ED 118 621 TM 005 111

AUTHOR Hyatt, C. J.; DeBerg, O. H.

TITLE Performance Evaluation: The Use of Scoring Systems in Adaptive Training.

PUB DATE [Oct 74]

NOTE 23p.; Paper presented at the Annual Meeting of the Military Testing Association (16th, Oklahoma City, Oklahoma, October 21-25, 1974)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage

DESCRIPTORS *Flight Training; Measurement; *Military Training; *Scoring; *Task Performance; *Tracking

ABSTRACT

Research is described involving the development of a scoring system for performance evaluation. The example used is aircraft landing. Tables are included which give a suggested method for establishing a relevant scoring system in relation to this example. (DEP)

PERFORMANCE EVALUATION

THE USE OF SCORING SYSTEMS IN ADAPTIVE TRAINING

By

Squadron Leader C. J. Hyatt, RAF and Captain O. H. DeBerg, USAF

INTRODUCTION

This paper is intended to describe work which we have done at the Crew Station Design Facility, in the field of scoring complex tracking tasks. It is intended to provide a down-to-earth approach to the development of a philosophy for scoring systems without getting deeply involved in mathematical approaches or analysis or results. The typical sort of tracking task we might look at in the facility is landing an aircraft in Instrument Flying Conditions, using various forms of approach aid, and with various task loadings.

The approach we use to this problem will obviously read across fairly readily to many other areas of study involving the use of motor skills. Our work to date has been entirely in the field of developmental studies rather than training, but since both these processes are amenable to treatment by an adaptive loop approach, they have many features in common, and in particular they both need a meaningful scoring system. Future projects which will require us to develop scoring systems specifically for adaptive training are the Simulator for Air to Air Combat, and the Advanced Simulator for Undergraduate Pilot-Training.

Figure 1 shows how the Crew Station Design Facility fits into the ASD organization. It is part of the Directorate of Crew and AGE Engineering, which is responsible for providing an advisory service to the individual systems program offices of ASD. You will see that we are in the same division as the simulator branch, and we also have simulators of our own, used almost exclusively for experimental work.

As the emphasis on simulators in training increases, and these simulators become more sophisticated, the need for sophistication in our scoring systems increases, and over the last year we have found it necessary to educate our 'customers'. This paper is a distillation of our studies.

GENERAL DISCUSSION OF PHILOSOPHY

I want to start by looking at some definitions; working definitions rather than academic ones. The first of these is what we mean by an adaptive process. Figure 2 shows how we have defined it for our purposes in very general terms.

Figure 3 shows how this process can be represented by a classic three block loop, and I think most workers in this field would accept this general concept. More specifically in the context of training, the three blocks can be more precisely defined as shown in the lower half of the figure. There already appears to be a fair consensus of opinion that the quality of the performance measure is "the make-or break" feature of the system. Most people would agree that it is unfortunately also heavily subjective. To make the performance measure as good as possible let us first look at what we want from a scoring system. We have made a list (Figure 4) of qualities which appear to be important:

The first and most important is that the scoring system should be directly related to the objectives of the adaptive process. To ensure this, we have to persuade the training or experimental director to define his objectives clearly.

The remaining qualities are not in any particular order: If the objectives are multiple, each aspect must have some element of the scoring system directed toward it.

Many parameters may be collected during a study, and the mass of data can be confusing. It is necessary to reduce this to manageable and comprehensible proportions.

Although some subjective decisions must be made in the formulation of the scoring system, the user should not have to make any when applying it-ideally a computer should be able to handle it.

Having achieved a numerical value for the score, it should be possible to relate the various ranges with an acceptability index, for example: Excellent, Good, Average, Poor, Unacceptable.

For the greatest benefit to be obtained from the adaptive system, knowledge of results is a valuable aid, and hence it is desirable if possible for the scores to be generated in real time.

To fully apply these principles we feel that it is necessary to develop the performance measure block of the classic adaptive loop (Figure 5). The classic adaptive training loop calls for the application of the philosophy of training in the formulation of the adaptive logic. This means simply how the next step of training is governed by present performance. Because this is the only block labelled 'Logic' there is a tendency to try and use this step to insert logic related to the training objectives. We believe that the performance measure block should be developed into three separate stages as shown in the lower part of the figure, and that the value logic and adaptive logic should be kept entirely separate from each other. It is these three steps that we want to concentrate your attention on. In simple and direct terms these three steps can be relabelled as shown in Figure 6. Our aim is to develop the philosophy which should be applied to the generation of the value or scoring logic step in this process.

DEVELOPMENT OF SCORING LOGIC

Having defined the concept of 'scoring logic', it is easy to see that this logic is the primary link between the raw data which can be obtained by monitoring performance, and the score which is used to determine subsequent progress.

It is the sole point in the loop at which the recorded performance is weighed against the fundamental objectives of the process. As a result, the quality of the scoring logic is the key factor in determining whether or not the adaptive cycle is efficiently directed toward the aims of the process, be it an experimental study or a training program.

It is only too easy to skip over the question of objectives when defining the logic, and there is, perhaps, an even more insidious risk of using existing scoring logic 'because it worked well last time'. Scoring systems tend to look similar to each other, especially those used in any one particular field of endeavor, and subtle but vital differences can go un-noticed.

One way of minimizing this risk - which we believe is a worthwhile investment of time - is to carry out an objective evaluation of the true aims of every new scoring system we devise, and to develop a sound rationale for the scoring logic. Better still, this rationale should be formally written up and included as an integral part of our description of the scoring system.

Let us then look at the various ways in which the aims of the adaptive process we are considering can affect the way we go about scoring it. There are a number of questions we have to ask ourselves, and some of the major ones are outlined in Figure 7.

The first - and in our developmental studies the most fundamental - question, and yet oddly enough the one most frequently overlooked, is which part of the man-machine system are we looking at: the man, or the machine, or the interface between them.

In the training context the answer is simple - the man, that is to to say the trainee, is paramount. I would like to digress for a moment, though, and speculate on the wealth of data which has, at one time and another, been collected within adaptive training systems and which if it is still stored, could have potential value for the study of the training machines used, or of the way they display information. It seems probable that much of this data might be tapped simply by running it through new scoring systems with appropriate changes in their aims - assuming that the original scoring system was correctly aimed at training.

Much valuable information on the merits or shortcomings of various monitoring or operating consoles currently in use might be acquired in this way.

However, as I said, this is a digression, and in the training context it is the trainee we are trying to assess.

The next question we have to ask is what are we trying to find out about the trainee. The basic answer here is obvious: we want to know how well he performs. But to decide how to measure this performance, we need to ask a number of subsidiary questions.

For example, we have to decide which of the parameters available to us - which means those parameters we can measure without undue expense - are relevant to performance. And the performance we use as a yardstick here must itself be performance which is directly relevant to the training objectives.

A good example of the sort of decision we have to make in this area arises when scoring a tracking task such as an Instrument Landing Approach. Should we consider distance along track as a measure of performance? This distance is one way of looking at the trainee's control of his speed; not just his instantaneous speed, but the integral of his speed with respect to time from the start to the present. Thus if he makes an error in speed, to minimize his resulting penalty score he must not only correct the speed, but make a suitable compensatory adjustment to bring him back to his correct position along track. The argument against using this parameter for scoring is that as long as the trainee stays on the correct line through space as defined by the landing system, it does not matter when he gets to various points along it. However, this obviously depends on the scenario. If the object of the exercise is not only to fly along the correct line and land at the correct point, but also to land at a specific time to fit in with an existing traffic pattern then the accuracy of his position along track must be considered of some importance.

Another decision we have to make, which is also related to what we are trying to find out about the trainee's performance, is whether we should be concerned with his continuous operating or tracking ability, or only with his ability to reach a certain point by any means at his disposal. This quite clearly will determine whether we want a continuous scoring system or what we term 'gate' scoring - that is to say a measure of his ability to pass through a gate in space, or perhaps a series of gates.

I realize that you may be thinking that the points we are making here are overly simple and obvious - they are simple, and they should be obvious, but we think it is vitally important to emphasize that a cold-blooded analysis of this sort should be made, rather than the sort of approach which we know from experience often does occur, based on the principle of "Well, we usually score parameters A, B, and C - it looks as if they should be OK again this time."

Having, we hope, established which parameters we want to select from the raw data, we next have to think how the values of these parameters can best be used to give us a measure of performance. The logical approach is to compare these achieved values with some ideal. Most scoring systems adopt this approach, but once again a vital step which may be missed is to ensure that the ideal we set is truly relevant to our objectives. It is useless to set an ideal value of some parameter at 100 feet plus or minus zero, when we know that the 'Ace of the base' can only achieve 100 feet plus or minus ten feet, and plus or minus twenty feet is quite adequate for routine performance of the task. This is also a useful point at which to consider the limitations of our measured data; we can run into serious problems if we try to score to the nearest foot, when the equipment - and I include here both the instructor's monitoring equipment and the trainee's operating equipment - can only measure to the nearest five feet.

Even having established an appropriate ideal and a valid measure of divergence from it, we still need to consider what the implications of this divergence are.

First of all we must look at the relationship between size of divergence and importance - for example in a certain situation a five foot error could be acceptable, a ten foot error merely embarrassing, but a fifteen foot error fatal. Clearly this is not a linear relationship - at least, not in terms of human values, and our score should reflect this. In an extreme case of this sort of situation, any error less than ten feet might be totally acceptable, while anything in excess of ten feet would be totally unacceptable.

This is an example of what is generally known as 'time on target' scoring. In some contexts such as air to air combat with guided missiles it may be a perfectly adequate measure of performance, although even here it is more suited to competitive scoring than to training. But for a complex tracking task it contains too little data to help the instructor or trainee to identify areas of weakness and plan remedial training accordingly.

Another thing we have to take into account when assessing the divergence of the achieved performance from the ideal is the fact that some parameters may be much more critical than others, and if we want a scoring system which assigns equal penalties to equally unacceptable errors, we must weight the measured errors accordingly.

A FORMAL PROCESS FOR CREATING A SCORING SYSTEM

I now want to recapitulate on the various sorts of decision we have discussed, and in so doing I want to formalize and develop a process for creating an effective scoring system.

I want to show how, starting from the raw data available to us, we can adopt a systematic approach to mould it to our purposes, and ensure that our training objectives are met.

There are various processes we can apply to a mass of raw data, but we believe that six of these processes, shown in Figure 8, if applied in sequence will go far toward producing a sound system.

The first step is to select the parameters which are relevant to our training objectives.

Next we must look at what data is available to us on these parameters and edit it. By editing it, I mean deciding which values of it to use, and the choice here ranges from using it all, through using it at regular intervals, to using it only at specific points which we consider relevant.

We must then compare this edited data with what we believe to be the ideal values we are seeking to achieve by out training program. This comparison will give us what we have chosen to term 'error values'.

These error values must now have two processes applied to them. These are Modification and Weighting. The dividing line between them is not clear cut, and for this reason - to avoid lengthy discussion of what each comprises, I will not separate them. The operations which I include under these headings are:

Ensuring that the error values reflect the trainee's performance rather than any shortcomings in the training equipment.

Ensuring that the size of the error and the importance of this size is suitably reflected in the score.

Ensuring that the more critical parameters carry appropriately heavier scoring penalties.

Ensuring that any inter-relation between parameters is accounted for - this for example would include any weighting in respect of range if this were considered relevant.

The result of modifying and weighting the error values is to produce what we call 'scoring elements'.

Finally the scoring elements we have arrived at can be combined to give a single comprehensive score, or they may be combined in groups to give sub scores related to particular parts of the training objective, or particular capabilities of the trainee. Part of this combination process will be normalization of the score with respect to time or distance if this appears to be appropriate.

We believe that this systematic approach gives the best chance of achieving a useful and meaningful scoring system.

PRACTICAL EXAMPLES

If our systematic approach is applied to a variety of training programs, a variety of scoring systems will naturally result, but in all probability they will fall into four or five broadly defined groups; for example time on target systems, cumulative error systems, gate score systems, or combinations of these. Within any of these groups, one can postulate a generalized scoring system which by manipulation of a series of constants can be used for various slightly different training tasks. This concept lends itself very well to a computer based scoring system within an evolving organization - for example a pilot training school - where the training objectives remain fairly constant but the equipment used, and the associated operating procedures, will probably evolve steadily over the years.

To give you a good example of this, I will offer a brief outline of the type of scoring system we are currently using in our developmental studies to assess a pilot's ability to carry out landing approaches under instrument flying conditions. Remember of course that in these studies our aim is to assess the man-machine interface: This does not affect the process of scoring, but only the detailed application of the scoring logic.

We have concluded that the appropriate method to adopt in this instance is continuous scoring, with the score a function of size of error. Time on target scoring simply does not tell us enough about the things we need to know. However we do also take several specific gate scores at appropriate points en route, depending on the nature of the particular study. Our terms are defined in Figure 9, and a generalized formula for this continuous scoring is shown at Figure 10.

You will see that this formula allows for easy adaptation to suit changes in the relative importance of the different errors by adjusting the constants K, changes in the relationship between magnitude and importance of errors by adjusting the functions of the error values, changes in normalization philosophy by adjusting the function of time, and changes in weighting for range by adjusting the range function.

To illustrate how this type of formula has been applied, I will use the Microwave Landing System as an example. The object of this exercise is to ascertain whether or not certain tracks in airspace can be flown, on a time schedule, in a safe and efficient manner. Different routes are to be flown, and the object of the scoring system is to indicate the relative desirability of each route. Some typical routes are shown in Figure 11.

An MLS route differs from the standard ILS approach in several ways. First it is time dependent. The pilot must be in the right place at the right time. Also he must fly several different headings, some on a command basis, some by dead reckoning. Finally the descent rate is not constant, but depends upon which portion of the approach is being flown.

Scoring these approaches depends upon our value logic (as is true in any scoring system). This logic is based upon the objectives defined for the study. As initially set forth the purpose of MLS approaches is to control aircraft throughout a given airspace both with respect to time and spatial orientation relative to the prescribed path. This leads to the following logic:

(1) The aircraft must be equally controlled throughout the airspace.

(2) The aircraft must be in the "right place at the right time."

(3) For safety considerations the further the aircraft is from track the closer it approaches a critical situation.

(4) Some types of error are more critical than others (i.e. low on altitude is worse than high on altitude).

(5) Different tracks must be compared to one another on the same basis.

Apply these criteria, all possible parameters are analyzed and either discarded or modified and included in the score as considered appropriate. The formula shown in Figure 12 is the result. No range weighting is included because he must fly just as accurately at great distances as he does in close. Time of arrival considerations are met by using along track error which is time dependent. Hence the error in this direction is taken to be the distance of his actual position from where he should be. The safety consideration states that big errors can be critical, hence a square law is used. This penalizes large errors very heavily. The weighting constants were arrived at subjectively by discussion with qualified personnel as to the relative importance of different types of errors to be considered in a safe approach. Finally, in order to compare different tracks the entire score is normalized with respect to time. The resulting equation can be taken as a whole or by parts to examine the quality of the approach.

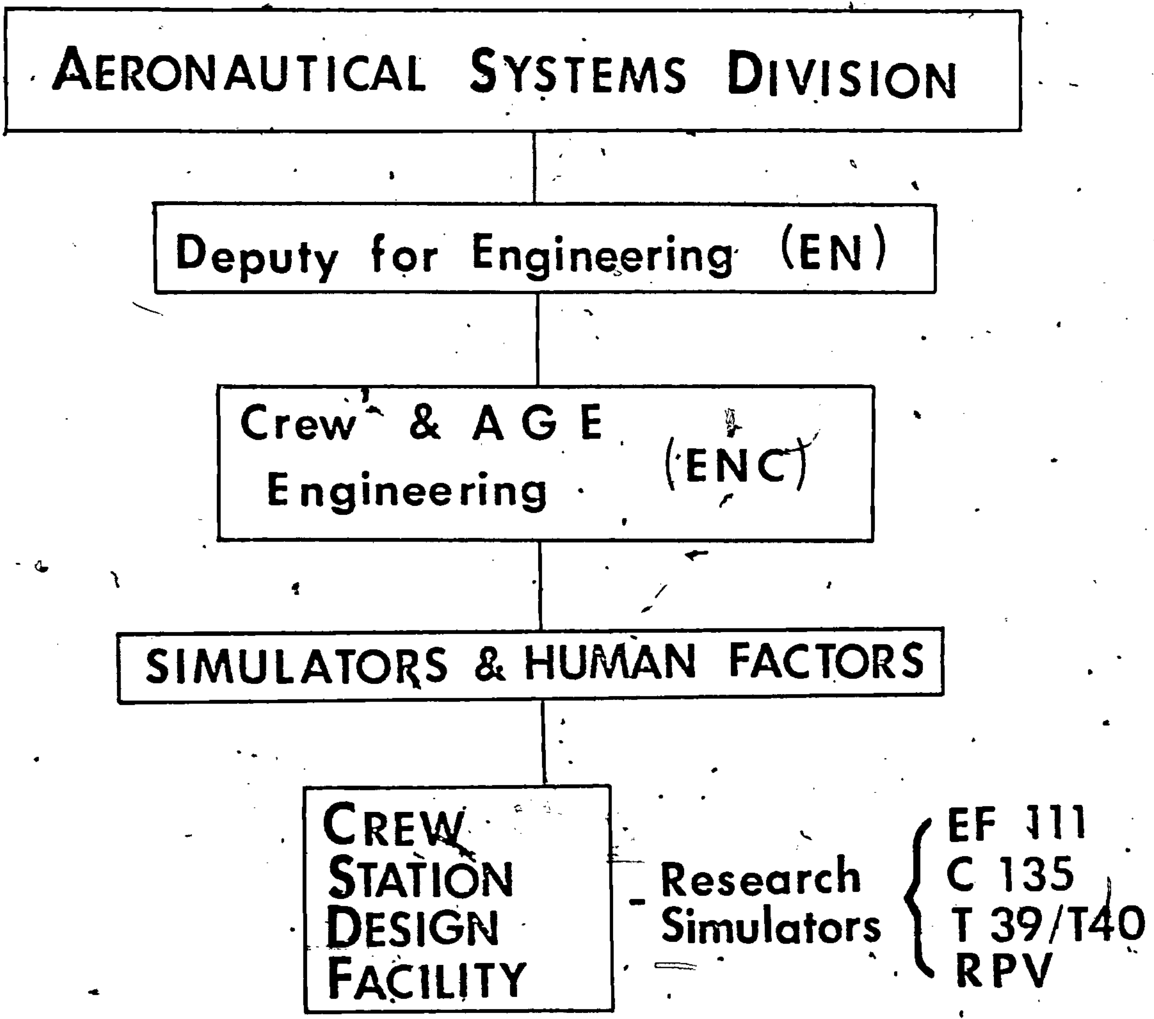
CONCLUSION

To conclude, we do not claim to have found all the answers, but we do feel we have gone a long way toward asking the right questions.

We have described a learning process that we have gone through, and which we suspect many people go through in the process of producing scoring systems. It has been a salutory experience for us to formulate our ideas into a systematic process. We hope that our presentation today will stimulate interest in the process and perhaps save others some of the time we have spent.

Remember: To produce an effective scoring system three things are essential. Good raw data must be collected, good value logic, or scoring logic, must be applied to it. The resulting components must be assembled in a practical manner. Figure 13 shows how our proposal sequence of operations generates the first three steps of the five step adaptive loop. The common thread running through the whole process is relevance - we must constantly ask ourselves if our scores relate to our aims.

We also hope that we have given you some food for thought, and that some of you will feel like contributing your own ideas in discussion now. Perhaps we may discover a few more of the answers to the questions we have posed. Thank you.



AERONAUTICAL SYSTEMS DIVISION

Deputy for Engineering (EN)

Crew & A G E Engineering (ENC)

SIMULATORS & HUMAN FACTORS

CREW STATION DESIGN FACILITY

- Research Simulators

EF 111
C 135
T 39/T40
RPV

FIG 1

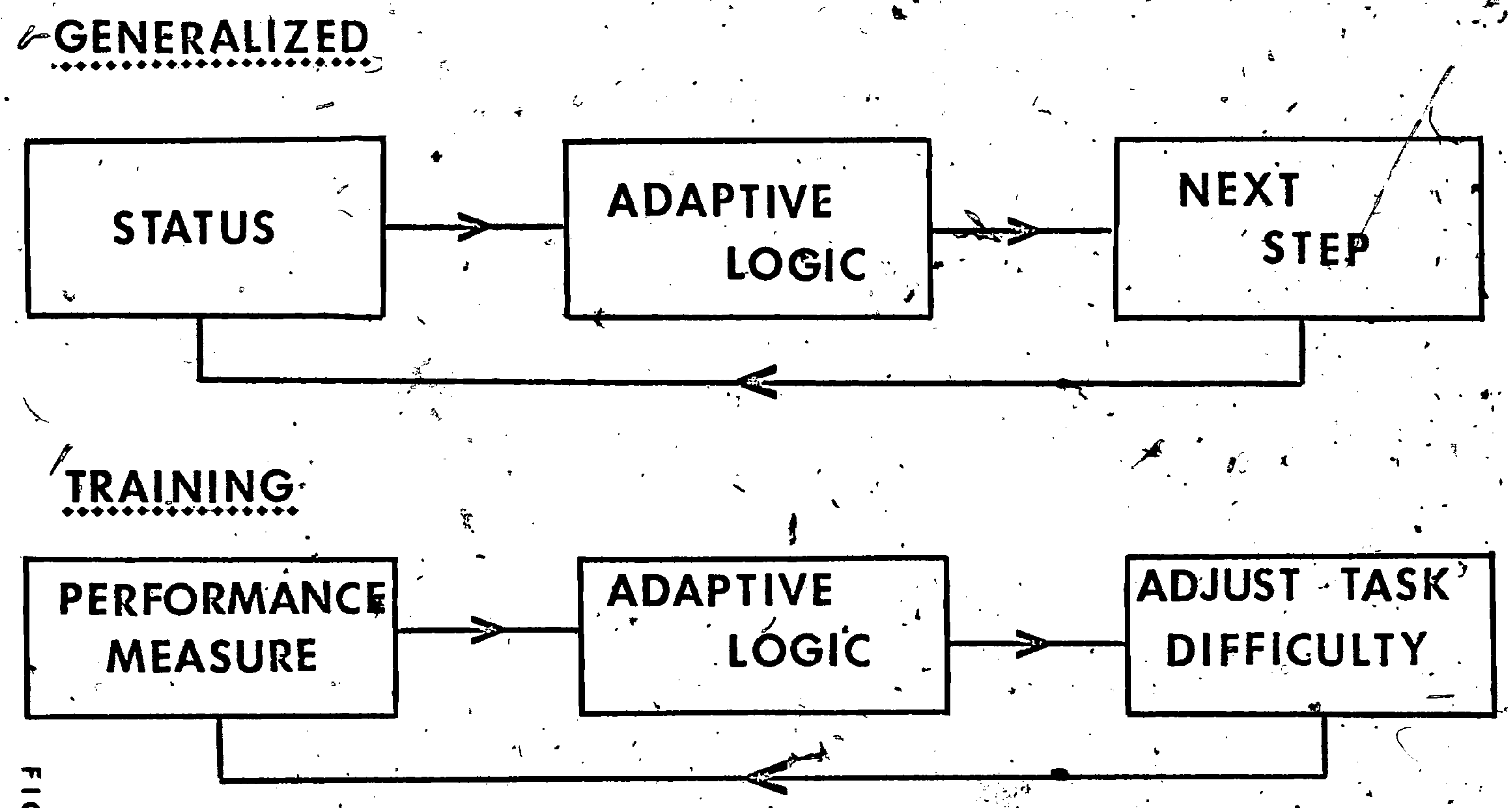
THE ADAPTIVE PROCESS

A PROCESS IN WHICH THE OUTCOME OF ANY STEP DETERMINES THE NATURE OF THE NEXT STEP.

THE RELATIONSHIP BETWEEN ONE STEP AND THE NEXT IS GOVERNED BY THE APPLICATION OF A PREDETERMINED LOGIC.

FIG 2

THE ADAPTIVE LOOP



GENERALIZED

STATUS

ADAPTIVE LOGIC

NEXT STEP

TRAINING

PERFORMANCE MEASURE

ADAPTIVE LOGIC

ADJUST TASK DIFFICULTY

FIG 3

AN EFFECTIVE SCORING SYSTEM SHOULD:

★ BE DIRECTLY RELATED TO OBJECTIVES

★ DEAL WITH ALL ASPECTS OF OBJECTIVES

★ CONDENSE RAW DATA INTO SIMPLE MEASURES OF CRITICAL SYSTEMS ACTIONS

★ NOT BE DEPENDENT ON SUBJECTIVE MEASURES

★ RELATE TO AN ACCEPTABILITY INDEX

★ PRODUCE RESULTS QUICKLY, PREFERABLY IN REAL TIME

FIG 4

THE ADAPTIVE TRAINING LOOP

Development of Performance Measurement:

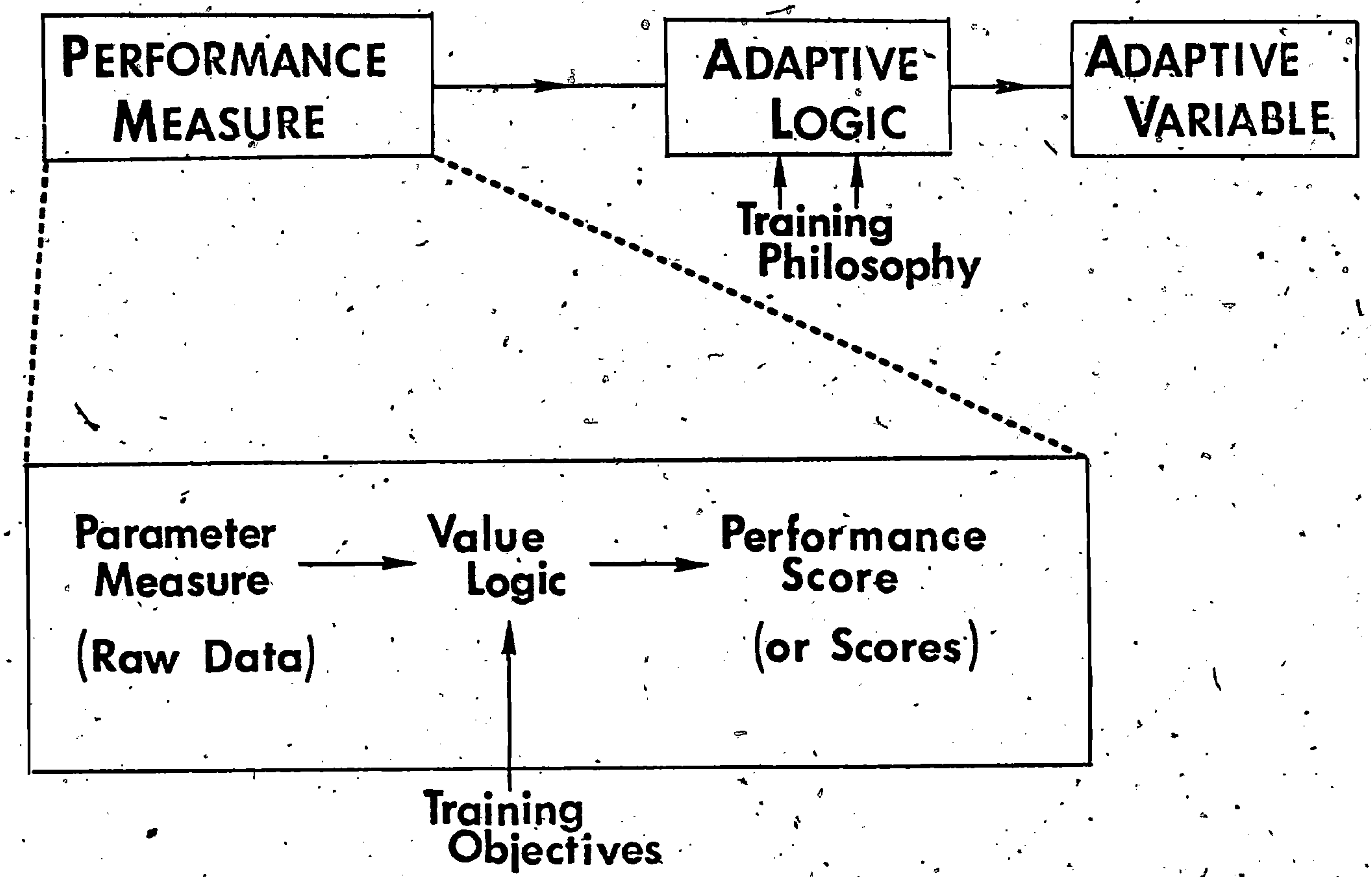


FIG 5

ESSENTIAL ELEMENTS OF A SCORING SYSTEM

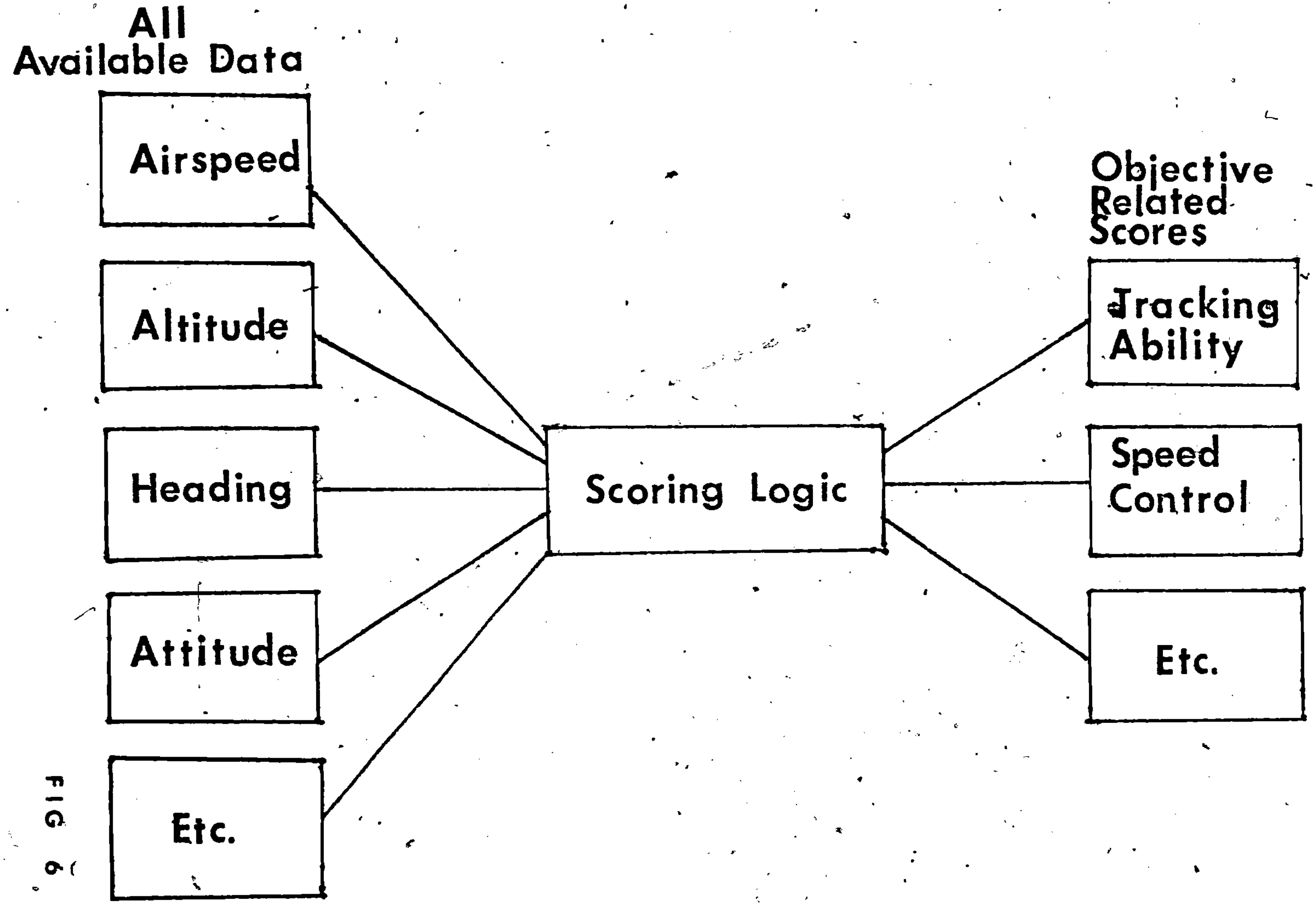


FIG 6

SCORING LOGIC MUST BE RELEVANT TO OUR AIM

Q. Are we looking at the man, the machine, or the interface?

A. In the training context - the man.

Q. What do we want to know about him?

A. How well he performs.

Specifically, Which parameters are affected by performance?

Which parts of the performance are relevant?

Q. How can we measure his performance?

A. Compare it with an ideal.

Specifically, How precise must the ideal be?

Are some apparent errors actually equipment limitation?

How are size and importance of error related?

Are some parameters more critical than others?

FIG 7

APPLICATION OF SCORING LOGIC TO RAW DATA

- SELECT
- EDIT
- COMPARE WITH IDEAL
- MODIFY
- WEIGHT
- COMBINE

FIG 8

SCORING SYSTEM DEFINITIONS

SUBSCRIPTS

x - (Along Track)
y - (Across Track)
z - (Vertical)
t - (Time)

TERMS

E → Error Value
F → Scoring Function
R → Range
$\left.\begin{matrix}k\\K\end{matrix}\right\}\rightarrow$ Constants
t → Time

Typically:

$$f(E) = E, E^{n}, \log E, E-K$$
$$f(R) = 1/R, 1, R$$
$$f(t) = t, 1, kt$$

FIG 9

The Scoring

Simplified,

$$S = \sum_{t=t_o}^{t_f} F_x + F_y + F_z \qquad \text{where } F = \frac{kf(E)f(R)}{f(t)}$$

Extended,

$$S = \sum_{t=t_o}^{t_f} \frac{k_x f(E_x) f(R)_x + k_y f(E_y) f(R)_y + k_z f(E_z) f(R)_z}{f(t)}$$

FIG 10

TYPICAL MLS FLIGHT PATHS

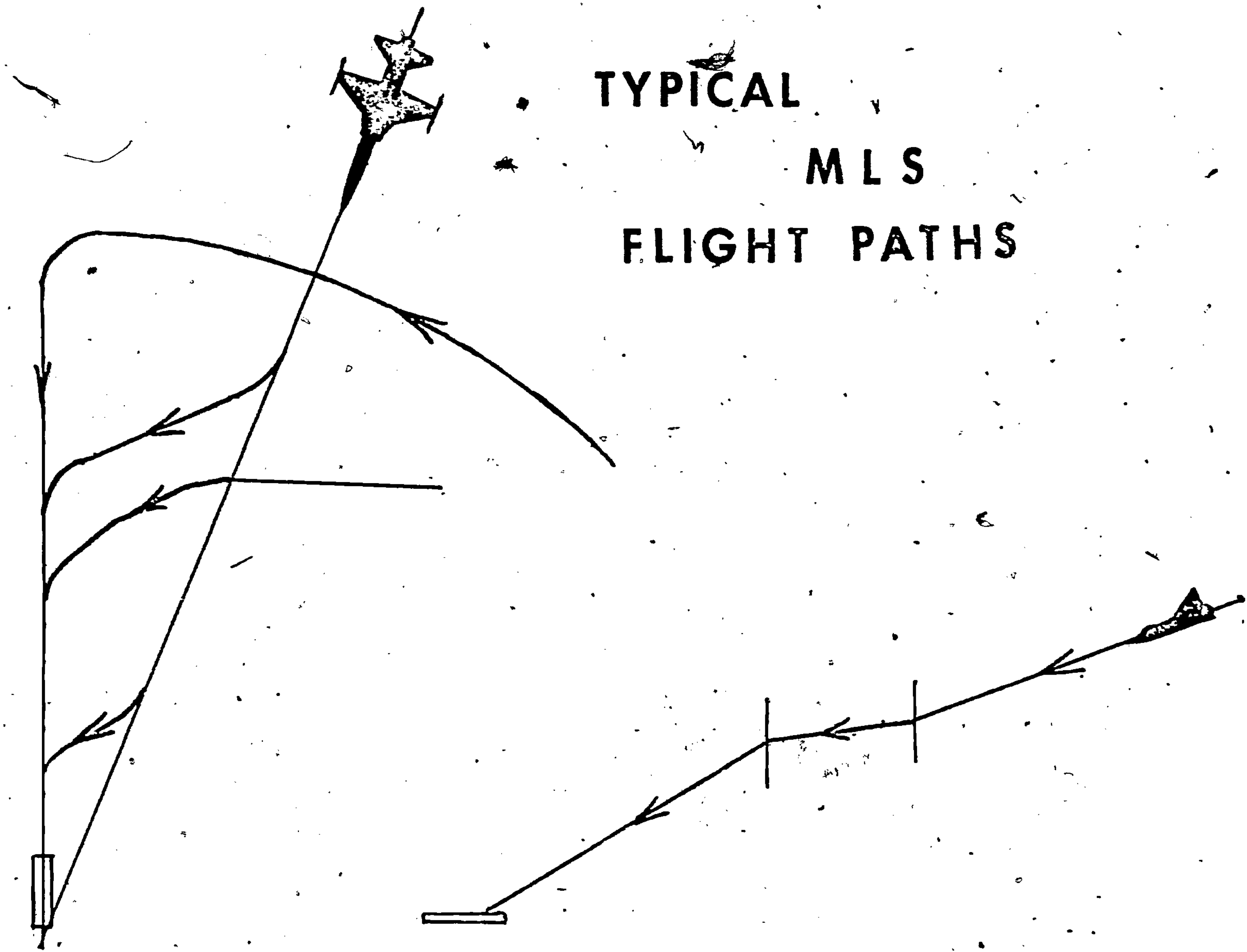


FIG 11

MLS Scoring

$$S_{MLS} = \sum_{t=t_o}^{t_f} \frac{\left(\frac{1}{32} E_x(t)\right)^2 + \left(E_y\right)^2 + \left(k_z E_z\right)^2}{t}$$

where

$$k_z = \begin{cases} 2 & \text{if } E_z > 0 \\ 3 & \text{if } E_z < 0 \end{cases}$$

FIG 12

THE COMPLETE PROCESS

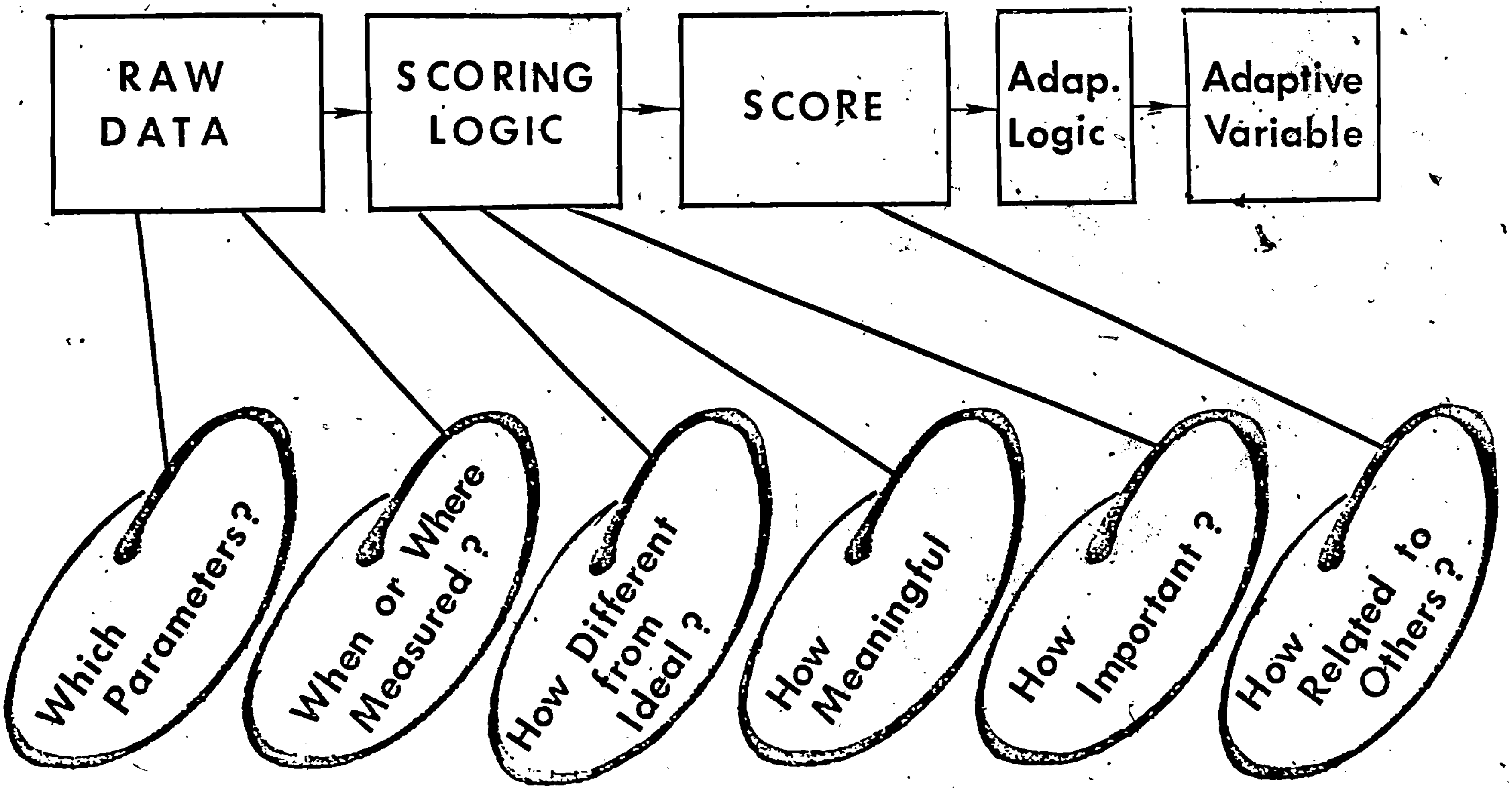


FIG 13