to bear more resemblance to the error function of the least square method. As a result of this kind of modification, we obtain

$$w_i = \pm \lambda^{i-1}(1\pm\lambda^2)(1 + \lambda^{i-1}z),$$

and you see the δ_i's on initial trials are more heavily weighted than before.

Now suppose that I obtained an estimate of w_1 by this method, although this is actually a false statement. Then the only remaining parameter u_1 can be estimated, for the first time, directly by the method of least square.

However, since the modified method of simple ratio didn't work, I dropped u_1 too, from the model by assuming it equal to $w_1/2$ for run class 1, and attempted to obtain the least square estimate of w_1 by successive approximation. The method is very simple. The reason why the least square method is usually impracticable is that the weight of the error function, $\delta_i'(\theta)$, is usually a fairly complicated function of θ. But if we replace this unknown θ in $\delta_i'(\theta)$ by its arbitrary estimate θ^*, then the estimation equation

$$\delta_i'(\theta^*)\,\delta_i(\theta) = 0$$

is often solvable. Then, if the estimate θ_1 obtained by solving this approximate equation is considerably different from θ^*, you will replace θ^* by θ_1 and repeat the same procedure, although I think repetition is usually unnecessary since it is easy to get a fairly good estimate θ^* to start out just by a trial-and-error calculation.

Anyway, this method again failed in my case. Any by now the reason for all the failures is clear. The guessing quotients on the first couple of cycles are completely beyond the descriptive framework of Decay Model II. (For a probable reason, see Toda, 1962.) And since all those improved estimation methods give heavy weights to those initial trials where the theoretical values are most sensitive to the variation of w_1, it is no wonder that I ended up with utterly incomprehensible estimates of w_1.

Anyway, these failures led me to an entirely new line of approach. I attempted to use the method of minimum absolute error, that is, to estimate parameters by minimizing the sum of absolute errors, and it turned out that this method is very useful. At any rate, the method of minimum absolute error should at least be as good as the method of least squares, and furthermore, it has a very nice property of disregarding exceptional data values. But this does not mean that this method innocently gives us estimated values no matter how exceptional values may exist in the data. On the contrary, it gives us precise information through the course of estimation about which values are exceptional and in what way they are exceptional. Unfortunately, I have no time to go into details of this method. But I am convinced that this relatively unknown method is worth more attention by the users of stochastic models.

REFERENCES

Bush, R.R., and Mosteller, F. Stochastic Models for Learning. New York: Wiley, 1955.

Estes, W.K. Toward a statistical theory of learning. Psychol. Rev., 1950, 57, 94-107.

Estes, W.K. The statistical approach to learning theory. In S. Koch (Ed.) Psychology: A Study of a Science, V. II. New York: McGraw-Hill, 1959, 383-491.

Hake, H.W., and Hyman, R. Perception of the statistical structure of a random series of binary symbols. J. exp. Psychol., 1953, 45, 64-74.

Mosteller, F. The mystery of the missing corpus. Psychometrika, 1958, 23, 279-289.

Toda, M. Micro-structures of guess processes. Part A. Scientific Report No. CS-2. Cambridge: Harvard Center for Cognitive Studies, 1962.