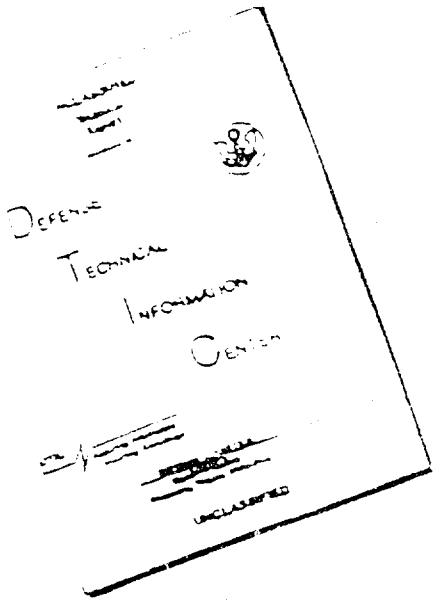
AD734432

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Security Classification

DOCUMENT CONTROL DATA - R & D

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

1. ORIGINATING ACTIVITY (Corporate author)
Human Factors Department
Hunker Ramo Corporation
Westlake Village, California

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Comparative Analysis of Human Reliability Models

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)
David Meister

6. REPORT DATE
30 November 1971

7a. TOTAL NO. OF PAGES
505

7b. NO. OF REFS

8a. CONTRACT OR GRANT NO.
Contract No. N00024-71-C-1287

8b. ORIGINATOR'S REPORT NUMBER(S)

b. PROJECT NO

c.

d.

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

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Naval Ship Systems Command
Washington, D. C.

13. ABSTRACT

The purpose of this study was to describe, analyze and compare available models and methods for making quantitative predictions of human performance in man-machine systems. The 22 methods reviewed were divided into those relating to operability and maintainability; operability models further subdivide into analytic (non-simulation) and simulation models. Each model was analyzed in terms of goals, assumptions, scope parameters, data requirements, procedures and validation/application studies. It was found that most models are reasonably effective for prediction, but are less effective for design analysis, selection and training purposes. Simulation models are more powerful than analytic ones. Choice of a model seems to depend on its particular advantages for solution of specific system development problems. The report provides requirements for development of input data banks and data presentation formats. The most recent studies and the state of the art of human reliability prediction are reviewed. Recommendations for further research are made, centering around a survey of user needs for predictive data.

DD FORM 1 NOV 68 1473

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Evaluation						
Human engineering						
Human factors						
Human performance						
Human reliability						
Maintainability						
Man-machine systems						
Methodology						
Models						
Operability						
Prediction						

L0074-1U7

COMPARATIVE ANALYSIS OF HUMAN RELIABILITY MODELS

Final Report
Contract N00024-71-C-1257

David Meister
Bunker Ramo

November 1971

Prepared for

Naval Ship Systems Command
Department of the Navy
Washington, D. C.

BUNKER RAMO
ELECTRONIC SYSTEMS DIVISION
31717 LA TIENDA DRIVE, WESTLAKE VILLAGE, CALIFORNIA 91361

ABSTRACT

PURPOSE OF THE STUDY

The purpose of this study was to

1. Describe and analyse behavioral models for predicting the performance of personnel in the operation and maintenance of military systems.

2. Compare the models in terms of objectives, assumptions, scope, parameters, procedures, inputs/outputs, uses and validation/application studies.

3. Provide a catalogue of available models among which users could select for their particular needs.

4. Summarize the present state of the art of human performance reliability prediction.

5. Suggest ground rules for development of input data bank(s).

6. Suggest needed further research.

METHOD EMPLOYED

This involved the following steps:

1. Collect, describe and analyze a total of 22 models.

2. Submit written model descriptions to developers for comments and correction of misinterpretations.

3. Secure a consensus of opinions by qualified personnel about criteria for evaluating the models.

CONCLUSIONS

1. The models described fall into the following classes:
   - A. Operability
     1. Analytic (largely reliability-oriented)
     2. Simulation
   - B. Maintainability

2. Simulation models are more powerful than non-simulation models because they provide time histories of system simulations that are useful for diagnostic purposes and because they describe functional relationships between parameters, thus permitting tradeoffs. Because they simulate operator/system processes, simulation models largely avoid the problem of task interrelationships and thus of combinatorial statistics. Most maintainability models do not predict/evaluate human performance efficiency directly.

3. A number of potential uses for predictive models are hypothesized. These include: (a) prediction of the absolute performance of operator/system performance; (b) comparison of predicted performance with a specified quantitative requirement; (c) comparison of alternative system configurations on the basis of predicted human performance; (d) suggestions for the redesign or initial design of the system configuration; (e) suggestions for manpower and training requirements.

Most models can be used for prediction and evaluation of missions/tasks/systems (uses a, b, c). However, they are relatively insensitive to equipment design parameters (use d) and supply relatively little information about manpower selection and training (use e).

4. There seems to be no general purpose model. Each model deals with some situations (e.g., types of tasks, systems, stage of system development) better than with others. Each has certain advantages and counter-balancing disadvantages.

5. A third of the models considered will accept data from any source half will accept data from experimental sources only. Most models require fairly detailed information. With the exception of the methodology associated with the AIR Data Store, no operability model has a standardized data base to be applied to the model. In most cases the parameters the input data must describe are not specifically indicated by the model. Only a few "performance shaping factors" are included in most models.

6. Almost all models use some form of function/task analysis as the basis for identifying the behavioral unit to which input data are to be applied. The task analytic parameters are, however, rarely described.

7. Most models output probabilities of successful task/system performance and completion time.

8. For most models predictive validation/application data are either lacking or incomplete, so that evaluative judgments based on precision/efficiency cannot be made. They provide little information about consistency or ease of use.

9. Development of an effective data bank requires consideration of:

a. The model with which the data will be used;

b. The particular uses to which the data will be put;

c. The parameters of interest to the model user;

d. The level of detail required of the data bank;

e. The output measures of the model with which the data bank will be used;

f. The scope of the tasks/behaviors to which the data must be applied.

10. Significant problems still remain with regard to

a. Subjective estimates of performance;

b. Task independence vs. interdependence;

c. Relative importance of time and error as dependent performance variables.

11. Among the 42 responses received to a questionnaire on evaluative criteria, there was significant agreement on the relative importance of the 16 criteria proposed (.01 level). In general, criteria describing how well the model corresponds to real world events are considered most important; criteria describing the adequacy of model structure and ease of use are considered much less important.

12. It is concluded that many of the models reviewed have considerable potential for solving system development problems relating to human performance, but that further research is required before they can be applied practically to weapon system development.

RECOMMENDATIONS

1. We assume that each of the models has been developed to respond to some assumed system development need and that they differ in terms of their capability of satisfying these needs. Since, however, we do not know what these system development needs are, it is impossible to determine which of the models will be most useful, or whether in fact any of these models will be useful and used. Therefore, it is necessary to determine

   a. Who will make use of these models;

   b. What are the uses to which these people will apply these models and at what system development stage, with what constraints, etc.;

   c. How precise/detailed must the information be which the model outputs and in what form;

   d. What parameters must the model output data deal with.

Answers to these questions will determine what requirements should be levied on models and data banks.

Our first recommendation is therefore that a study be performed to answer these questions. By presenting sample outputs of the various models, their input data requirements, to a broad cross section of potential users, it should be possible to compare the models in terms of their applicability to actual system development problems.

2. In view of the fact that so many models, lacking other sources, make use of subjective data estimates, it is highly desirable that the most effective method for securing such estimates be developed. Although such standard methods (e.g., paired comparisons) exist, it is necessary to determine how much reliance one can place on them (i.e., their validity), what parameters one can expect personnel to include in their judgments, in relation to what tasks, etc.

3. Since the focus of human factors efforts must be on system design, and since all the models we have surveyed lack sensitivity to equipment design parameters, high priority should be given either to the development of a model which is focused on design parameters or to the development of a data bank which specifically includes such parameters.

4. A longer range recommendation which is directed specifically at governmental sponsoring agencies is that emphasis be given to the validation/application of the most promising of the models presently available. The lack of validation is the one most severe deficiency of available models, and makes a comparison among them dependent upon secondary criteria. No model development effort should be considered complete unless it terminates in one or more application-oriented validation studies.

5. Many other studies should be performed. These include studies of task interrelationships, conditional dependencies and the effects of feedback on performance. However, the ones suggested are considered the most immediately pressing.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to a number of people without whom this study could not have been performed:

(1) To the developers of the various predictive models, many of whom supplied research reports of their models, reviewed the author's preliminary draft descriptions and commented (some in great length) on the technical accuracy of these descriptions. The author alone, however, must bear final responsibility for any inadequacies that the reader may find in this report.

(2) To the many human factors specialists who were kind enough to fill out the author's questionnaire on criteria for evaluating man-machine models. The names of these respondents are given in Appendix A.

(3) To Mrs. Dorothy Finley of Bunker Ramo's Human Factors Department who performed the statistical analysis of the responses to the criteria questionnaire.

(4) Not least, to Mr. James P. Jenkins, Head, Systems Effectiveness Branch, Sonar Technology Division, Sonar Directorate, Naval Ship Systems Command, who supported the efforts that led to this report.

(5) Special thanks are due to Mrs. Carolyn Bagdonas who uncomplainingly and with great persistence typed several versions of the manuscript.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

SECTION I

INTRODUCTION AND PURPOSE

For the Navy to exercise any significant effect on the design of man-machine elements during system development, it must apply methods of quantitatively predicting the performance of operators and technicians when the system goes into operation. The term generally used to describe that prediction, "human reliability" (HR), connotes a personnel capacity — akin to that of equipment reliability — to perform equipment operation and maintenance tasks in accordance with system requirements.[1]

Over the past 10 years or more a number of human reliability methods and models have been developed. This is not the place to review past history (which has in any event been done by Swain, Ref. 4); it is sufficient to say that the potential model-user may choose among a number of differing solutions to the problem of human performance prediction. He may choose between operability and maintainability models; between those that simulate behavioral processes and those that do not; between those that function at a relatively molar task level and those whose elements are quite molecular, etc. Since each of these models may be useful for somewhat different system development purposes, or may have advantages and disadvantages depending on the user's needs, it is necessary to analyze and compare the various approaches systematically before making a choice.

Such a comparison is made all the more necessary because the development of these techniques and the research associated with their development has uncovered a number of questions, the relevance or importance of which depends on the particular methodological approach one takes.

Among these questions are:

(1) What metric should be employed in describing human performance?

---

1. This definition of human reliability- or human performance reliability, as the Human Engineering Division of the Aerospace Medical Research Laboratory phrases it- must be differentiated from the same term as used by some psychiatrists in the military services to designate the capability of personnel to resist emotional breakdown under stress. The term as used here has connotations of accuracy.

(2) To what level of detail in behavior and equipment should the method attempt to predict?

(3) How should predictions at one level of man-machine system functioning be combined with predictions at another level?

(4) What kinds of system development and use problems should an HR predictive method attempt to solve?

(5) What should be the characteristics of the historical data (bank) to be used as the basis for making performance predictions?

Perhaps the one critical question that has generated most controversy is how much effort should be expended on the development of a data bank to be used for predictive purposes, and what parameters that data bank should contain. Swain (Ref. 5) among others has called for an intensive effort in developing such a bank based on empirically gathered data, and has suggested the parameters that should be included. However, other workers with a different approach, e.g., Blanchard and Smith (Ref. 1), Knowles, et al. (Ref. 4) would concentrate on the development of techniques for securing subjective judgments.

Assuming that a data bank is required, what data items should it contain? Here one has a choice between the very molecular equipment characteristics of the AIR Data Store (Ref. 3) and Swain's more molar performance shaping factors (Ref. 5). And to what behavioral elements should that bank be applied? Obviously, since any data bank must be applied using a method of some sort, the nature of that method will at least in part determine the characteristics of the data bank.

From that standpoint, any data bank may be considered a behavioral model itself (or at least implies such a model), so that if one is to answer this data bank question logically, it is necessary to adopt a specific methodological approach; and this requires an analysis such as the one described in this report. The fact is that any research problem selected is determined at least in part by a methodolcgical strategy, even when it is apparently independent of that strategy. In consequence a meaningful direction in which to pursue human reliability research cannot be determined until one knows what the various approaches are and what they imply.

The specific purposes of the study described in this report were then

(1) To summarize the present state of the art of human reliability predictive methodology. A major part of the summary is a catalogue of the various predictive methods which a potential model-user [2] can examine to find that method which best satisfies his particular needs. The catalogue contains descriptions of the various models, but only so much detail is provided that a user can acquaint himself with the elements of the methodology. Since some of these methods are highly complex, it is impossible to present them as fully as they deserve. However, references are given to basic documents from which details needed for utilizing the model can be found.

(2) To perform a comparative analysis of the various methodological approaches, to indicate their similarities and differences, to extract their implications for model use, and to infer the problems which further research must solve. As part of this analysis questions are raised about each method. There is no intent to denigrate any method by raising these questions; they are simply to examine the fascinating theoretical and methodological problems that still remain to be solved and to indicate where further research or development is required.

(3) To examine the methodological problems raised by these problems so that it will be possible to outline a program of research required to solve them.

This report is organized in several sections. After discussing the criteria used to select the methods under review, and the criteria employed to evaluate the selected methods, each method will be described. These descriptions are the bulk of the report. A summary of the state of the art of human reliability research is followed by a set of guidelines for the development of data banks. The conclusions reached from the analyses performed previously lead then to recommendations for research to implement the Navy's human reliability program.

2. The term "model-user" refers to those personnel involved in system development who can and should make use of human performance predictions to solve developmental problems. These include (the list is probably not exhaustive) military system planners and project managers, contractor project managers, design engineers, reliability and human factors specialists.

REFERENCES

1. Blanchard, R. E. and Smith, R. L. Man-Machine Modeling: Some Current Deficiencies and Future Needs, in Jenkins, J. P. (ed.) Proceedings of U. S. Navy Human Reliability Workshop, 22-23 July 1970. NAVSHIP Report 0967-412-4010, February 1971, 183-198.

2. Knowles, W. B. et al. Models, Measures and Judgments in System Design. Human Factors, 1969, 11, 577-590.

3. Payne, D. and Altman, J. W. An Index of Electronic Equipment Operability, Report AIR-C-43-1/62-FR, American Institute for Research, Pittsburgh, Pa., 31 January 1962.

4. Swain, A. D. Overview and Status of Human Factors Reliability Analysis. Proceedings, 8th Reliability and Maintainability Conference, Denver, Colorado, July 7-9, 1969, 251-254.

5. Swain, A. D. Development of a Human Error Rate Data Bank. In Jenkins, J. P. (ed) Proceedings of U. S. Navy Human Reliability Workshop, 22-23 July 1970. NAVSHIPS Report 067-412-4010, February 1971, 117-148.

SECTION II

A. METHOD OF CONDUCTING THE STUDY

Two questions summarize the problems the author faced at the start of this project:

(1) What is meant by the term "model" and what criteria can be used to select the human reliability models to be analyzed?

(2) What criteria can be used to evaluate model effectiveness and how do we develop these?

What is a Model?

The reason we were concerned about this question is simply that the term is too inclusive; too many people take the term "model" in vain. When someone wishes to conceal his jejune thinking behind a facade, he calls it a "model". Models may be anything: abstract or concrete, shallow or sophisticated, qualitative or mathematical.

Obviously in this project it was impossible to consider all behavioral models, simply because there are too many of them; many of them are irrelevant to our interest in human reliability and many simply do not have the substance even if they have the name.

Consequently, out of sheer self-preservation, it was necessary to define the subject matter of the project rather closely.

It would be ideal if we had a satisfactory verbal definition of a model, but all-inclusive definitions of models turn out to be rather abstract and therefore not very clear. Siegel and Wolf (Ref. 4) tell us that "... the much-bandied term "model" has been so broadly defined as to incorporate virtually any form of abstraction used to represent concrete phenomena." The procedure adopted to define a model was to describe in a series of steps what a model does and use this as a sort of "template" against which to measure our models. This is illustrated in Table 1, which lists the requirements for a model and then what the model does in relation to a system (presumably a man-machine system).

TABLE 1. DEFINITION OF A HUMAN PERFORMANCE PREDICTIVE MODEL

REQUIREMENTS	ACTIVITIES
The model	The model
1. Starts with certain goals and assumptions; based on these, the model defines the class of systems and system elements it can represent.	
2. Has rules/procedures for isolating * relevant system elements.	* Analyzes the system into its structural elements
3. Specifies the data base it requires.	
4. Specifies the measures for which the data will be used.	* Selects appropriate data from available data sources.
5. Indicates how data will be applied to system elements.	* Applies selected data to system elements.
6. Has rules/procedures for exercising system functions, i.e., making the system work as in the real world.	* Exercises system functions
7. Has rules/procedures for synthesizing (integrating) system operations, elements and measures.	* Develops system output (terminal) measure(s).

* It is necessary to "scope" the system by classifying those essential aspects of the system which are to be considered by the model. For example, in a man-machine system all the human behaviors that occur in system operations can be categorized as functions and tasks of varying complexity. This enables the analyst to assign quantitative values to the performance of these behaviors and to organize them in meaningful ways. In the case of the man-machine system the classification/isolation process is called task analysis.

  Moreover, it is manifestly impossible for any model to deal with the totality of the very large number of parameters that may influence system functioning. Consequently some selection of these parameters is required. The model should specify what the selected parameters are and the criteria for making the selection.

The model is a way of representing a real world system in a quantitative (preferably) fashion. Here we follow Chapanis' (Ref. 2) usage, in which a model is defined in terms of its representation of behaviors. A model of a system is an abstraction which reproduces (simulates) symbolically the way in which the system functions operationally.

From that standpoint of reproducing the behavioral process, the author confesses that some of the "models" to be reviewed in this report are not "true" models. A number of them include as part of their methodology a representation of system operations (e.g., a function flow diagram of how the system performs), but others lack even this characteristic.

In consequence we have "models" that are merely procedural techniques or methods for applying predictive data; we have techniques or methods that include certain "model" characteristics; and of course we have a number of "true" models.

We have not eliminated those methods which are only partial or incomplete models, because to have done so would have meant reducing our sample size by half or more. Moreover, the fact that one method is a model whereas another is not is really irrelevant to the question of how well they predict, although not to the way they predict.

We have compounded the fault- if it is a fault- by referring indiscriminately to the methods reviewed as "models" or "methods" or "techniques", but most often as "models". Only the purist should be disturbed by this usage.

There are three planets in the model universe: the model, the system (rather, a class of systems) the model is designed to represent, and data which permits the model and the system to interact quantitatively.

The left hand column of Table 1 indicates what the model must be in order to do its job; the right hand column describes what the model does when applied to a system.

To summarize what we learn from the left hand column, the model includes: (1) goals and assumptions; (2) definitions of the systems and system elements it can deal with; (3) procedures for isolating or analyzing these system elements; (4) specifications for the data base the model requires to act upon the system and the measures it derives from these data; (5) rules for applying these data to the system elements; (6) rules

for exercising system elements as in a Monte Carlo simulation; and (7) rules for synthesizing (combining) system operations and elements to derive a terminal system output (again in quantitative form).

Eventually it will be seen that this listing of model requirements supplies a set of model-inherent criteria, i.e., to be maximally effective a model must contain these elements.

For the moment, however, we are most concerned with item (2), definitions of those systems and system elements the model can represent, because this tells us which models deal with human reliability.

There are at least three types of models, and some subvarieties:

(1) Those models that predict human performance but in which there is little or no consideration of equipment characteristics.

(a) One subvariety deals with the full range of human task behaviors. An example is a model which describes how people learn or react to stress.

(b) Another subvariety deals only with individual isolated human functions. For example, in this category we would place visual reconnaissance, signal detection, vigilance and decision-making models.

(2) There are models that predict system performance but in which there is little or no consideration of human functions, except possibly indirectly. Examples of such models are: economic utility, cost, reliability, availability.

(3) There are models that include in their operations both human and equipment parameters (i.e., the effect of the human on the equipment and the effect of the equipment on the human, both of these as they affect system performance).

(a) One subvariety deals with the full range of human functions, i.e., perceptual, motor, cognitive, all wrapped up in what we call "tasks" or "functions".

(b) Another variety deals with isolated individual human functions. In this category the most common example is manual control (tracking) models.

What the model deals with depends upon its goals and assumptions. For example, the classical equipment reliability model is interested in predicting reliability at the component, and equipment level. It follows therefore that it does not include human behavior (except as reflected in human-initiated hardware malfunctions).

What are the goals of the models we were interested in? To answer that we had to ask ourselves what our goals in developing human reliability models are. Those goals are:

(1) To measure/predict the effect of the human on equipment/system performance, and the effect of equipment/system elements on human behavior.

(2) Ultimately to determine those equipment (and to a lesser extent manpower selection and training) characteristics that maximize the likelihood of most effective human performance (i.e., human reliability) in the system context.

The two goals are independent but interrelated. One can predict human performance in a system context without determining effective equipment characteristics, and similarly one can determine effective equipment characteristics without predicting human performance in relation to those characteristics (this is classical human engineering). The author happens to believe, however, that human engineering is inefficient unless it is based on measurement and prediction of human performance. He also believes that merely to predict human performance without being concerned about the equipment correlates of that performance is inconsequential, since one can do very little with measurement or prediction by itself. To be meaningful therefore, measurement/prediction must be tied to design consequences, or it is essentially only an amusing game.

From that standpoint any model that involved human performance alone, i.e., that had little or no equipment reference, e.g., decision-making models, or that involved equipment/system performance alone, e.g., availability models, was not germane to our purpose. Moreover, since we were interested in the relation between human and system performance, any model that dealt only with an isolated subset of human functions, even if these had an equipment reference, like manual control, was also not germane, since systems do not function without the full range of human behaviors. In other words, if we want to predict the performance of an operator in a command/control system we cannot be concerned solely with signal detection or solely with decision-making or solely

with tracking capability, etc. because the system involves all of these in an integrated whole.

It was now possible to establish two major criteria of the models to be selected for analysis:

(1) The model (method) must be quantitative or attempt to supply a quantitative value in some way related to the performance effectiveness of the human component of a man-machine system. This automatically eliminated purely qualitative models and those that described social (man-man) systems; it also eliminated models concerned only with machine components.

(2) The model (method) must describe or attempt to describe a relatively full range of human behaviors. It may not handle all of them equally well, and it may handle them only by implication, but we wished to avoid models/methods specialized for a single function like Bayesian decision-making or signal detection, etc. The reason for imposing this constraint was simply that systems almost invariably require the full range of behaviors; hence a model that could predict for only one or two of these behaviors - even though it could do this superbly - would be meaningless for our purposes.

It would have been possible to apply additional model-selection criteria, such as:

The model must contain explicit procedures for application to actual systems; or

The model must be predictive or evaluative rather than merely descriptive. However, models vary in terms of the specificity of their procedures, and only if the model was hopelessly general or abstract was it ignored. Moreover, even if the model were not explicitly predictive or evaluative, if its results could be utilized for predictive or evaluative purposes, it was considered.

B. CRITERIA FOR MODEL EVALUATION

Any analysis of the models described in this report inevitably involves evaluation of these models and hence the application of evaluative criteria. There is no possibility of avoiding these because they are inherent in the analytic process. Some might reject the concept of evaluation, but to do this is to permit no selection of the better from the less good. The analyst therefore has merely a choice between making his criteria overt (and thus more susceptible to control) and allowing them to influence his judgments unconsciously.

The basic problem one runs into with evaluative criteria is that they are unavoidably subjective, representing value judgments on the part of those who analyze the models. Although there is no way of completely eliminating this subjectivity, it is possible to attempt to constrain this subjectivity by securing a consensus opinion from a large number of qualified specialists as to which criteria are most important. This in itself can be considered a worthwhile activity, since it has never been done before.

The procedure adopted involved several steps:

(1) Analysis of evaluative criteria applied to models by other workers in the field;

(2) Development of a comprehensive list of all potential criteria, together with detailed definitions;

(3) Submission of the list of criteria to qualified specialists with the request that they rank these in order of importance;

(4) Development of a subset of criteria for evaluating the models reviewed in this report on the basis of an analysis of the responses made in (3).

In developing the list of potential criteria two questions arose:

(1) Should one be idealistic, and include all possible criteria, or more pragmatic, and limit the selection of criteria to only those that appear most important?

(2) Is it possible to expect to find agreement among a cross-section of specialists with different orientations and uses for models? This ties in with a point that will be raised a number of times in subsequent discussion: that the choice of a model, based on its "goodness", may well depend on the chooser's use-orientation, or what he wants the model to do for him.

The first question was answered by attempting to be all-inclusive (within limits, of course). Rather than the author exercising any a priori judgments, it was decided to allow the "experts" who would rank the criteria maximum opportunity to express themselves.

The second question had to be answered positively; otherwise any means of evaluating models would be impossible. If one adopts the point of view that there are no general evaluative standards, it is impossible to make any meaningful comparisons of models, because judgment becomes purely idiosyncratic. We had to reject this point of view and assume (hope, rather) that despite various points of view some generally accepted standards exist among specialists of various persuasions. The reader will see whether this assumption was justified.

In the development of the list of potential criteria a large number of sources were consulted, three of which were most useful in supplying criteria. The three sources are Altman (Ref. 1), Siegel (Ref. 5) and the author's own criteria paper (Ref. 3). In order that the reader may see the degree of commonality in the standards suggested by these three authors, Tables 2, 3 and 4 list the three sets of proposed criteria. Note the considerable degree of similarity among the three sources.

The various criteria examined appeared to fall into three classes:

(1) Those that describe how well the model corresponds to the real world or predicts operational system performance. Altman's "homomorphism" is an example of this. What we are actually talking about here is the validity of the model.

(2) Those that relate to the structure of the model being evaluated, e.g., its parameters, assumptions, outputs, etc.

(3) Those that describe how efficiently the model can be used, e.g., its ease of use, its applicability early in the developmental cycle.

On the basis of an analysis of the three criteria sources referred to above, as well as other sources (which were, unfortunately, not as productive) a list of 16 criteria was developed for submission to the sample of respondent specialists. This is presented in Table 5.

Certain things must be said about these criteria. It recognized that some of them are more abstract than others, that some are implied by or dependent upon others, (i.e., if the model contains certain qualities, then other related qualities must automatically exist in that model). We make no apology for the less than completely satisfactory status of these criteria. Criteria describing such complex qualities cannot be considered hard and fast evaluative standards. However, they can be useful in reviewing the models.

TABLE 2

Altman's Criteria for Effective Quantification of Human Performance

▲ COMPREHENSIVENESS--the extent to which quantification techniques are capable of dealing with the full spectrum of significant types and aspects of performance.

- Roles--the variety of different kinds of performance (as determined by systems functions or purposes) accounted for.
- Behavioral aspects--the variety of different stimulus inputs, mediating processes, responses, and behavioral feedback mechanisms covered.
- Performance aspects-- the extent to which all of the relevant dimensions of performance are measured.
- Design-performance correlation--the extent, nature, and ease of translation between system design characteristics and performance requirements.
- Environment--the extent to which performance-affecting environmental variables are taken into account.
- Individual differences--the extent to which techniques can deal with performance differences associated with individual performers.
- Developmental versatility--the variety of different stages of design and development to which techniques can be applied.
- System versatility--the variety of different types of systems to which techniques can be applied.

▲ EFFICIENCY--the extent to which performance quantification techniques accomplish useful purposes without waste.

- Data stores--the extent to which the legacy of information available from previous experience is applied to new situations.
- Behavioral frameworks--the degree to and ease with which commonalities from one set of performance requirements to another are recognized.

- Flexibility of behavioral levels--the facility with which it is possible to gauge the size of behavioral units measured to the use to which measures will be put.

- Sensitivity--the amount of impact design variations have on performance estimates.

- Ease of use--the reciprocal of costs involved in administration of performance quantification techniques.

▲ HOMOMORPHISM--the extent to which the structure of quantification matches the structure of performance in practice.

- Directness--the clarity and ease with which quantitative estimates can be understood within the context of actual performance in the system.

- Lawfulness--the extent to which quantification procedures are consistent with accepted principles of human behavior.

- Error detection and correction--the ability of quantification techniques to account for performance error being detected and corrected before full consequences are felt.

- Dependent probabilities--the extent to which quantification procedures are able to account for interdependencies among performance requirements.

▲ PRECISION--the closeness of agreement between quantitative estimates of performance and the actual performance which occurs in systems operation.

- Objectivity--the extent to which quantitative procedures are free from bias on the part of their administrators.

- Reliability--the extent to which quantitative procedures agree from one independent application to another.

- Validity--the extent to which quantitative performance estimates conform to expectations.

TABLE 3

Siegel's Criteria for Evaluating Man-Machine Models

It is assumed that an effective method

1. Will yield numerical probability estimates.
2. Will allow statement of work sequences yielding low (high) reliability.
3. Is applicable early in developmental cycle.
4. Has practicality.
5. Has generality.
6. Is compatible with other human factors techniques; minimum additional analytic requirements.
7. Has validity.
8. Has psychometric reliability.
9. Will yield time as well as probability value.

TABLE 4

Meister's Criteria for Evaluating Human Reliability Techniques

1. Usable by non-specialists.
2. Should not require excessively tedious calculations.
3. Should not require the application of performance data which are not readily available.
4. Must lead to usable design recommendations.
5. Capable of being utilized at all stages of system development; should be able to handle all system elements.
6. Answers must be formulated in task performance terms.
7. Capable of being validated by the collection of performance data in the operational setting.
8. Outputs compatible with those of equipment performance predictive techniques.
9. Capable of assimilating data from various sources.

TABLE 5

MAN /MACHINE MODEL CRITERIA

This questionnaire is part of a Navy project (contract N00024-71-C-1257) to develop a catalogue of models for predicting man-machine performance (i. e. , how effectively humans perform in operating and maintaining systems) and to assess their usefulness.

In its broadest sense, a model is any physical or symbolic representation of a process; but for purposes of this study we define it more narrowly as an organized statement of procedures for reproducing man-machine system operations symbolically by applying quantitative data to the physical and behavioral elements of the model in order to predict the resultant performance of these elements when the model system is exercised.

In orde to evaluate the many models that exist, it is necessary to establish evaluative criteria; this is the purpose of this questionnaire. We are therefore asking you to indicate the relative importance of the criteria below for evaluating the effectiveness of these models, and also to add your own criteria if you feel we have left any out.

Please read the criteria descriptions below and rank them in order of decreasing importance. In other words, the most important criterion would be ranked (1), the next most important criterion would be ranked (2), etc. Although you may feel that two or more criteria are of apparently equal importance, please do not give them the same rank. In order to develop highly discriminating criteria, it will be necessary for you to choose among them, however equal you may feel them to be.

If you feel that we have left out any significant criterion, please add it, describing it as clearly as possible, and rank it along with the others. If you feel that any of the criteria below are insignificant or irrelevant, put an X beside it, and do not rank it.

One last caution. When you consider these criteria, please think of a model which ideally satisfies your concept of what a model should be and do, not of actual models that you are personally aware of and any defects they may have. Please read and consider all the descriptions below before you rank them.

RANK

_____ Compatability

The quantitative output of the man-machine model (that is, after it has been exercised) should be capable of being compatible with or combined with other system outputs like equipment reliability measures to yield a measure of total system performance. For example, if the output of the system model is the probability that an operator will accomplish a particular job, this measure should be capable of being combined with the probability of equipment failure in the same system.

_____ System Development Applicability

Exercising the model should yield results which are useful in answering system development questions, such as: the comparison of alternative system configurations; the selection of equipment design characteristics; the determination of training requirements; or the determination of the manpower needed to run the system. Although the model results need not be phrased directly in terms of such recommendations, it should be possible to derive these recommendations from model outputs.

_____ Timing

The model technique should be capable of being applied (used) early in system development (e.g., conceptual planning stages) as well as in later, detail design and testing stages.

_____ Ease of Use

Personnel who do not specialize in model building (as well as specialists, of course) should find it possible to exercise the model and derive answers with only that reasonable effort required to learn the technique. In other words, the model should not require such specialized background knowledge as complex mathematical techniques or sophisticated computer programming.

Comprehensiveness

The model should be applicable to (supply answers for) a variety of equipment systems (e.g., command/control, fire control, sensing, etc.) and to a variety of behaviors (perceptual, motor, cognitive); and should account for the performance of operator/maintenance teams as well as individuals, individual differences among personnel, varying environments and the effects of a number of task performance variables.

Technique Availability

The model should be able to make use of outputs from conventional human factors analytic techniques, such as task analysis or operational sequence diagrams. It should not impose a requirement for special types of analyses other than those currently available to the man-machine specialist.

Data Availability

The model should be able to make use of data to perform its operations (i.e., to exercise the model) that are reasonably available from common sources. It should not require the performance of elaborate or time consuming or costly additional studies to gather the necessary data.

Effectiveness

The ability of the model to predict man-machine performance with some degree of effectiveness should be demonstrated by comparing the model outputs with data from an external criterion. For example, the predictions of a command/control model might be compared with empirical data from an actual command/control system. Note that we do not use the term validity for this criterion because it is considered that no man-machine model presently is or can be assumed to be completely valid. The criterion assumes only two things: (1) attempts should be made to demonstrate model effectiveness; (2) some degree of model effectiveness (enough to suggest that the model has potential) should be found.

_____ Assumptions

The assumptions made by the model and its procedures for operation should reasonably accord with or at least not violate generally accepted behavioral principles.

_____ Clarity

The model structure should be such that goals, assumptions and procedures for exercising the model should be sufficiently detailed and clear that they can be readily understood and used by others than those developing the model.

_____ Internal Consistency

The model structure must be such that goals, assumptions and procedures for exercising the model are consistent with each other.

_____ Reliability

The model should possess reasonable reliability, defined as: (1) different users of the model technique should obtain the same results when applying the technique to the same system; (2) comparable (although not identical) results are achieved by the same user when the model is applied to several similar systems.

_____ Job-Relatable Measures

The model should supply outputs which can be directly interpreted in terms of relevant system job performance (e.g., accuracy, error, task completion time, reaction time, etc.). In other words, the outputs of the model should be phrased in terms of measures that are directly relatable to actual system job performance.

Intervening Variables

If intervening variables (e.g., aptitudes, cognitive factors) are utilized in the model, these should be linked to and translatable into observable system behaviors.

Objectivity

The procedures for exercising the model and deriving model outputs should be such that they are substantially uninfluenced by subjective processes on the part of the model user.

Analysis and Synthesis

The model should, where required by the nature of the prediction problem, contain analytic procedures capable of breaking down larger units of behavior (e.g., functions) into smaller ones (e.g., tasks); and of combining (synthesizing) measures of more detailed behavioral units to create measures of more molar behavior-units. For example, if it is desired to predict performance at the task level, it should be possible to decompose functions into the tasks of which they are composed in order to output task-relevant data; and, conversely, to take the task outputs and combine them to derive function outputs.

Others. (Add your own, defining them as precisely as possible.)

_____

_____

_____

The questionnaire presented in Table 5 was sent to a list of 52 potential respondents, representing a cross-section of governmental, industrial, academic and consultant personnel selected because of their seniority in their respective fields. The list was developed by the author and checked by the technical monitor for the project, J. P. Jenkins, and his staff.

Of the 52 specialists to whom the questionnaire was sent, about 42 responded; their names are listed in Appendix A. We say "about 42" because several specialists responded whose rankings could not be used because they did not follow instructions. On the other hand, several respondents induced members of their staffs to complete the questionnaire, which added to the total N. Taking 42 as the number who responded, the percentage of return for this questionnaire was approximately 80%, which is quite high for a mailed questionnaire.

The reader will note that respondents were asked simply to rank the criteria in order of importance and to eliminate tie-ranks to avoid clustering of responses. Several respondents asked why the questionnaire did not require paired-comparison selections, or ratings of each criterion on a scale of importance. Neither of these alternatives was selected, although either would have been preferable to forced choice-rankings; however, it was felt that if excessive demands were made on respondents, too many would fail to respond.

Provision was made for respondents to add any criteria they felt were not included in the list of 16, and to eliminate any criterion which was felt to be irrelevant. About 10% of the respondents each suggested as many as four additional criteria; however, many of these duplicated the criteria in the questionnaire (although in different words). Of the 672 criteria ranked (42 respondents x 16 criteria), only 25 were eliminated as irrelevant, or less than 4%, suggesting that the overwhelming majority of respondents felt that the criteria presented were relevant.

Table 6 and Figure 1 present the results of the analysis of the responses. In addition to the mean and standard deviation of the rankings for each criterion, Kendall's W coefficient of consistency (Ref. 6) was applied to the matrix of 16 x 42 criterial rankings. The purpose of the W analysis was to determine whether there was overall agreement among respondents as to the relative weighting they would apply to the various criteria. The W value was .36, which, with an N of 16, is significant at the .01 level. Inspection of the "raw" rankings in the 16 x 42 matrix also reveals a high degree of correspondence among the rankings. Thus, our original fears that there would be wide disagreement among specialists, based on differences in professional orientation, turned out to be unwarranted.

TABLE 6

Summary of Rankings Given to Potential Criteria

CRITERIA	MEAN	STANDARD DEVIATION
Real World Correspondence		
Effectiveness	2.8	2.6
Reliability	3.7	2.5
System Development Applicability	4.1	3.1
Job Relatable Measures	8.2	2.9
Compatability	6.5	3.8
Model Structure		
Intervening Variables	12.1	3.1
Analysis and Synthesis	11.7	3.5
Internal Consistency	7.7	3.9
Comprehensiveness	11.1	4.4
Assumptions	8.5	3.9
Objectivity	7.7	3.5
Model Use		
Clarity	9.2	4.1
Timing	9.9	4.5
Ease of Use	10.4	4.7
Technique Availability	10.8	4.2
Data Availability	8.7	3.8

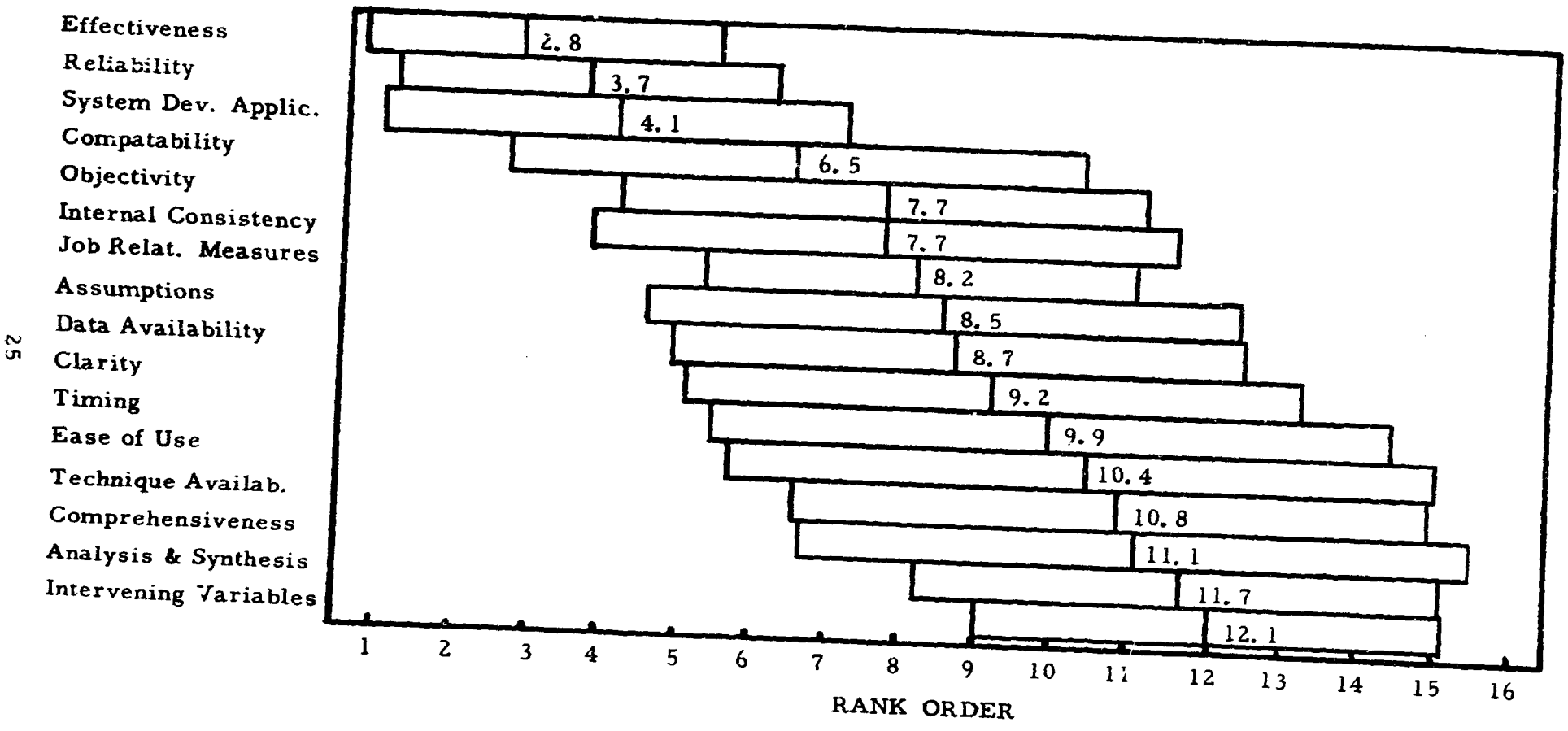


FIGURE 1. MEAN AND RANGE OF RANKS ASSIGNED TO MODEL EVALUATION

Recalling the three classes of criteria noted previously, the 16 criteria can be divided as follows:

Those involving real-world correspondence:

(1) Effectiveness;
(2) Reliability;
(3) System development applicability;
(4) Job relatable measures;
(5) Compatability

Those dealing with model structure:

(6) Intervening variables;
(7) Analysis and synthesis;
(8) Internal consistency;
(9) Comprehensiveness;
(10) Assumptions;
(11) Objectivity

Those describing model use:

(12) Clarity;
(13) Timing;
(14) Ease of use;
(15) Technique availability;
(16) Data availability.

Table 6 and Figure 1 indicate that 4 of the 5 real world correspondence criteria were ranked highest; the only exception is job relatable measures, thus indicating that specialists place most emphasis on the capability of behavioral models to predict efficiently. The situation is a bit more muddled with the other two classes of criteria. Of the 6 model structure criteria, 3 were considered on the whole as being least important; however, internal consistency and assumptions followed hard upon real-world correspondence criteria. Model use criteria occupied a somewhat intermediate-low position in the rankings.

These judgments must be tempered by the high degree of response variability found, as shown in Figure 1. Differences in criterion weighting corresponding to differences in professional orientation show up markedly. However, greatest consistency (as shown by smaller standard deviations) is present in real-world correspondence criteria. Although a measure of the significance of differences between rankings for individual criteria was not attempted, it is apparent that the 4 real-world correspondence criteria are markedly set apart from the remaining criteria, whose ranks are closely bunched together.

What can one say about the results of the questionnaire responses?

(1) There appears to be general agreement that criteria describing how well models predict or relate to real-world performance are most important;

(2) Specialists feel that model structure is of less importance than model use;

(3) There is a high degree of variability in rankings of the importance of the various criteria which tends to support the suggestion that the way in which one views a model depends to a large extent upon one's professional orientation and the use to be made of the model.

Based on the preceding analysis and partly in consideration of the practical problems of applying criteria to models, a number of criteria among those tested were selected to evaluate the models reviewed in this report.

Of the 16 criteria originally developed, 7 were finally selected, several of these being combined because they seemed to represent different aspects of the same criterion.

Two of the criteria selected (validity and reliability) achieved the highest ranking in the responses made to the questionnaire, and had the lowest variability in responses. The criterion of system development applicability, which had the next highest ranking in the responses, was considered to be related to a number of other criteria, i.e., comprehensiveness of systems/task being evaluated, and to timing, i.e., stage of system design at which the model could be applied. In consequence a third criterion was developed, also termed system development applicability, which consisted of the original system development applicability, comprehensiveness and timing criteria. A fourth criterion was developed, termed model characteristics, which combined the original objectivity criterion and a new criterion which appeared to represent the essential discernable feature of the former criteria dealing with assumptions, clarity, internal consistency and intervening variables. This last (new) criterion relates to the extent to which the structure of the model was described in detail.

The other criteria were eliminated because it appeared that they would not discriminate among the models (i.e., all the models possessed the quality in common) or because it would be inordinately difficult to make the judgments required. For example, it was assumed that all the models output job relatable measures. Criteria dealing with the individual aspects of model structure were considered overly complex to discriminate in their original form.

Originally it had been intended to scale the four evaluative criteria and thus to provide a quantitative evaluation of the models reviewed. On more reasoned reflection, however, this seemed unwarranted, since the subtleties of the models make scaling of their dimensions largely uninterpretable.

Table 7, however, which presents the criteria finally selected does suggest that a model may possess different amounts of each criterion quality. The criteria listed in Table 7 are the basis for the Evaluation subsection concluding each model description.

TABLE 7

CRITERIA FOR EVALUATING PREDICTIVE MODELS

VALIDITY

Validity is defined as the demonstration of (or the attempt to demonstrate) the model's ability to accomplish the objectives for which it was developed. It is considered (1) that no technique is likely to be completely validated, but that degrees of validation exist; (2) that there are various ways of validating a technique, among which are formal experimental studies, correlational studies (concurrent validity) and application to system development problems; (3) that the most effective validation is prediction of an effect which is then demonstrated in the "real world".

-- No validity data exist or are available, nor has the method been applied to system development problems.

-- Although formal validation of the method has not been performed, the method has been applied to system development problems and users report reasonable success in its application.

-- Formal validity studies have been performed and show a reasonable degree of correspondence between predicted and observed values.

-- The method has been formally validated and has been successfully applied to system development problems.

RELIABILITY

Reliability is defined as the ability of various users to apply the method with reasonable consistency among the users and to achieve comparable results when the method is applied to several similar systems. It is considered that (1) the most effective demonstration of reliability is a formal correlational study but that (2) reports by users indicating consistent results are also acceptable as demonstrating reliability.

-- The method has never been utilized by anyone other than its developers, nor are data available reflecting on the consistency of its use by different analysts.

-- The method has been applied by various analysts and although no formal reliability studies have been performed, it is reported that the method can be applied by users.

-- Controlled studies have been performed that indicate that the method can be applied with reasonable consistency by various analysts.

-- Controlled studies indicate that not only do various users get consistent results in applying the method, but that when the model is applied to similar systems, consistent results are achieved.

SYSTEM DEVELOPMENT APPLICABILITY

This criterion is composed of three dimensions. Dimension A measures the applicability of the method to various types of equipment/system/tasks/behaviors. Dimension B measures the applicability of the method to various kinds of system development uses. Dimension C measures the applicability of the method at various stages of design.

In A the types of tasks to be considered are: discrete and continuous (e.g., tracking) tasks; perceptual, motor and cognitive behaviors. In B the types of system development uses the method can be applied to are: evaluation of operational systems; prediction of future system effectiveness; comparison of design configurations; design analysis and redesign suggestions; selection/training requirements. In C the stages of system design are operational only; later design only; all stages.

A. Comprehensiveness

-- Method is limited in its application to specific types of equipment/systems (within its field of specialization to operability of maintainability);

-- Model is limited in its application to specific types of behaviors/tasks (again within its field of specialization to operability or maintainability)

-- Model can be applied to all types of equipment/systems/tasks/behaviors.

B. Applicability

-- Model does not predict future performance but only measures on-going system performance;

-- Model does not output a prediction of equipment/system or mission/task effectiveness, but is descriptive only of future system performance;

-- Model outputs a prediction of equipment/system/mission/task performance effectiveness, but cannot be used (or used only with difficulty) for other purposes (e.g., design analysis, training/selection requirements);

-- Model outputs a prediction of system performance effectiveness and can also be used for other purposes such as design analysis, etc.

C. Timing

-- Model can be applied only to systems that have become operational.

-- Model can be applied only to systems in their later stage of design.

-- Model can be applied to systems in early as well as later stages of design.

MODEL CHARACTERISTICS

The characteristics of the model are based on two dimensions, A, objectivity, B, definition of structure. All other things being equal, a model is assumed to be best which requires fewest subjective judgments (and the method of securing these should be explicit); and whose conceptual structure (assumptions, parameters, etc.) are explicitly defined and described in detail.

A. Objectivity

-- Model requires many subjective judgments and the method of securing these is not explicit;

-- Model requires some or only a few subjective judgments but the method of securing them is not explicit;

-- Model requires some or only a few subjective judgments but the method of securing these is highly explicit;

-- Model requires practically no subjective judgments at all in it application.

B. Structure

-- The assumptions and parameters underlying the model are not explicitly defined or described in detail.

-- The assumptions and parameters underlying the model are explicitly defined and described in detail.

REFERENCES

1. Altman, J. W. Progress in Quantifying Human Performance. Paper presented at the Electronic Industries Association System Effectiveness Workshop, Chicago, Ill., 18-20 September 1968.

2. Chapanis, A. Men, Machines and Models. American Psychologist, 1961, 16, 113-131.

3. Meister, D. Criteria for Development of a Human Reliability Methodology. In J. P. Jenkins (Ed.) U. S. Navy Human Reliability Workshop, Washington, D. C. 22-23 July 1970, NAVSHIPS Report 0967-412-4040, February 1971, 25-64.

4. Siegel, A. I. and Wolf, J. J. Man-Machine Simulation Models. New York: John Wiley and Sons, 1969.

5. Siegel, A. I. The Applied Psychological Services Program Plan for Developing a Human Reliability Prediction Method. In J. F. Jenkins (Ed.) U. S. Navy Human Reliability Workshop, Washington, D. C., 22-24 July 1970. NAVSHIPS Report 0967-412-4010. February 1971, 246-260.

6. Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

SECTION III

DESCRIPTIONS OF THE PREDICTIVE MODELS

The models described in this report are listed below in the order in which they are presented:

A. Operability Prediction Models

1. Analytic Methods

(I) American Institute for Research (AIR) Data Store
(II) THERP-Technique for Human Error Rate Prediction
(III) TEPPS-Technique for Establishing Personnel Performance Standards
(IV) Pickrel/McDonald model
(V) Berry/Wulff model
(VI) Throughput Ratio
(VII) Askren/Regulinski model
(VIII) DEI-Display Evaluative Index
(IX) Personnel Performance Metric
(X) Critical Human Performance and Evaluative Program(CHPAE)

2. Simulation Methods

(I) Digital Simulation model
(II) TACDEN
(III) Boolean Predictive technique
(IV) MOS-Human Operator Simulation
(V) ORACLE-Operations Research and Critical Link Evaluator
(VI) Personnel Subsystem effectiveness model

B. Maintainability Prediction Models

(I) ERUPT-Elementary Reliability Unit Parameter Technique
(II) Personnel Reliability Index
(III) MIL-HDBK 472 prediction methods

We do not maintain that this collection of models represents all those that might exist; they are, however, all that we could find. To be completely safe, therefore, the set of methods reviewed in this report should be considered only as a sample of all existing models. However, because of the many sources that were examined, it is felt that the ones included represent the greatest majority (e.g., 90%) of presently available man-

machine system effectiveness models, especially of those published in the "open" literature. Among them, moreover, are those best known to workers specializing in the area.

As far as sources that were examined to locate models, the following possibilities were canvassed:

(1) Letters were sent to other specialists requesting leads.

(2) The published human factors and reliability literature was reviewed, e.g., Ergonomics, Human Factors, J. Applied Psychology, Annals of Reliability and Maintainability, government reports, etc.

(3) The American Psychological Association performed a computer search of its abstracts for the author. However, this covered only the last three years because its information retrieval system is still in embryo form.

(4) Various other abstracts and review documents were persued.

(5) Finally, a notice was published in the Human Factors Bulletin (a journal of the Human Factors Society), requesting anyone developing or aware of a behavioral system effectiveness model to contact the author.

The list above has been divided between models that predict system operability (i.e., the ability to operate the system) and those that predict system maintainability (i.e., the ability to perform maintenance on the system) and system downtime.

The operability models are further differentiated in terms of whether they (a) develop predictive indices by analyzing data banks to select and assign an appropriate value to the behavioral unit being predicted or (b) develop predictive indices by simulating behavioral processes (usually on a computer). Category (a) operability models do not employ simulation methods; category (b) models do. Obviously analysis is involved in both types of models, if only in performing the task analysis which is the customary exordial phase in development of the terminal prediction. However, the reason we call the first category "analytic" is because the determination of the predictive value (e.g., .9997 for task X) is made only on the basis of an analysis of available data. For example, in the AIR methodology one must select from its Data Store the predictive values

corresponding to the significant characteristics of the equipment components involved in the task. Once those values are selected, the terminal task/equipment/system prediction is determined solely by combination of the original selected value(s). Obviously simulation models make use of data banks, but in their case the analysis phase is followed by a simulation phase (which corresponds to the combinatorial phase of the analytic models) and the development of the terminal task/equipment/system prediction arises directly from that simulation, not from the analysis.

The maintainability models reviewed are primarily analytic (the one possible exception being Siegel's Personnel Reliability Index as applied to his 1-2 man digital simulation model).

The above distinctions are made without prejudice to the fact that computers may be used in the combinatorial operations of the analytic models. The essential distinction is that in one case computers are used to simulate behavioral processes, in the other case they are not so used.

There are other ways of characterizing these models, for example, in terms of the range of behaviors and tasks they cover, but the categories selected seem to be the most meaningful.

The following are the category headings around which the model descriptions are organized:

Introduction A capsule description of the model designed to orient the reader to further details. This section includes any special characteristics that distinguish the model being described.

Goals Describes what the developer of the model is attempting to do with the model. This is important because the model can only be evaluated in terms of what its objectives are. This section ties in with a later section on Anticipated Model Uses, because only if one knows what the model is designed to do can one ask whether these goals are actually accomplished. The implications of these goals in terms of what they require in the way of model assumptions, parameters, procedures and data are examined.

Assumptions This section describes any behavioral or non-behavioral assumptions which the model makes. These are examined in terms of whether they are reasonable (accord with experimental evidence or real-world experience). If model assumptions are unreasonable, doubt is cast on the validity of the model. This section also contains any definitions of special terms used by the model.

Methodological Scope This section asks whether the model will cover the range of tasks and behaviors commonly found in systems. In particular, the question is asked whether the model will deal with continuous as well as discrete tasks; with cognitive as well as perceptual/motor behaviors; with the range of systems commonly found in the military, e.g., command-control, sensing, piloting, etc.

Parameters This section asks what the elements of the model consist of and any factors that play a significant role in the operation of the model. For example, certain models make a special point of stress as a factor in mission accomplishment. The implications that these parameters have for such things as data requirements or type of output measure are also examined.

Data This section includes subsections on input data required, the sources of those data and outputs of the model. Specifically, we ask: what kinds of input data are needed; how easy or difficult is it to secure these data; is a data bank required or not; what kind of measures does the model output; how useful are these measures?

Procedures for Model Application The main subsections are: (1) analytic method; (2) method of synthesis; (3) the behavioral unit to which data apply. Analytic method refers to the manner in which the model user must analyze the system in order to define the behavioral unit being predicted and to determine what dimensions the input data should have. Method of synthesis deals with the process that, given that the model user has analyzed the system down to its component units (e.g., behaviors, equipment components) to which predictive data will be applied, he uses to reconstitute or rebuild lower level units into higher ones. The section describes the analysis-synthesis process and what its implications for data outputs are. Finally the section asks to what behavioral unit (presumably the most molecular) the predictive data are applied?

Anticipated Model Uses This section deals with what the model can be used for, in terms of the goals already noted. The model's applicability is discussed in terms of the following potential uses: prediction of system effectiveness; design analysis; manpower selection; and training requirements. This section examines how effectively the model can accomplish its goals and satisfy system development requirements.

Validation/Application Studies The evidence for the validity and applicability of the model is examined in terms of formal studies performed and the various systems to which the model has been actually applied. The adequacy of the validation evidence is examined.

Summary Evaluation This section sums up the preceding examination of the model and presents an evaluation of the model in terms of the criteria described in Table 7.

In order for the reader to evaluate the following model descriptions correctly he should know that most of them were submitted to their developers in preliminary draft form for review. This permitted correction of any technical inaccuracies and also allowed developers to comment on and rebut any judgments made by this author. In a few cases the whereabouts of the developers were unknown and so they could not be contacted. However, all but one or two of the operability models were presented to their developers for review and their comments, if they were made and were appropriate, have been included in the written descriptions. This could be done, however, for only 2 of the 6 maintainability models.

A. OPERABILITY PREDICTION MODELS

ANALYTIC METHODS

I. INDEX OF ELECTRONIC EQUIPMENT OPERABILITY (AIR DATA STORE)

INTRODUCTION

The Data Store was developed in 1962 by the American Institute for Research (AIR, Dr. J. W. Altman and a number of colleagues). Basically, it consists of a compilation of data describing various characteristics of controls and displays and is phrased in terms of the probability of successfully operating these equipments as a function of their characteristics. The minimum time needed to operate the equipment, together with increments of time required by individual equipment characteristics, is also provided. Figure 2 presents a sample Data Store card for a class of controls (joysticks). After system tasks have been analyzed to determine applicable equipment characteristics and behaviors, the probability and time information provided in the Data Store, derived from a review of relevant literature, is applied to these characteristics and behaviors. A measure of equipment operability is developed by multiplying the probabilities for the individual equipment characteristics and behaviors (and adding the times needed for their operation) to determine the operability of each task. Individual task reliabilities are then multiplied to determine the operability of the entire equipment or system under consideration.

The preceding paragraph is merely a gross description of the method; further details are given in the remainder of this section; and for the most complete description of the method, readers are urged to refer to the source documents listed at the conclusion of the section.

GOALS

The goals of the technique are to:

"1. Predict the time and reliability (accuracy) of operator performance.

2. Identify specific design features which degrade operator performance.

3. Provide general guidance concerning selection and training of operators for evaluated equipments."
(p. 3, Ref. 3)

JOYSTICK

(May move in many planes)

Base Time = 1.93 sec

TIME ADDED (sec)	RELIABILITY	
		1. Stick length
1.50	0.9963	(a) 6-9 in
0	0.9967	(b) 12-18 in
1.50	0.9963	(c) 21-27 in
		2. Extent of stick movement (Extent of movement from one extreme to the other in a single plane)
0	0.9981	(a) 5-20 degrees
0.20	0.9975	(b) 30-50 degrees
0.50	0.9960	(c) 40-60 degrees
		3. Control resistance
0	0.9999	(a) 5-10 lb
0.50	0.9992	(b) 10-30 lb
		4. Support of operating member
0	0.9990	(a) Present
1.00	0.9950	(b) Absent
		5. Time delay (time lapse between movement of control and movement of display)
0	0.9967	(a) 0.3 sec
0.50	0.9963	(b) 0.6-1.5 sec
3.00	0.9957	(c) 3.0 sec

Figure 2. Sample Data Store Card

Other goals are implied in the criteria which guided development of the operability index (p. 3, Ref. 3). For example, one of the criteria was that specific design features should be explicit in the evaluation process, so that the model should be a diagnostic as well as an overall evaluation tool. Moreover, it was intended that every factor of known importance should be included in the procedure.

We shall see later whether all these goals can be accomplished by the method. We should point out, however, what these goals imply.

The operability index is a prediction of operator performance in relation to specific tasks and operations required by the system mission. Although it might be possible to consider the index as a somewhat arbitrary figure of merit for equipment (in much the same way that Siegel's DEI measure can be considered), the developers insist on its relationship to actual operator performance. Consequently, in evaluating the index, its assumptions and parametric interrelationships can be considered reasonable only if they conform to what is known about human performance.

The goals of the technique also include design diagnosis and selection/training information. In evaluating the adequacy of the method it is therefore necessary to consider whether one can indeed secure meaningful information from the index relative to design diagnosis and selection/training.

The above statements are made, not to disparage the Data Store technique, but because there seems to be a general tendency on the part of model/method developers to extend the goals of their methods beyond what may reasonably be expected of these techniques. A technique may be of considerable value for a limited objective; but if the objective is extended to cover a wide range of uses, the technique may suffer because it cannot live up to all that is expected of it.

ASSUMPTIONS

The following assumptions are implicit in the technique:

1. Operator performance is influenced by molecular equipment design features, e.g., joystick length. Recall that one of the criteria applied in the development of the index was that all factors of known

importance should be included. Obviously, operator performance is directly or indirectly influenced by a multitude of factors. However, the fact that the Data Store includes primarily molecular equipment features suggests one of two possiblilities: either the developers feel that only design features are significant for operator performance; or the Data Store structure is influenced by the availability of design feature information in the literature and the non-availability of information on other factors in that literature. The second supposition is probably correct: the Data Store was developed on the basis of only 164 studies.

Although more molar processes are implied by the Data Store (see the mediating processes included), the amount of data on these molar processes is very slight- which is probably a reflection on the adequacy of the literature. However, the inability to include central processes (e.g., task factors such as specificity of instructions, amount of feedback, etc.), means that the claim of the index to performance prediction can be accepted only in part. Obviously it fails to consider many factors which do influence performance.

If one were using the index solely to evaluate human engineering adequacy, instead of predicting operator performance (the two are vastly different in scope), the failure to include central processes would be far less important.

2. Behavior can be broken down into a Stimulus-Organism-Response (SOR) framework ("O" refers to the central processes referred to in (1)) and each aspect can be handled separately, i.e., can be measured separately for information reception, internal processing, responding. This means that behavior at a molar level must be broken down into these individual elements and then resynthesized. Table 8 (Behavioral Levels) illustrates the analytic process and the recombinations required to get from dimensions to mission performance (what we call synthesis).

A major aspect of the model (we use the term to refer to S-O-R framework rather than to the overall technique) must therefore deal with this analysis-synthesis process. A basic question that must be raised about this assumption is:

Can one in fact meaningfully extract behavior dealing with information reception without simultaneously considering its effect on responding, its relationship to internal processing, and vice versa?

TABLE 8

Behavioral Levels

Mission: operate fuse jammer
  Phase: prepare for operation
    Task: activate amplifier
      Behavior (or step): throw S11 to ON position
        Aspects of Behavior: (inputs, mediating processes, outputs)
          Components: (specific categories of an aspect) toggle switch as a component of output
            Parameters: (relevant characteristics of components) angle of the throw from position, as a parameter of the component toggle switch
              Dimensions: (specific values or characteristics of parameters) 40^{o}

If the operator looks at a display, is not his performance in viewing that display determined not only by the physical characteristics of the display but also by his mental set, the type of perceptual task he must perform in relation to that display, etc? Which means, in effect, that any perceptual performance value in the Data Store also includes (or must include) a value representative of the central process determining the operator's perception.

Similarly, in synthesizing the performance of the operator to describe more molar tasks, (going from dimensions to the mission), the question one must ask is whether a combinatorial process which assumes independence of the S-O-R elements is in fact tenable.

During the process of analyzing system/mission operations (as in Table 8) it is necessary to decide which behavioral parameters and dimensions are relevant to the behaviors being studied. We do this constantly in using task analysis methods, which suggests that it would be unfair to criticize the AIR method for this requirement. However, the parameters and dimensions from which one must select the ones relevant to a particular task are limited by those included in the Data Store. In other words, the Data Store parameters and dimensions may be insufficiently inclusive. (This was implied previously in referring to the lack of data on mediating processes.)

3. Our comments so far have focussed on the interaction among behavioral parameters. The developers of the method recognize (p. 7, Ref. 3) that the consequences of this interaction are unknown and that this is a major limitation of the approach. They indicate that the determination of interaction effects is currently beyond the state of the art. However, they assume that "interaction effects will tend to balance out so that results of evaluation will not be consistently in error."

One must ask whether this last assumption (balancing out of interaction effects) is a tenable one. Does the literature indicate that statistical interaction effects are generally non-significant? (Incidentally, one wonders whether it would be possible to take account of statistically significant interaction effects in the original studies to modify the probability values associated with the various dimensions?)

The assumption of non-significant interaction effects considerably simplifies the methodology and it may in fact be necessary to accept it, if one is to do anything practical with a data bank developed on the basis of experimental studies. If we are reluctant to accept this assumption, it is because the complexity of human behavior strongly suggests important interaction effects.

4. Ideally, because the AIR model breaks behavior down into its S-O-R elements, it follows that data should be secured that reflects those behavioral elements, e.g., data describing visual perception, decision-making, psycho-motor performance, etc. However, the data that are available do not describe these behavioral aspects (or at least in the experimental literature they are so scattered across or contaminated by machine variables that one cannot readily equate performance with behaviors independent of machine variables). As a consequence, it is assumed that "a careful study of the sources of machine outputs would provide the information concerning the range of stimuli with which man would be expected to cope. Similarly, a study of machine inputs, essentially controls, would identify a majority of the characteristics of man's response," (p. 6, Ref. 3). In other words, the range of perceptual behaviors included in the Data Store is determined by the characteristics of machine sources of stimuli, and similarly response behaviors are determined by machine mechanisms of responses. One may ask whether one can in fact equate a behavior with the machine characteristics that lead to or influence that behavior? More pragmatically, what this leads to is that the range of behaviors considered by the Data Store is essentially determined by available machine displays and controls and even

more by those controls and displays which were selected for performance testing by experimenters.

The fact that machine characteristics are used to organize behavior values means that central processes are likely to be overlooked. On the other hand, since one cannot isolate behavioral responses from the stimuli (e.g., equipment characteristics) that elicit them, there is some justification in using these equipment characteristics to determine the range of the behavior responses. Our objection is directed more to the fact that all of the dimensions relevant to equipment are not included in the Data Store, because experimenters have not systematically tested all of the available dimensions. (This is not mere carping; a data bank must be examined in terms of the parameters it includes.)

5. The conceptual structure implied by the S-O-R framework involves for the AIR developers four levels of classification: aspect of behavior, components, parameters and dimensions. For definitions of these, see Table 8. It is important to note that aspects of behavior represent S-O-R elements at a very gross level. Aspects are not equivalent to individual functions, such as detection, classification, counting, etc., but to complexes of behavior such as perception. In consequence, individual differences in behavioral functions[1] are ignored, although it is well known that they influence perception or motor responses considerably.

It is understandable why differences in behavioral function are ignored in the Data Store structure: the performance in the Data Store is organized by machine characteristics (components, parameters, dimensions) and for these individual functions are irrelevant. In other words, the characteristics of a meter, for example, remain unchanged regardless of the manner in which one views the meter.

One side effect of ignoring the individual functions is that the Data Store model does not provide explicit guidance with regard to the identification of functions, tasks, and subtasks; in other words, how one should go about breaking down gross system operations into the behavioral units to which one attaches performance estimates is not indicated, although the general principles are those of task analysis.

---

1. (e.g., whether one is viewing a CRT display- perceiving- to count the number of data items of a given category vs. viewing that display to add new data)

6. If one looks at the Data Store structure (see Figure 2), it would appear as if each of the equipment dimensions had a performance reliability of its own. This is not true. The value provided represents the effect of that dimension or characteristic on the operator's overall performance. In other words, a performance reliability of .9999 for a given dimension presumably means that 1 out of 10,000 times an error would occur as a result of the given dimension. It does not mean that if that equipment dimension is included in a design that the operator's performance will be .9999. It does not mean that one can describe the operator's performance in a given task by means of that single dimension, because many dimensions will affect that performance. Since the individual dimensional value does not represent the operator's performance, but rather the effect of that dimension on his performance, it does not reflect operator performance directly and is actually what can be termed a "constructed" index, one which assumes a meaning because a decision has been made to give it that meaning.[2]

7. The effect of that dimension is assumed to be independent of any other dimensional effect of a given parameter of the same component. This assumption is crucial for the process of synthesizing or combining subtask and task values to secure an equipment or system value. The independence assumption is related to the assumption that interactive effects between parameters cancel out. If they do, then independence is a viable concept.

The assumption of independence permits application of the "product" rule (i.e., multiplication of the individual equipment/task probabilities. $R_{opi} = r_{t1} \times r_{t2} \times r_{t3}$, etc., where t_i is the ith task; $R_{t_n} = r_{s_1} \times r_{s_2} \times r_{s_3}$, etc,, where r_s = step reliability. $R_{s_n} = r_{c_1} \times r_{c_2} \times r_{c_3}$, etc., where c = component. R_{p_n} = the reliability of the selected relevant dimension, where p = parameter. This means that the reliability of any operation i is a function of the individual reliabilities of the tasks comprising that operation. In turn the reliability of task i is a function of the reliabilities of the individual components used in that task; the

2. All of which raises a question that has both theoretical and practical interest. Constructors of data banks must consider what the relationship of the parameters included in their banks is to predicted task performance.

reliability of the component is a function of the parameters that apply to that component; and the reliability of each parameter is assigned on the basis of whichever dimensional value is selected as applicable to that parameter. Figure 3 summarizes the process.

In general, workers in the field have not accepted the independence assumption. Behavior is, if anything, interdependent rather than independent. However, independence is often to be found underlying man-machine prediction models because the complexity posed by interdependence is very difficult to handle with present modeling concepts and data.

One of the side effects of the independence assumption is that each additional task, component and parameter decreases the predicted reliability of the operation. Moreover, one element with a substantially reduced reliability will significantly lower the estimated reliability of the entire operation, even if all the other elements are quite high. This has highly undesirable effects in attempting to calculate the performance reliability of a complex equipment or system. As the number of components and tasks increases, the multiplicative process tends to degrade the estimated reliability of the overall operation to values far below what one would realistically expect of that operation. We shall see similar effects manifested in TEPPS.

8. We shall discuss the output metric later, but it is instructive to consider what the probability value derived for a component or task really means. The metric is derived from errors made in the original experimental study, so that it is assumed that r = 1 minus the error frequency. As Regulinski (Ref. 4) has pointed out, this assumption applies only under very rigidly delimited conditions. But we should continue further to ask what the probability value means, even if the assumption holds. If the value reflects the probability that the act being described will be performed without error, then it assumes that every error on which the metric is based has significant effects on operator performance. This is highly unlikely because acts may often be performed erroneously and still be completed successfully. Many errors have non-significant effects (e.g., they do not fail the mission) or errors can be noted by the operator and the acts redone correctly. From that standpoint the Data Store methodology is not a completely efficient predictor of behavior.

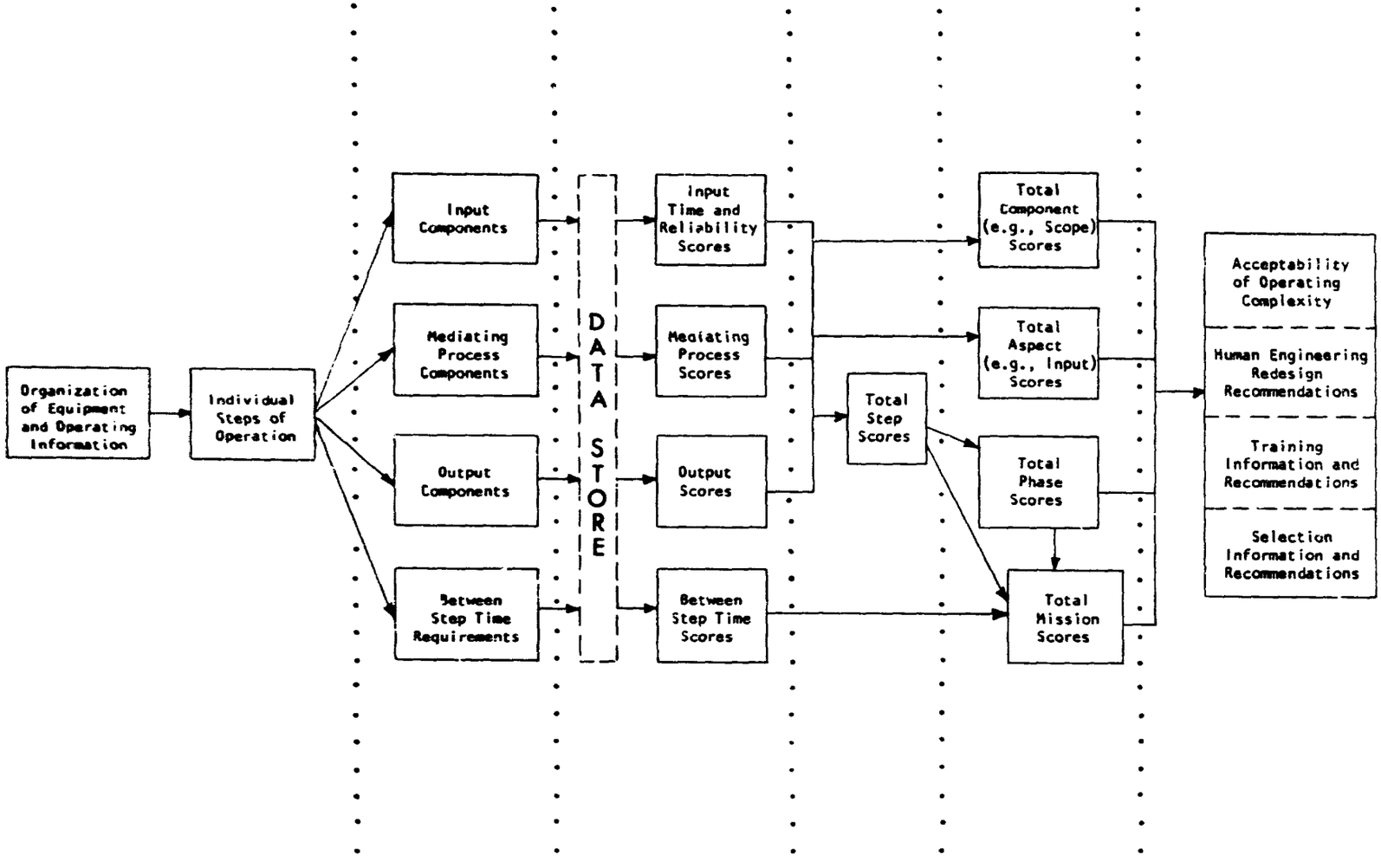


Figure 3. Graphical Summary of the Basic Evaluation Process

If we have gone into as much detail as we have in relation to the assumptions made by the predictive model, it is because similar questions must be asked by anyone developing a predictive model and a predictive data base. It is apparent from this review of Data Store assumptions that a data base does in fact imply a conceptual structure and methodology.

There are two major ways of evaluating the validity of a model: (1) conceptually, in terms of what has been termed construct validity or the reasonableness of the model assumptions; (2) empirically, in terms of the model's ability to predict variations in performance as a function of variations in model parameters. Ultimately, empirical validity is what the model developer must rely on for his justification. It is a question however, whether empirical validity can be achieved without construct validity.

Consequently, it is necessary for the developer in starting his work to ask of his model (or his data base) questions such as

(1) What is the conceptual structure implied by the body of data gathered and the manner in which the data have been gathered?

(2) Does that conceptual structure realistically reflect human performance?

(3) What effects do the limitations of data-gathering opportunities have on the compromises that must be made with relation to that conceptual structure?

We shall discuss validation studies performed with the Data Store later.

METHODOLOGICAL SCOPE

The Data Store can be applied to any equipment/system which contains controls and displays. The limiting factor here is the type of behavior involved in the system operation. The behaviors the Data Store deals with are discrete operations; it has difficulty handling continuous (e.g., tracking) and decision-making behaviors, the latter because the Data Store is very limited with regard to data on mediating processes.

However, this limitation is not peculiar to the Data Store; we will encounter it later in other models. The work of Irwin et al. (Ref. 1) has shown that the model can be applied to maintenance operations (although not troubleshooting, because this is largely decision-making).

PARAMETERS.

The major parameters utilized in this technique have already been noted. There are two classes of parameters: (1) structural parameters, describing the elements in the Data Store; and (2) process parameters. Structural parameters are aspects of behavior, components, parameters and dimensions, which have been defined in Table 8.

Process parameters are those relating to the various behavioral levels into which system operations must be analyzed. These describe the various units of behavior for which one is attempting to predict. Process parameters include the mission, the phase, the task and the behavior at progressively more molecular (detailed) levels of description. "The finest unit for which reasonable performance data can be established is the individual step, act or behavior" (p. 5, Ref. 3).

This unit corresponds to what others have called the task element or the simple stimulus-response act. The behavior or step is described by the individual stimulus to the operator and the individual response he makes to that stimulus. Obviously, if one is to include molecular equipment dimensions in one's predictions, the behavior predicted must be at a level commensurate with those discrete dimensions. Moreover, this is a level at which complex conceptual behaviors should not be required. If one throws a switch from one position to another, only memory is involved. (Although some would argue that memory is a complex conceptual behavior.)

There is comparatively little difficulty in determining what that behavioral level is, because it is usually the level at which operating procedures are written. In contrast to other predictive techniques this is actually an advantage; whereas it may be difficult to define the task level precisely, this is not true of the task element level.

It has other advantages also. If the task element is the irreducible substratum of behavior, the data applied to it can be applied to any system or equipment because every system/equipment will contain the same behavioral level (i. e., the same discrete task elements). Data gathered at more molar levels might be more difficult to apply across equipments.

DATA

1. Measures Used

There are two basic measures employed in the Data Store:

(a) Time to perform or time to complete the behavioral step or task element (in seconds). This is the absolute minimum time needed to complete the step. This time holds only if all the parameters listed for a given component possess optimum dimensions. If the dimensions are not optimum, an increment of time must be added to the base minimum. What makes a dimension non-optimum is the fact that experimentally it required more time to complete the task using a component with that dimension; in that sense, a non-optimum dimension is merely one which requires more time; the characteristic is not defined independent of the time measure. It should be noted that no range is provided for this time measure. Presumably this is because the times presented are minima. Ideally, however, where data are based on a distribution of empirically derived values, the range, e.g., σ, of the distribution should be presented.

(b) Performance reliability which in the case of the Data Store is derived from error data, is in fact equivalent to error data. For example, if 5 errors occur over 10,000 opportunitites for error, this gives an error percentage of .0005 or .9995. Where reliability is defined as 1 minus an error percentage, the resultant reliability can be thought of only in terms of error likelihood. There is nothing wrong with such a definition of reliability; it is necessary, however, to recognize that this measure deals only with a sub-class of all possible performances. That is, as indicated previously, it reflects only those situations in which the occurrence of an error in essence fails the mission, or in which one is not concerned with task or mission completion but only with error-occurrence. Again, the range of the reliability distribution is not indicated.

Because the error data from which the reliability measure is derived does not vary as a function time, this reliability measure is not equivalent to that commonly employed by reliability engineers, which is $e^{-\lambda t}$. It is, however, very similar to a measure of achieved reliability, which is essentially s/n, where n = the total number of attempts to complete a task and s = the number of successful attempts.

Despite the fact that the performance reliability measure described in the Data Store does not exactly correspond to the measure of equipment reliability, there is no reason why the former cannot be used meaningfully as long as the nature of the measure and what it represents is recognized. It should be understood that no "reliability" measure used in any of the prediction techniques to be described corresponds exactly to the measure of equipment reliability. The reason for this is the fact that the error data from which human reliability measures are derived do not take the time factor into account as does the equipment reliability measure. This also has implications for the combination of human reliability and equipment reliability measures (to secure a measure of system effectiveness); although the numerical form of the human reliability measure permits combination with the equipment reliability measure, it is a bit like combining apples and oranges.

The two measures (time and reliability) were derived independently of each other, and the reliability value (e.g., 9963) is not a performance reliability as a function of the time required to complete the step (e.g., 1.5 seconds). Although the two measures are independent, they should be somewhat related; i.e., as reliability decreases, there is a tendency for performance time to increase.

2. Data Sources

The error and time data used to derive Data Store values were secured from 164 experimental studies in the literature. There has been some criticism of the Data Store because its data sources have been so few; but apparently this represents the distillation of several thousand research reports. The author has found in his own research (Ref. 2) that many studies do not contain data translatable into the probability metric and must therefore be discarded (if the metric is to be employed); and many of these studies are not well defined in terms of control-display dimensions. This casts some doubt on the feasibility of using the experimental literature as a data source; but it should be noted that the developers carried their literature search only through the late 50's, and there is now an additional 10 years worth of literature that might be used as a data source.

Data defined in terms of relatively molecular equipment dimensions must be secured from carefully controlled experiments. Field operational

testing situations (e.g., as in military exercises) do not supply data describing these dimensions because in the less well controlled field test situations the subject responds not to the individual equipment dimension but to the entire equipment/task complex comprising a multitude of interdependent factors. In non-experimental testing it is almost impossible to separate out mediating processes and individual components, and to ensure that error effects are truly independent. Only a laboratory set-up will provide the type of data suitable for a Data Store type of data structure.

1. Model Output

The output of the Data Store (e.g., a reliability value) can be viewed in two ways:

(a) As the probability of correct performance by personnel in operating specified equipment;

(b) As a figure of merit for the equipment when operated.

Because the performance reliability metric output by the Data Store takes into account only a few of the factors influencing human performance, and is primarily oriented to equipment characteristics, the author feels that the Data Store is more appropriately used in the sense of (b) rather than (a). The reasons why we say that the Data Store reflects human performance only partially have been presented earlier: the inability to account for molar (central) processes, for interdependence of behavioral parameters and for the factors that determine task completion success. For example, it should be noted that the Data Store performance reliability does not take into account differences in the tasks to be performed with an equipment. It makes a difference to performance reliability whether the operator is merely monitoring a CRT display or classifying the stimuli on that display. Consequently, if one wishes to be a purist, the meaning of the Data Store measure in terms of (a) is inadmissable. These reservations apply of course only to the "construct" validity of the Data Store model. If empirical validation studies (i.e., comparison with actual performance values) were to demonstrate a high degree of relationship between Data Store predictions and actual reliable performance, one could forget these objections.

As a figure of merit reflecting the operability of an equipment (which was, we feel, the original intent of the methodology), the objections raised previously do not apply. Any index may be constructed of whatever elements one wishes, and if they do in fact serve to differentiate different equipment configurations, that is all that one asks of the index. The figure of merit index does not imply a model of actual performance, whereas the other does, in which case the latter must be considered in the light of its relationship to known data on human performance.

The understandable desire of workers in the field to predict human performance rather than merely equipment operability (however important the latter is) has led to a confusion between the two meanings of the Data Store output. This confusion has been intensified by the form of the Data Store output which on the surface (but only on the surface) seems to be the human equivalent of equipment reliability. In any event, because of this confusion, more has been expected of the Data Store than it can reasonably be expected to supply.

All the reservations mentioned previously do not apply to the time measure. On the other hand, much less use has been made of the time measure than of the reliability output.

PROCEDURES FOR MODEL APPLICATION

1. Procedures for Analysis

The process by means of which one analyzes mission/operations into subtask or step elements has been described previously. It is no different from what one does in any application of task analysis methodology.

Operations which may be considered akin to functions or gross tasks must be broken down to individual tasks; tasks must be broken down to steps or task elements; each task element must be analyzed in terms of all the components involved; then the components involved must be broken down to parameters; and finally the appropriate dimensions of each parameter must be selected from the Data Store.

A number of questions need to be asked: (a) how accurately is it possible to determine which tasks belong in which operations and which task elements in the task; (b) how accurately can one determine which components are involved in each task element? The breakout to the task element level can be accomplished with reasonable accuracy, assuming clear, detailed operating procedures and skilled analysts. A problem arises with regard to determining which components are involved in the task element, particularly for perceptual tasks, because even though the operator is responding (perceptually) to a single display (mainly), he may be aware of (or in part responding to) the entire complex of other displays. For example, if he is monitoring one of several meters on a display panel, his response to that meter may be affected by the presence of the other meters. This is, in part, taken care of by the parameter category, (e.g., number of lights in visual field), but in many cases, this information (which depends on the operator's mode of perceptual response) is lacking. The same problem does not exist to the same degree for the response mechanism.

There are, however, potential sources of confusion in the analysis process as a result of the feedback parameter. Two types of feedback are recognized by the Data Store:

> "a. Directing Feedback. This type of information resulting from operator performance serves only to direct or aid control manipulation. Stimuli or information that primarily aid control manipulation, such as labels or associated display features, are assessed as parameters of that control.....
>
> "b. Initiating Feedback. Feedback which signals the end of one step of behavior and serves as the input to the next step, such as an indicator light or scale value, is assessed with the step of behavior it initiated. Such feedback may be ignored, however, ..."

However, there are other types of feedback, which are much more intractible, e.g., that which indicates the degree of success with which a task has been completed. Some consideration will be given to this type of feedback in later discussions of other predictive techniques.

In general it can be said that feedback does not pose much of a problem for the Data Store user because he tends to ignore it, at least in its more complex forms. No attention is paid in the analysis to other

modifying task parameters, such as speed and precision requirements, exposure time, etc., for two reasons: they are difficult to identify with specific equipment components and the experimental literature does not provide much data on them.

One task factor that is explicitly noted is perceptual shift. This refers to the time required for the operator to shift his attenition from one control or display to another. "Normally the time required for perceptual shift can be ignored, since it is of very short duration..."

2. Procedures for Synthesis

What we mean by synthesis is the re-combination of the individual step and task reliabilities and times to derive values for larger behavioral units and for the system as a whole. This procedure is based on simple multiplication for reliabilities and addition for times. As Smith et al. (Ref. 5) point out, time addition is entirely reasonable and the error of estimate for times is small in terms of its effect on the total prediction. The multiplicative procedure is much less defensible (because it ignores the interdependence of behaviors) and because of the multiplication an error in predicted reliability at an element level will have compounding effects on the accuracy of the total prediction.

3. Data Application

Reliability estimates and times are applied only to the individual component dimensions. Values for subtasks, tasks, phases and missions are secured by multiplication and addition, which assumes that any interactive effects among these subtasks, tasks, etc. are minimal.

ANTICIPATED MODEL USES

1. Prediction of Human Performance

A good deal has already been said about the limitations of the Data Store as a method of predicting operator performance. We would prefer not to use the Data Store in this sense, although there is some evidence (to be discussed later) for its utility as a predictive model.

Should the Data Store be used as a predictive device, it should be reserved for tasks involving relatively simple discrete control-display operations, in which the reservations noted earlier are less pressing. The Data Store cannot handle predictions of contingency events or those involving conceptual or decision-making operations. For example, if alternative contingency or decision pathways involve the same equipment, the user will get the same prediction in either case.

Can one in fact use the techniques to predict operator performance? Yes, but only if one recognizes the tremendous simplifying assumptions one must make. This simplification does not matter quite so much in the design use of the model.

2. Design Analysis

The Data Store is particularly adapted for problems encountered in detail design. Because of its emphasis on equipment characteristics, it may be used to

(a) compare alternative design configurations

(b) select components.

In comparing alternative designs the procedure is to calculate the performance reliability/completion time for each configuration and to select the one with the highest reliability and the shortest completion time. In the selection of components the engineer would presumably select a component with those dimensions providing the highest reliability. In such a selection the actual performance reliability of the dimension is less important than its value relative to other dimensions. Thus, if one component dimension has a reliability of .9999 and another .9996, one would presumably select the first.

The Data Store is considered to be of particular value in generating redesign recommendations. Here the procedure is to identify those comparisons that contribute most significantly to total mission time and reliability. These then become candidates for redesign.

The Data Store will supply information useful in making function allocations (i.e., deciding whether or not a given task should be automated) but only when equipment details are specified. It supplies no information for deciding on the manpower required to operate an equip-

ment, because the methodology is not geared to differences in number of personnel unless they use different equipments. However, a formula is included in Irwin et al (Ref. 1) for including the effect of one operator checking another's performance.

3. Training

The Data Store does not pretend to provide information which is applicable to a total training program. Rather it applies only to those tasks and components which contribute most to total mission scores and which are not amenable to redesign. These would presumably receive additional emphasis during training.

Because the Data Store dimensions relate solely to equipment, it is difficult to see how much it can contribute to the specification of required training. Certainly the nature of any required training, its content or duration could not be inferred from the Data Store.

4. Selection

The comments on training apply also and even more so to selection of personnel. Selection relates to aspects of behavior, and these are phrased so grossly that they supply little information. For example, if it is found that the greatest contributors to mission reliability are inputs, one could recommend that personnel be selected on the basis of visual or auditory aptitudes. The specific nature of these aptitudes would not be indicated by the Data Store and would have to be inferred from other sources.

VALIDATION/APPLICATION STUDIES

Different types of validation should be considered in the evaluation of any predictive method:

1. Construct validity or the determination that the model contains those factors critical to operator performance. Although the developers say that construct validity of the index "seems assured", our examination suggests this is the case only if one is willing to accept highly simplifying assumptions.

2. Content validity or the extent to which the model contains all those parameters and dimensions needed for operator performance. This has been shown not to be the case.

3. Empirical validity (some would prefer the term "predictive validity") or the extent to which Data Store predictions correspond to actual operator performance. Empirical validity can be determined in several ways:

(a) by comparing Data Store predictions against observed/measured performance of operators

(b) by comparing Data Store predictions against subjective estimates of operator performance made by experienced personnel

(c) by comparing Data Store predictions against ratings of equipment operability by experienced personnel; this indicates the extent to which the index differentiates among various equipments.

The best estimate of empirical validity is by comparison of predictions against observed/measured performance. This was not possible during the development of the Data Store. However, Irwin et al (Ref. 1) performed a validation based on comparisons with observed performance and found that observed reliability was .9989 as against a prediction of .9972.

The developers determined validity by comparing the predicted time and reliability with rankings of equipment in order of complexity (method (c) above). Note that this validity index measures equipment operability rather than operator performance. There was fair agreement between the Data Store scores and the rankings.

There is consequently some indication that the Data Store methodology provides valid estimates of equipment operability and human performance, but the evidence is far from satisfying. We shall see in later discussion that although predictive methods are developed, they are often not validated so that it is difficult to arrive at a realistic estimate of the utility of these methods.

Inter-user reliability, i.e., the consistency of Data Store scores developed by various users, was found to be high.

The Data Store has been applied to at least one system development project the author is aware of (and probably many more about which he has no knowledge). Recently the author was talking to a human factors specialist who was involved in the system development of the P3C aircraft. The Data Store was used to develop predictions of response times and the error likelihood of personnel acting as crew of this aircraft. A mockup study indicated good correspondence between predicted and actual response times, although apparently no data were collected concerning error responses. The Data Store in this application was used to create a sort of baseline with which actual personnel performance could be compared.

A point that should be noted is that at present the Data Store is the only standardized data bank available to system developers. Consequently they will tend to use it despite any inadequacies we may have pointed out. This is a strong argument for the development of formalized data banks.

SUMMARY EVALUATION

Validity As noted previously, formal validity studies have been performed and show a reasonable degree of correspondence between predicted and observed values. The Data Store is one of the few techniques for which formal validation studies have been attempted.

Reliability Controlled studies have been performed that indicate the method can be applied with reasonable consistency by various analysts. The rather structured nature of the technique tends to improve its use-reliability.

System Development Applicability

A. Comprehensiveness: The method is limited in its application to specific types of behaviors and tasks, principally to those of a control/display nature.

B. Applicability: The method is primarily useful for predictions related to equipment design features and hence is specailly valuable for design analysis.

C. Timing: The method can be applied to systems in early as well as later stages of design, provided that system design has been

detailed down to molecular equipment characteristics. Otherwise it can be employed only to systems in the later stages of design.

Model Characteristics

A. Objectivity: Requires very few subjective (i. e., "expert") judgments.

B. Structure: Assumptions and parameters underlying the model are reasonably well defined and explicitly described.

REFERENCES

1. Irwin, I. A. et al. Human Reliability in the Performance of Maintenance. Proceedings, Symposium on Quantification of Human Performance, Albuquerque, New Mexico, 1964.

2. Meister, D. and Mills, R. G. Development of a Human Performance Reliability Data System: Phase I. Final Report Contract F33615-70-C-1518, July 1971.

3. Payne, D. and Altman, J. W. An Index of Electronic Equipment Operability, Rpt. AIR-C-43-1/62-FR, 31 January 1962.

4. Regulinski, T. L. Quantification of Human Performance Reliability, Research Method Rationale. Proceedings, Workshop on Human Reliability, 22-23 July 1970, Washington, D. C.

5. Smith, R. L. et al. Subjective Judgment as a Means of Generating Corrective Maintenance Data. Final Report, F33615-69-C-1396, Integrated Sciences Corp., Santa Monica, Calif., (no date).

ADDENDUM

In reviewing the preceding description of the Data Store Dr. Altman made the following points (personal communication to author, 8 July 1971):

Some points which may be of interest to you for historic perspective are:

1. All of our original models for combining reliability estimates involved the use of interaction terms. In no case were we able to find anything like adequate data for estimating interactions. We reluctantly went to the simplistic multiplicative model because it was the only way we could avoid, in effect, saying to the evaluator that we could give him at least some rough guidance on the easier part of performance estimating, but that he was strictly on his own when it came to the tougher part.

2. We did not want the Data Store part of the Operability Index to be essential for the limited-use procedure developed for the Army--let alone a major export for generalized use. The Data Store was intended to provide guidance to the evaluator when he could not obtain more appropriate performance data. In a sense, the Data Store was intended to be a final fallback position. Our qualms about the Data Store were largely a result of our conclusion that the existing behavioral sciences and human factors literature are mostly lousy for generalizing performance data.

3. Aside from our immediate commitments to the Army to come up with a practical tool, our main objective was to demonstrate the feasibility of combining equipment analysis and task analysis for purposes of human engineering evaluation. Although we had also hoped to find existing bodies of human performance data which would suffice to support good quantitative estimates when equipment and tasks had been appropriately analyzed and juxtaposed, review of existing literature quickly led us to essentially negative conclusions except for stopgap purposes.

4. My own feeling was that the results were highly favorable toward the notion of combining equipment and task analysis for human engineering evaluation. Given the essentially negative conclusions relative to the state of existing literature, it seemed to me that three priority developments were needed to bring human engineering evaluation to the minimum level required if it was to be a technology rather than essentially ad hoc or "art":

   a. New conceptual and data gathering approaches to establish nominal values for performance expectations.

   b. Better defined and articulated study of the effects of "conditions of performance," such as speed stress, boredom and fatigue, environmental stress, etc.

   c. Empirical study of individual differences, both training-experience background and ability variables.

5. I would consider models for combining performance estimates to be essentially part of the task description and analysis problem. A proper task analysis should tell one what the performance components are and how they relate to each other. Much more can be done to make task analysis more relevant to and a more powerful tool for performance forecasting. The most immediate need is for approaches to task analysis that will support the description of tasks involved in generating a given body of performance data. That is, I feel task analysis has been much neglected as a tool for specifying "experimental" conditions.

6. Since the time of our work on the Operability Index, I have felt that human factors needed to develop new approaches to generating dependable performance data--approaches which might have quite disparate philosophical bases from the usual psychological experiment. Over the years, I have made a number of

suggestions for such approaches--from use of synthetic tasks with randomized assignment of characteristics to incremental building of a data base from in-depth study of a small number of tasks and addition of new tasks only when sufficiently well understood to permit prediction of their performance within specified accuracy for stated conditions.

II. THERP-TECHNIQUE FOR HUMAN ERROR RATE PREDICTION

INTRODUCTION

THERP is probably the best known of what can be termed the "human reliability" predictive techniques, having been described and commented on in a number of texts, including the author's (Ref. 3) and in evaluative reports (Ref. 1). For that reason, perhaps, we have given it what may appear to be a more intensive scrutiny than the others. It is a method for predicting human error rates and for evaluating the degradation to a man-machine system likely to be caused by human errors (in association, of course, with other interactive factors like equipment reliability, procedures, etc.). Although historically[1] the method was associated with and initially built upon the AIR Data Store methodology, it has since developed independently of that methodology. THERP is one of a number of techniques strongly influenced by equipment reliability concepts.

Its procedure involves 5 steps which are repeated until the system degradation resulting from human error is at an acceptable level.

1. Define the system or subsystem failure which is to be evaluated.

2. Identify and list all the human operations performed and their relationships to system tasks and functions.

3. Predict error rates for each human operation or group of operations.

4. Determine the effect of human errors on the system.

5. Recommend changes as necessary to reduce the system or subsystem failure rate as a consequence of the estimated effects of the recommended changes.

These steps are considered "typical of the usual system reliability study if one substitutes 'hardware' for 'humans'" (p. 7, Ref. 4)

---

1. Dr. Swain, the primary developer and exponent of THERP, comments:

   "THERP is not an extension of the AIR model. All we did was to see the AIR Data Store as a source of data; but we later changed our minds and no longer use it...

NOTE 1 (Continued)

Following is how THERP originated. I arrived at Sandia in February of 1961 and began to do human engineering work on nuclear weapon systems. It quickly became apparent that my recommendations for equipment design sometimes were not accepted because I could not tell the designers how much benefit (in quantitative terms) my recommended changes would provide to the system. After L.W. Rook arrived at Sandia in May of 1961 he and I got to talking about the need for quantifying human performance influences on system reliability. We were naturally very much influenced by the reliability model used at Sandia as we were part of the Reliability Department and were in the Systems Reliability Division of that department. We saw that if we could find measurable human behaviors analogous to measurable equipment behaviors we could use the conventional reliability model. In other words, we were looking for p_i's, where p_i is the probability of human error. Reliability problems at Sandia generally were time-independent (or essentially so), though the model could handle time-dependent events as well. But generally, the problem was stated something like the following: What is the probability that Equipment Item A will function in its intended mode when called upon to do so. And, generally, these equipment items were one-shot devices or at least devices with a very limited life span so that considerations such as mean-time-to-failure were not appropriate. So we thought of human behavior in much the same way: What is the probability that Human Behavior A will occur correctly when it is supposed to occur? Often Human Behavior A had to occur within a very definite time frame. In such cases the question was changed to: What is the probability that Human Behavior A will occur correctly within Time Span T? Usually Human Behavior A could be thought of as occurring completely correctly (in terms of its system consequences) or completely incorrectly. But if there were degrees of correctness, we would fractionate the question (or, in terms of tree diagramming, provide different branches) to provide estimates of degrees of behavior correctness. In nearly all of our work (and in the Sandia applications to date) this type of fractionization has not been necessary to answer the system design or planning problems with which we had to cope.

In addition to the influence of conventional reliability technology, we were influenced by Herman William's article: Williams, H. D. "Reliability Evaluation of the Human Component in Man-Machine Systems," Electrical Manufacturing, 1958, 61(4), 78-82. Williams had been at Sandia Labs and had participated in the first human reliability analysis done by anyone, so far as we know. This 1952 classified study is described in SC-R-66-906 by Swain (also see p. 688 of Human Factors, Dec. 1964). Although we went much farther in developing the THERP model than the William's approach, we owe him a debt of gratitude.

In those early days THERP as a model was not influenced by any other outside work. Sometime late in 1962 someone sent us a copy of the AIR Data Store and the related three documents. We evaluated their model as too simplistic and not at all appropriate to practical human reliability work. But we were happy to see all the data on human performance in the AIR Data Store which could be used to derive our p_i's. (One minus the AIR figures gave the error equivalent for a small lump of behavior, and is described in SCR-685.) So we used the AIR Data Store rather extensively for a few years, but never used their model and THERP was never influenced by it. Therefore, to say that THERP is an extension of the AIR model is quite incorrect.

About 1964 Rook was struck by the narrow range of reliability figures in the AIR Data Store and did the Monte Carlo analysis which I reported in SC-R-68-1697. Rook left Sandia in 1965 and I continued to develop the THERP model. Since then Rigby has added to the model (SC-R-68-1875 and SC-R-69-1208). (End Note 1)

GOALS

The goals specified in the basic document describing THERP are:

(1) To derive "quantitative estimates of the degradation to a man-machine system resulting from human error" (p. 2, Ref. 4).

(2) Or, "to evaluate the human error contribution to system degradation" (p. 7, Ref. 4).

(3) To predict human error rates (p. 8, Ref. 4).

Since the fifth of the steps listed in the Introduction to this section is to "recommend changes" (to the system), another goal might be phrased as

(4) To determine those design changes to the system necessitated by the system failure rate.

In general, the goals we saw applied to the AIR Data Store-to predict the time and reliability (accuracy) of operator performance, to identify design features which degrade operator performance and to aid in selection and training- also apply to THERP, with the possible exception of selection and training. However, there is a flavor in THERP of application to the system rather than to the individual operator and equipment, as in the AIR Data Store. In this connection one might note THERP's concept of evaluating the human error contribution to system degradation. Moreover, in a recent letter to the author from Swain, the developer indicates that not only behavioral data are included in the application of the technique, but also equipment failure data, environmental factors, and other non-human events as well. This is also indicated in the footnote on p. 7 of Ref. 4, and in a secret report (Ref. 8). Whereas the Data Store methodology on which THERP is based provides a figure of merit for the individual equipment, THERP is an effort to achieve a true system measure. If equipment data are included in the use of the technique, then one must consider THERP as a technique for predicting system performance. For example, in the study described in Reference 1, the probability was estimated that environmental and other factors would preclude successful radio transmission from a ground station to a pilot in the air. However, all the reports describing the technique focus largely on prediction of human performance.[2]

2. Although human reliability predictive models aim at predicting "system" effectiveness, in most cases the prediction is carried out only as far as the operator subsystem. Theoretically, once an operator performance prediction is available, it can be combined with that for the equipment subsystem to provide an overall system value. However, because the primary focus of human factors specialists has been on human performance predictions, and because the combination of human with equipment performance predictions is not as easy as a simple multiplicative relationship would imply, the combination has not often been performed. See Verdi, A. P. The Manned Orbital Laboratory (MOL) Man-Machine Effectiveness Model, Appendix B to Report SM-51356, Human Engineering Program Plan, Dunlap and Associates, Santa Monica, California, 15 March 1966, under Contract AF04(695)-904 for the MOL System Program Office
for an example of an attempt to perform this combination.

ASSUMPTIONS

There is some difficulty in determining the basic assumptions underlying THERP, because THERP is very pragmatic about its assumptions. For example, "THERP makes no assumption as to the dependence or independence of behaviors. It is up to the user of THERP to make these assumptions....." (p. 12, footnote, Ref. 4) presumably he is skilled in behavioral technology.

Presumably, analysis of the specific system operations being evaluated will indicate whether these are dependent or independent; and their treatment in probability equations will follow this determination. From this we can infer that THERP assumes both independence and dependence of behaviors and applies the appropriate treatment as necessary.

In contrast to the Data Store methodology, instead of being restricted to molecular equipment design features, THERP assumes that many factors, many of them very molar, influence behavior. Among these factors (Ref. 7) are various psychological and physiological stresses, training, motivation, and situational factors, etc.

These factors must be taken into account in the gathering of error rate data and the error estimates derived should be modified in accordance with the presumed effect of these factors on performance. One difficulty that arises, however, in accounting for these molar factors on performance is the difficulty of recognizing their influence and estimating the extent of that influence. We shall deal later with THERP's procedure for quantizing the effects of these factors.

If we examine the hypothetical example provided in Reference 4 (p. 28) dealing with pilot communication over a radio, we see that the probability estimates are applied to relatively gross events, e.g., pilot will perform his operations on the ground, high stress condition occurs for pilot's air operations, etc. If operational data are available for these events or tasks, the analytic procedure does not proceed to any level more detailed than this. However, if operational data are not available, it is necessary (1) to conduct the necessary laboratory studies to obtain useful data, (2) to collect error rate data from operations which have tasks similar to the tasks for which such data is needed, or (3) to derive error rate data from expert judgment, preferably involving the use of psychological scaling.

To the purist there might appear to be some inconsistency about the combination of data from various sources, such as laboratory data and expert judgment data. For example, it is quite possible that various types of data may have different errors of estimate; but in most cases these errors of estimate will be unknown.

The explanation of this apparent discrepancy seems to lie in the heuristic nature of the technique. Whatever assumptions need to be made to secure data to exercise the technique will be made, on the very reasonable premise that in the absence of more definitive data or ground rules, one must do the best one can with what one has.

It is only fair to note that many of the techniques reviewed in this report make use of all available data, whatever their source. Those data that appear to be most valid or firmly based on experimentation are preferred, of course; but should such data be lacking the user of the technique will rely on less desirable sources.

One of the consequences of the technique's heuristic orientation is a certain degree of reliance on expert judgment which one finds in the use of the technique. In contrast to the Data Store, whose data are secured from experimental sources, THERP speaks about "estimations" of error rates or "judgments are made"(p. 9, Ref. 4). Inevitably certain assumptions are implied in such judgments, however, the very precise rules for securing expert judgments that one finds in TEPPS are not noted in THERP.

(Swain notes that his reports contain several references where details concerning the gathering of expert judgments are pointed out. Nevertheless, the author still feels that THERP's procedures for gathering expert judgments are not spelled out as they are in TEPPS.)

Because of the greater flexibility in data sources permitted by THERP, the possibilities for securing applicable data are expanded. Data may be secured from empirically observed performance of personnel; from historical records of operational performance; from "expert" judges. The manner in which these data are combined or in which human operations are assumed to interact is determined by expert judgment (as reflected in the analysis leading to the development of the probabilility tree) as well as by reliance on the experimental literature.

An initial step in the procedure is the listing and identification of the human operations to be evaluated. As a consequence, THERP makes use of the same analytic technique used by the Data Store and most other techniques reviewed, i.e., task analysis. However, the technique does not require breaking system operations down to the task level, although in actual practice this is often done.[3]

---

3. Swain comments: "Our technique does not require breaking system operations down to the task level, but we usually do. As a matter of fact, we usually get down to the step level, where a "step" is a typical step in a procedures document which is reasonably well human engineered -- that is, where a step consists of approximately one S-O-R unit. Example:

> Step 1. Adjust the XYZ until peak voltage is indicated on the voltmeter and record the resistance from the digital multimeter on line 17 of Form ABC.

Generally, a step is one that has about 10 of the types of elements you would look up if you were using the AIR Data Store. But this is only a gross approximation.

Just to continue this example a bit, in the absence of hard data, we would normally assign a .01 error rate to this step. However, this .01 could be modified considerably depending on various performance shaping factors. Suppose, for example, that instead of a digital readout for resistance, the usual backward reading resistance scale would have to be used. We would have to crank in some extra error rate. How much? Depends on what data we have. Sandia has conducted some unpublished studies on reading of test equipment. But suppose there is no Sandia data, and we cannot find data any place in the literature (which I think we could). Now we would get down to the matter of judgment and more of subjectivity would have to be used. In similar cases, several of us have independently made our own estimates, each with a rationale. And we have argued and arrived at a committee consensus. That might shock non-practitioning purists..... But if we were off by a factor of 2 or even 5, it would be seldom that the overall outcome of the reliability analysis would differ to any important extent. And if more accuracy would be needed, we would have to run our own study -- as we have done several times.

The task analysis procedure is, however, required in addition to uncover all the possible human actions and procedures entering into the evaluation. These include contingency operations which might be substituted for required operations.

The necessity for considering all operations (including those of a contingency nature) makes the use of a graphic mode of describing or presenting these operations extremely desirable (see Figure 4). In this respect, THERP is similar to other techniques like TEPPS or the digital simulation models to be described later.

Practicality, however, requires that the number of operations to be considered be restricted. "At this point, the analyst ordinarily makes some restriction in the human operations to be considered further. . . . he drops from consideration those human operations for which it is apparent (our own emphasis) that no significant degradation to system. . . . failure rates would result as a function of their incorrect performance. . . " (p. 9, Ref. 4). Although the developer indicates that one must be conservative in dropping irrelevant human operations, it is apparent that certain judgments are required based on "expertise". Again this ties in with the heuristic elements of the technique.

The lack of information on the ground rules to be applied in making these judgments has certain implications for the user of the technique: on the one hand he is granted considerable freedom; on the other, he must develop his own criteria.

It was noted in our review of the Data Store that each error was assumed to have essentially the same effect on operator performance. THERP on the contrary requires that the effect of a given error on system performance must be determined if it is to be evaluated properly. Hence, one of the major steps in the procedure is to determine the effects of human errors on the system. This is the probability (F_i) that an error or class of errors will result in a system failure, failure being defined as mission non-accomplishment resulting from error or equipment malfunction. F_i is any failure mode of interest.

This determination is itself a prediction, since the presumed effect is something that does not invariably occur; in other words, it is a probability. The determination of error effects may be based on empirical data; or where these are not available, it may be based on expert judgment.

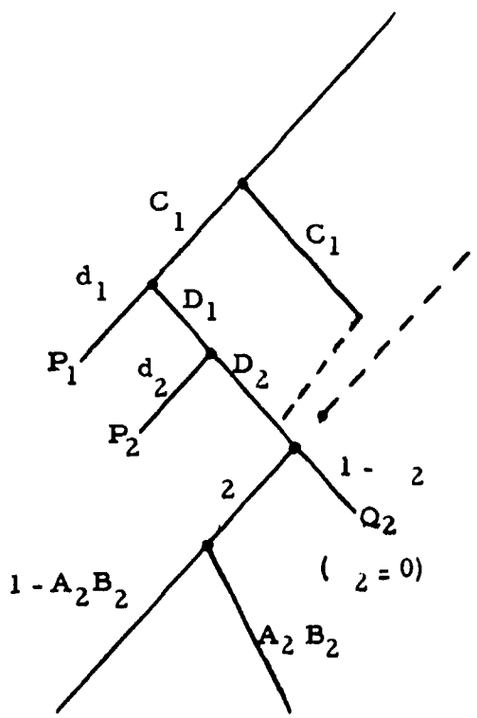


FIGURE 4. PROBABILITY TREE ILLUSTRATING BRANCHING TECHNIQUES (Taken from Swain, 1964)

(P = success. Q = failure. Small English letters represent human successes. Capital English letters represent human failures. Greek letters represent probabilities of events not under direct control of the humans in the system.)

Obviously F_i represents a degree of methodological sophistication which one does not find in the AIR Data Store. On the other hand, it permits again a certain element of subjectivity in THERP operations.

THERP recognizes, as do other techniques, that behavior is more interdependent than independent. In fact, a study performed by the developers (Ref. 6) reveals the inadequacies of the independence assumption. However, no explicit rules are given for accepting the assumption of independence or interdependence. This is left to the user of the technique. However, it is indicated that "often.... independence of certain behaviors can be assumed even when it is known that the assumption is incorrect. In such cases, it is judged that the resultant calculation... is sufficiently accurate for the purpose at hand." (p. 13, footnote, Ref. 4). This suggests that independence may be assumed in order to simplify the application of the technique. However, Swain notes that "we always make an evaluation of the need to assume lack of independence."

Again, the pragmatic element in the technique leads to a degree of subjectivity in its use.

Certain highly specific assumptions are made by the developers of THERP:

(1) One assumption deals with one operator monitoring another's performance. Here, an error probability (that he will not detect an error made by the second operator) of .15 is assigned to the operator monitor. This is based on studies of inspector accuracy in industrial assembly line situations. The .85 probability of inspector accuracy may seem a little high (in view of data supplied by Harris and Chaney, Ref. 2), but is certainly acceptable as a first estimate.[4]

4. Swain notes: The .85 figure is based on a series of rather practical studies, initially based on a review of the literature in SCTM-53-61(14) and later on studies reported in Ergonomics. This .85 estimate applies only to a passive type of inspection task where the actual defect rate is low (i.e., .01 or less) and where the inspection task is simple. Harris and Chaney obtained different results for different types of inspection tasks. For example, they show inspector accuracy rates varying from 15 percent to 75 percent as a function of equipment complexity. In our

(2) A second assumption relates to the self-correction of errors. "If a man has X probability of error for an important time-critical task on which errors occur infrequently, then his probability of making an error on Trial 2 (after he has made an error on Trial 1) is 2X, for Trial 3 (given errors on Trials 1 and 2) it is 4X. . . . . etc. , until the limiting condition of a 1. 0 error probability is reached." (p. 21, Ref. 4). The factor producing this doubling of error probability is stress. If stress is not anticipated, the error rate for both trials is assumed to be the same, on the grounds that any increased error probability in Trial 2 would be compensated for by greater attention.

(3) A third assumption deals with behavior under high stress conditions. The degree of stress is a parameter which apparently enters into all THERP calculations (as it does in a number of other techniques). "An estimate of 10-20 percent error rate for pilot tasks analogous to the critical behaviors of SAC pilots is felt to be reasonable. . . ." (p. 23, Ref. 4). Again, judgment is required in the determination of whether stress exists, the degree of that stress and the effect of that stress on performance, as reflected in increased error rates. Such judgments are necessitated by the inadequacies of the studies performed on this parameter.

Other very specific assumptions are made in the solution of a given system problem, but these assumptions are peculiar to the system situation and are not inherent in the methodology itself. These are "reasonable" assumptions which are developed for purposes of simplifying what would otherwise be an inordinately complex process. For example, the assumption is made that a written procedure used in an operation is the one supposed to be used and that it is correct. If such assumptions are not made, the range of possible error situations to be evaluated would become almost endless.

---

work dealing with assembly and shop type tasks done on nuclear weapons, we often use a .95 or .99 accuracy rate for the inspector, depending on the nature of the task. These latter estimates are for inspection tasks in which the inspector's role is not passive and in which he is looking for a limited number of clearly defined and recognizable defects. These various estimates merely indicate that estimates of inspector accuracy may differ considerably when different types of inspection tasks are considered.

The author would not like to be unfair in seeming to emphasize the number of judgments required in utilizing THERP, although other reviewers (e.g., Freitag, Ref. 1) have made the same point. Since the developer of THERP views his approach as being strictly an empirical one- "if it enables us to make predictions sufficiently accurate for the purpose at hand, we use it" (p. 17, Ref. 4)- the looseness in the conceptual structure underlying THERP is a necessary consequence of its pragmatic orientation. It is a defensible point of view that the lack of applicable data on human performance makes flexibility an advantage rather than a disadvantage.

The flexibility we have noted in THERP does have one pragmatic consequence with which we must deal. That consequence is that the THERP user is often left to his own devices (must develop his own ground rules) in making crucial methodological decisions (e.g., independence/interdependence; degree of stress, etc.). If he makes these decisions correctly he is "home free"; if he does not.....

Pragmatism - the acceptability of any method of arriving at any answer which solves a problem- is itself an assumption that must be examined. It must be recognized that this assumption makes it difficult to quarrel with the details of any methodology, as long as that methodology appears to "work". Since the author's point of view is that the purpose of human reliability predictive models is first to solve system development problems and only secondarily to serve as conceptual tools to explain man-machine behavior, pragmatism is acceptable to him, but only to the extent that it is buttressed as much as possible by detailed clarifying procedures.

METHODOLOGICAL SCOPE

As in the case of the AIR Data Store, any type of system, function or task can be handled by THERP, provided the error data for that system, function or task are available. Although the Data Store has limitations in terms of not being able to handle continuous type tasks (e.g., tracking) or cognitive tasks (lacking appropriate data and being constrained by its assumption of element independence), this limitation does not apply to THERP, simply because the conceptual and methodological structure is so flexible. Whereas the Date Store methodology is limited by its sources, the same cannot be said of THERP, because any set of "reasonable" data, from whatever source, permits one to apply a numerical value for any type of behavior.

This flexibility in data source is therefore an advantage, provided one accepts the validity of the data from these sources. In other words, for example, if the problem is to determine an error rate for a strategic decision involving three alternatives, the user of THERP would - in the absence of any other data - accept estimates (using paired comparisons, perhaps) by strategic planners of the error rate they have observed for this type of decision. If this error rate seems reasonable (on logical, empirical or experimental grounds) to the THERP user, he can include this kind of behavior in the class of problems which THERP can deal with.

In the same way, if he must take into account different types and amounts of feedback (a problem which, we will see, besets other methods), the user of THERP can develop a probability value for this feedback based on whatever data sources he can find. Consequently, THERP is not system/task-limited as are other methods.

PARAMETERS

Besides the customary response time and error measures, considerable emphasis is placed on stress in establishing an error rate for a given behavior. This parameter is defined in terms of personnel response to emergency situations, and can be categorized in terms of high and low stress producing conditions. In order to account for the stress factor data from SAC pilot situations have been adapted to provide an estimate of 10 to 20% error for the pilot and 5 to 15% error rate for in-flight tasks performed by aircrewmen.

What the error rate would be to account for stress occurring in situations other than these is somewhat unclear.[5] Presumably, if the need arose, applicable data would be secured. The stress parameter is

5. Swain notes: SC-R-69-1208 by Rigby and Edelman (see also pp. 475-482 in Vol. 10 of Human Factors) presents a later treatment of stress, and it is not restricted to flight tasks. Using former crew members of multiengine aircraft as subject matter experts, Rigby and Edelman, scaled the AIR data form the Ronan AIR report and note, "In the absence of better data or information to the contrary, it may be practical to apply these stress levels and error rates to non-aircraft situations. It is necessary to show only that the emergency of interest

also related to the increased probability of error in repetitive trials, that is, twice the probability of an error occurring in trial 2 when an error was made on trial 1, 4 times the probability of error for trial 3, given that errors were made on trials 1 and 2, etc. Although stress is not explicitly indicated as the cause of this increased error rate, it is implicit in the explanation given: the "error resulting from operator tenseness..." (p. 21, Ref. 4). Where non-stressful conditions do not exist, the error probability remains the same on repeated trials.

Obviously stress (or the lack of it) is a basic parameter entering into all THERP error probabilities. It should be noted that in contrast to Siegel's use of the parameter in his digital simulation model, stress in THERP is not operationally defined by and does not vary in terms of the specific system situation, but is rather a central (e.g., emotional) mediating process. The use of a standard error rate for stress (even though there are apparently 3 categories of stress: high, low and none) in less precise than Siegel's use of the parameter.[5A] However, Swain (personal communication) indicates that the 2X error rate relationship for repeated trials is analogous to the maximum value found in Siegel's theoretical stress distribution.

In his most recent work (Ref. 7) Swain has pointed out the importance of what he calls "performance shaping factors" (PSF) in determining or at least explaining error rates. Many of these are extremely molar, e.g., motivation, training, psychological and physiological stress, etc., and require sophisticated judgments for their measurement. A number of 7-point rating scales have been developed to quantize PSF. Swain specifically emphasizes task difficulty, personnel redundancy (treated earlier) and manner of use of performance aids. With the exception of

---

is indeed comparable to some point on the scale. This can be done by judgment alone, where that is necessary. Empirical estimates can be obtained by having subjects (1) insert the given emergency into the scale and use the mean position or (2) compare the given emergency to two or more judiciously chosen items on the scale and convert the obtained proportions to scale positions. The validity of such results, of course, will depend upon the degree to which the subjects are experienced in both the aircraft and non-aircraft situations."

5A. Swain notes: In our earlier work, we used three error rates for the three stress levels (see SCR-685), but Rigby and Edelman have developed a scale which has five stress levels (see SC-R-69-1208).

personnel redundancy, very little use is presently made of these "performance shaping factors" because the data relating to them are not readily available.[6]

It is our impression that although Swain recognizes the importance of various parameters for the determination of error rates, these parameters with which THERP deals are not defined with very great precision.[7] Although this is in line with the empirical, heuristic orientation of the method is is unfortunate because the parameters which impact on human performance are extremely complex.

---

6. Swain notes: We do make use of (i.e., consider) all the performance shaping factors listed in my various reports. It is true that data of the best experimental study caliber is not available for these PSFs, and besides that, they tend to be situation-specific anyway, but any human reliability analyst worth his salt must consider them. My data store report (SC-R-70-4286) notes that when one has error rate data which one wishes to apply to reliability predictions for some set of tasks, one has to judge how comparable the PSFs are between the tasks of interest and the tasks for which the error rates are available. To the extent they are comparable, then one has to modify the error rate data to estimate the influence of the PSFs. How is this done? Largely judgment, helped by whatever studies there are which show in general how different levels of one PSF affect task performance. (We) are trying to work out a more rigorous method in this regard, but there will still be a lot of "expertise" involved. We hope to answer your correct statement that these PSFs "are not defined with very great precision." Our general approach will use the 7-point scaling of PSFs, some examples of which are given in SC-R-70-4286.

7. This is a situation which one finds in almost all the models reviewed.

DATA

1. Measures Used

THERP employs two primary measures: the probability that an operation will lead to an error of class i (P_i); and the probability that an error or class of errors will result in system failure or failure of that part of the system being evaluated (F_i). P_i is based on what is termed an "error rate", which is the frequency of error occurring during a block of time. It is not error occurring as a function of the effects of time (as in the sense of equipment wearing out over time), but error as a function of number of repeated trials, e.g., 5 errors occurring over 100 trials gives an error rate of .02. There is some question whether error rate is analogous to equipment failure rate, since the latter is more directly a function of time.

P_i then is an error rate transformed into an error probability simply by deriving a percentage of error occurrence. $1 - P_i$ is the probability that the operation will be performed without error. One can derive an error rate from a probability of successful task accomplishment simply by subtracting that probability from 1.0, e.g., 1.0 - .9998 = error rate of .0002. Similarly, a probability of success can be derived by subtracting the error percentage from unity (e.g., 1.0 - .0002 = .9998). In that sense probability of successful task accomplishment and error likelihood are mirror images of each other. F_iP_i is the joint probability that an error will occur in an operation and that that error (or class of errors) will lead to system failure. $1-F_iP_i$ is the probability that an operation will be performed that does not lead to error and consequent system failure. $Q_i = 1 - (1-F_iP_i)^{n_i}$ is the probability of one or more failure conditions existing as a result of class i errors occurring in n_i (independent) operations. When one simplifies the mathematics the measure employed is simply an error rate (P_i) modified by an effect probability (F_i).

Both P_i and F_i are point estimates, that is, a single value for a task; they do not consider a range of values (in the sense of a standard deviation), nor is there any confidence level associated with these estimates. There is no distribution of P_i or F_i as a function of other parameters such as time, stress, etc., to which one can refer. In that sense we can think of them as being essentially "static" values.[8]

---

8. Swain comments: You are quite right in implying that normally we use point-estimates for our estimated error rates and forget about

2. Output Metric

The output metric is the failure rate associated with the system or part of the system being evaluated; it is not a probability of successful performance, although that can easily be determined by calculating 1 minus the system failure rate. To determine the probability that any given task will result in an error leading to failure, the measure is 1 minus the probability of no failure or $1-(1-F_iP_i)^{n_i}$ which is Q_i. Total system or subsystem failure rate resulting from human error is expressed as $Q_T = 1-\left[\prod_{k=1}^{n}(1-Q_k)\right]$ where the quantity in brackets is $(1-Q_i)(1-Q_q).....(1-Q_n)$.

The output metric is therefore simply the combination of the individual F_iP_i's for the individual task behaviors, the combination being based on conventional probability theory. Where the tasks are assumed to be independent, the combination is performed through simple multiplication. Where the tasks are interdependent the probability calculation becomes a bit more complex. We do not propose in this review to repeat the very lengthy probability equations required to arrive at a system failure rate. These can be found in the basic reference (4).

---

distributions and confidence levels. But the THERP model can handle distributions and confidence levels, and I have used it for this purpose. The problem with distributions (unless you have an empirical distribution based on lots of trials) is that they are so iffy.

There is nothing in the THERP model that precludes the use of Monte Carloing, but given the usual iffy data, why gild the lily? What I have done on occasion is to use +1 S.D. error rates, based on estimates of mean error rates, an assumed normal curve, and an assumption of a range ratio (Wechsler's) of 3:1... The use of +1 S.D. error rates in a tree diagram and the subsequent mathematical treatment gives one a type of "worst case" method. But even that is unnecessarily complicated for many purposes. I have obtained judgments of error rates and deliberately inflated them by an order of magnitude to see, even with this inflation, if a certain horrendous system outcome would result. If one is still safe, then the analysis need go no farther. If, on the other hand, there is an unacceptable probability of the unwanted system outcome, then one has to buckle down and do a more detailed analysis.

3. Data Sources

As indicated earlier, THERP will accept data from any source. The fact that error data describe the task, rather than equipment characteristics as in the case of the AIR Data Store, makes it much simpler to gather the requisite data. On the other hand, the fact that the error data do not directly describe equipment characteristics, although the latter are considered in the assignment of an error rate, may make THERP's estimates somewhat insensitive to equipment design variations.

The determination of F_i presents the THERP user with an entirely different situation. Because the effect of an error on the system is involved, F_i probability estimates are largely system-peculiar and therefore must be developed anew for each system being considered. For example, depending on the way in which circuits are hooked up, an error may be either inconsequential or catastrophic. This is one of the reasons why individual reliability "failure modes and effects" analyses are often performed during system development. It would therefore seem unlikely that one could develop a "universal" data bank describing F_i error effects. Here the user must depend largely on the judgment of the system developer (e.g., reliability specialist). One can only point out that these F_i estimates may involve some inconsistency between "experts".

The concept of F_i is, however, a definite plus for THERP; without it every error is equivalent to a task or system failure, which would be difficult to justify.

The fact that empirical, experimental and subjective probability estimates are utilized indiscriminately in developing THERP estimates may lead to error in determining these estimates. (Swain claims no more accuracy in prediction than a factor of five, p. 12, Ref. 4). Moreover, regardless of the data source, each probability estimate is considered as equivalent to every other, which simplifies their combination. Swain notes, however (p. 12, Ref. 4) that the assumptions behind each data estimate are provided so that the user of the reliability study will not be deluded into assuming an accuracy not present in the estimates. No rules are provided for securing expert judgments as one finds in TEPPS, for example.[9] Again, however, THERP's pragmatic orientation is important here; only that accuracy needed to solve a particular problem is required, and the implication is that for most practical system development problems a high degree of accuracy would be necessary.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Methods

The system operation to be analyzed is defined in terms of a potential error/failure of that operation. In other words, operations that are unlikely to lead to error need not be considered. This considerably simplifies the analytic procedure, but one must be conservative in doing so, Swain warns us. Although the system operation to be evaluated may be at any level (the total mission, a general function (e.g., take off, landing) or a specific task (e.g., communicate message to tower), in practice whatever level at which the operation is described, it is reduced to a set of discrete tasks needed to perform the operation. Consequently, the analytic method used is that of task analysis. No particular task analysis variation or specific analytic procedures are specified; in this respect THERP is similar to all other predictive methods.

All the human operations (i.e., tasks) to be performed are identified and listed, including contingency events or decision-alternatives. In this last connection THERP is a considerable advance over the Data Store. Note that although these tasks may not be formally analyzed (in terms of written checklists, etc.) in terms of such factors as time pacing factors, stimulus exposure rates, accuracy requirements, conceptual demands, etc., the analyst is supposed to consider these in terms of his evaluation of task difficulty or stress level.

Analysis is combined with the development of a probability branching tree (Figure 4) which is a graphic means of representing steps in the operation and contingency events. The probability tree is necessary for the user to keep in mind the relationship among the alternative pathways. The reader should compare the probability tree with an analogous device to represent system operations, the Graphic State Sequence Model of TEPPS.

---

9. This subsection should not be concluded without mentioning the psychological scaling method frequently employed in THERP and described more fully in Reference 5. The method itself is too lengthy to include in this report, but Dr. Swain has kindly provided an example of ho v the method might be used for deriving error rates for SAFE-GUARD (system under development) tasks. The example is provided as an appendix to this subsection. The reader is cautioned that this appendix is based only on a draft paper.

2. Methods of Synthesis

Two methods of combining data are employed. The first (simple multiplication) is used when tasks are believed to be essentially independent. The second (somewhat more complex) is used when the analyst decides that the operations to be combined are interdependent.

The interdependence of operations is handled by analysis of the branches in the probability tree. If, for example, the branches indicate that the error probability of task C is a function of the combined error probabilities of tasks A and B, then this relationship is transformed into mathematical form and the equation is solved. Dependency relationships are decided upon by the user who makes up the probability branching tree. In any event, THERP's claims to deal with interdependency appear correct.

Such interdependency relationships refer only to the probability that a task will or will not be completed successfully. In other words, they deal only with a terminal binary state- successful accomplishment of the task or failure to accomplish the task. It is not clear whether any predictive method can handle degrees of successful accomplishment, e.g., the situation in which a preceding task A can be performed more or less accurately and the degree of accuracy in A will determine the successful accomplishment of following task B.[10]

3. Data Application

Data are applied at two levels: the task element level, but more commonly at the task level.

---

10. Swain comments: THERP can handle degrees of successful accomplishment by the use of the branching technique. No problem here. But we haven't found it necessary to go to such a level of complexity. I have always preferred to use binary decisions in my analytic work... As long as one has to make subjective judgments, it's easier if they can be fractionated into a series of yes-no judgments. This is one reason I prefer the paired-comparison scaling technique. And it's one reason my tree diagrams may look complex but really make the judgments in the reliability analysis much simpler. (There are just more of them.)

ANTICIPATED MODEL USES

In this subsection we consider the uses that may be made of the various models. The following are the major outputs that a model may have:

(1) Prediction of system effectiveness

Under this heading we include the following:

a. Determination that the system will perform to specified requirements. For example, given a system requirement, e.g., that the mission must perform with a reliability of .9995, can the system (including the human operator, of course) perform the mission with that reliability?

b. Prediction of the absolute level of efficiency the operational system will achieve (whether or not an operator performance requirement has been specified). For example, assuming that the system under development is built to present design, what is the numerical reliability that the system will accomplish its mission (e.g., .9995, .9870, .8993)?

(2) Design analysis

Under this heading we consider the following as possible uses of the model:

a. Comparison of alternative design configurations in terms of the effect on operator capability.

b. Suggestions for redesign of a system configuration (assuming that the model has been exercised on a system and has indicated potential deficiencies).

c. Suggestions for initial design of the system configuration.

(3) Manpower selection

Under this heading we ask whether a model will suggest the types and numbers of personnel needed to operate the system, and in particular the aptitude/skill level required of these personnel.

(4) Training requirement

We may also ask whether a model supplies information relative to the content or duration of the training required to support the system.

The first model application is quite obvious: this is what predictive models should be able to do. The meaningfulness of the remaining applications may, however, not be so immediately apparent. In the design analysis area, if two or more system configurations are available, between which a choice must be made, applying the model to each configuration should produce a prediction of system effectiveness (as in 1 b) which can be used as a figure of merit. The two predictions can then be compared and the configuration with the estimated greater effectiveness can then be selected (all other design parameters being equal). Since this application is only a variant on (1), any model should be able to do this.

If a model can predict, it should also be possible to use it as a diagnostic tool, since the factors responsible for a prediction should be discernable upon detailed examination of the model exercise. If equipment deficiencies are responsible for a lowered estimate of effectiveness, then those deficiencies should be available from the various model outputs. From these the analyst might infer design changes that could improve the system. Although this does not mean that the model will output human engineering recommendations as such, the more sensitive the model is to the equipment factors that influence effectiveness, the more apparent the items that should be changed will appear. Similarly, if manpower and training factors are responsible for a lowered prediction of system effectiveness, then the model should provide clues in this area also.

It should be apparent that any prediction must be caused by one or more factors entering into that prediction, whether it be equipment, the number of personnel, their aptitudes, training, etc. For a model to be maximally effective, those factors should be deducible from the model outputs. It would be unreasonable to provide the user simply with a numerical estimate and let it go at that; few users are likely to be satisfied with this.

Obviously the list of model uses represents an ideal set of uses. Models will vary in the degree to which they can satisfy these uses. Moreover, they may be specialized for one use rather than another. It is essential, however, to examine a model in such terms because the model is only a tool to provide answers, of which the above represent the most important.

One last point. It may seem unreasonable that a predictive model should be used for initial design analysis, in the sense of suggesting particular components, design layouts, etc. To the extent, however, that a model makes use of a data bank for assignment of error probabilities/completion times, that data bank can also be used as a tool for initial design analysis. If, for example, task A, involving certain behaviors and components X, Y and Z, has a substantially lower error probability than task B involving other behaviors and components L, M, N, it would make sense for the system designer in any situation in which he can use either task A or task B, to select task A. That is what we mean by initial design analysis.

1. Prediction of System Effectiveness

THERP's intended purpose is to evaluate system operations in terms of the effect of human error. A comparison is made between the predicted system failure rate and the performance required by the system. In the course of doing so, it predicts what the performance of those system operations will be- at least that part of it representing the human's contribution. The obverse side of the prediction of human error is the prediction of human success. Manifestly, if the technique is valid, it is a powerful instrument for prediction of system effectiveness.

2. Design Analysis

What one eventually does with the prediction in (1) above is to recommend changes to the system. If, after having predicted the system's failure rate due to human error, that failure rate is unacceptable (in terms of a specified requirement), then the technique should lead to recommendations for system redesign.

Here one enters an entirely new ball game. The question one must ask is whether THERP indicates as an inherent part of its technique what changes should be made. For each system operation (task) being evaluated a Q_i value is derived which represents the error probability associated with the operation. If, then, one ranks the tasks in terms of their Q_i values, changes should be made in those possessing highest error probabilities. After the changes are made, Q_T is recomputed; if the changes are effective (i.e., Q_T is lowered to an acceptable level), the technique has achieved its goal.

The determination of what changes should be made are not, however, readily apparent in THERP. THERP points out the task needing redesign, but does not suggest what those changes should be. Remember that the task description with which the error probability is associated contains comparatively little information about equipment or task characteristics (at least at a molecular level). For example, in the example cited in Reference 4, the pilot's reception of the code was found to be the weakest link (p. 48). The possible system redesign changes noted were "to increase the reliability of the pilot's reception, to bypass him, or.... (eliminate) the requirement to receive the code in the air" (p. 48). These are all logical alternatives, but they provide the system developer with relatively little guidance, assuming that the pilot must receive the code in the air and cannot be by passed.

The point we wish to make is that THERP, like many other models, pinpoints a source of difficulty but goes no further. To make the task more reliable by redesigning it requires attention to equipment and task characteristics which THERP may not provide because of its level of description. Any recommendations for redesign would have to arise from an independent human engineering analysis of the task and equipment.

This is not to say that the guidance THERP provides in pointing out the operation requiring attention is unimportant. What we do say is that the guidance may not be at a level detailed enough to make it very usable to the system designer. In this respect THERP is similar to other performance measurement and predictive methods.

If we emphasize this point it is to suggest that a technique specialized for prediction of performance may not necessarily be the best for design needs (except at a very gross level of design). Performance prediction is one sphere of activity; design guidance is another, and the two may not easily cross over.

THERP should do well in comparing alternative design configurations, provided that the configurations to be compared involve different task operations, rather than the same operations with different items of equipment. In essence, THERP develops a prediction of each alternative configuration and then compares predicted system failure rates. Selection is made of the one with the lower failure rate.

Where the same tasks are performed, but only the equipment configurations change, it is unlikely that significantly different error probabilities for tasks will be derived, unless the equipment changes are quite marked.[11] If the task is communication via radio, will the error probabilities differ significantly unless the two radio sets are vastly different? They will not, unless the error probability estimates input to the technique are particularly sensitive to equipment details.

THERP makes no pretense to supplying recommendations for initial design, since it makes no use of a standard data bank, as does the AIR Data Store.[11A]

Obviously the error rates derived for the various tasks are in part determined by equipment characteristics. However, unless one could partial out that part of the error rate dependent on specific equipment/task characteristics, it might be difficult to work backwards from the error rate to a concrete design recommendations.

3. Manpower Selection

THERP makes no claims as a technique to aid in the selection of personnel. Since the level of task description does not permit easy interpretation in terms of required aptitudes, nothing can be done in this area. However, the user might infer from his knowledge of the system what personnel aptitudes etc. would be needed.

4. Training

THERP makes no claims as a technique to predict training needs. However, should a task be found to have an unacceptable error rate, and redesign is not possible, it might be assumed that additional training of that task would be required. Such a recommendation would, however, do little more than point out the need for training.

---

11. Swain notes: It is indeed true that significantly different error probabilities for tasks will not be derived when equipment changes are moderate or minor. This is one of the findings that really hit us. And it was true whether we used the AIR Data Store or any other source of data or just plain expertise. The important point is that a great many human engineering niceties don't make an important difference in effects on system reliability under most operating conditions.

11A. Swain insists, however, that THERP can be used for initial design.

At this point the expertise of the analyst would undoubtedly come into play and he would determine the additional training requirements from the characteristics of the system. The point, however, is that the model as a model would not suggest these changes unless it were highly sensitive to training parameters.

VALIDATION/APPLICATION STUDIES

It was pointed out in the section on the Data Store that it is possible to evaluate the validity of a model on either conceptual and/or empirical bases. Obviously, if a model has demonstrated empirical validity, by predicting effects which are in fact demonstrated operationally, then the question of conceptual validity becomes irrelevant. However, no model has completely demonstrated empirical validity (although in a few cases highly indicative experiments have been performed). In the absence of demonstrated empirical validity, additional confidence can be secured in a model by examining the concept structure on which the model is based.

Because a significant characteristic of THERP is that it appears to be tailored to the system problem which it addresses, it is difficult to evaluate the methodology on the basis of conceptual validity. Consequently, empirical validation studies are crucial if one is to make an appropriate judgment of the technique.

A problem that one runs into, however, in establishing the empirical validity of a model which was derived from and is applied largely to problems arising during system development is that it is difficult to establish a controlled experimental design to measure validity. Moreover, it is difficult (although not impossible) to differentiate the purely predictive aspects of a model from its deductive aspects, e.g., from the human engineering design changes that are recommended and implemented.

To the extent that a model should find its use in system development, it is possible that one would not wish to differentiate predictive from deductive aspects. It is also entirely possible that for such a model the concept of validity should be subordinated to that of utility. Swain indicates "our own experience shows that it (THERP) is a practical, workable and useful method".

The difficulty of making appropriate judgments about THERP is intensified because the systems to which THERP has been most often applied are highly classified and therefore information concerning the success of THERP in dealing with these systems is not available to the author.

It is known that THERP has been and is being applied to a number of system development projects, e.g., Safeguard, but again any reports developed during these projects have been unavailable to the general reader because of security classification. With the exception of the illustrative example presented in Reference 4, therefore, which is not a validation study, there have been no illustrations, except in passing references, of how well the technique has worked, although the developer has indicated in conversations that the method works well.

The case of THERP, like that of the Data Store, and many other models, illustrates the fact that all of the documentation and validation one would wish to have is not available. Like the Data Store, THERP has apparently been applied quite often, but data have not been provided by its users to others working in the field to determine just how promising the technique is.

EVALUATIVE SUMMARY

Validity - Although only one formal validation study has been reported (to the author's knowledge), the method has been applied repeatedly to system development problems and is reported to work reasonably well. In fact with the exception of the Air Data Store, this is the only method that has been applied repeatedly in system development, although most of that application has been within the developers' own shop (Sandia Laboratories).

Reliability - No formal data are available on consistency with which the method can be applied, although the developer reports no difficulty in this area.

System Development Applicability

A. Comprehensiveness: Model can be applied to all types of equipments, tasks, behaviors.

B. Applicability: Method outputs a prediction of system effectiveness and can, with the aid of standard human engineering techniques, be used for design analysis.

C. Timing: Model can be applied to systems in the early as well as later stages of system design.

Model Characteristics

A. Objectivity: Judgments are required and the ground rules for securing these are not as detailed as one would desire.

B. Structure: Conceptual structure is somewhat informal.

REFERENCES

1. Freitag, M. Quantification of Equipment Reliability: I. Review of the Recent Literature and Recommendations for Future Work. TM-940, U. S. Navy Electronics Laboratory, San Diego, California, 2 June 1966.

2. Harris, R. and Chaney, W. Human Factors in Quality Assurance, Wiley, 1969.

3. Meister, D. Human Factors: Theory and Practice, Wiley, 1971.

4. Swain, A. D. A Method for Performing a Human-Factors Reliability Analysis. Report SCR-685, Sandia Corporation, Albuquerque, New Mexico, August 1963.

5. Swain, A. D. Field Calibrated Simulation. Proceedings of the Symposium on Human Performance Quantification in Systems Effectiveness, Naval Materiel Command, Washington, D. C., January 1967, pp. IV-A-I through IV-A-21.

6. Swain, A. D. Some Limitations is Using the Simple Multiplicative Model in Behavior Quantification. in Askren, W. B. (Ed) Symposium on Reliability of Human Performance in Work, AMRL-TR-67-88, Aerospace Medical Research Labs., Wright-Patterson AFB, Ohio, May 1967.

7. Swain, A. D. Development of a Human Error Rate Data Bank. Paper presented at the U. S. Navy Workshop on Human Reliability, Washington, D. C., July 1970.

8. Swain, A. D., "Human Factors Associated with Prescribed Action Links (U)," (CONFIDENTIAL article), Proceedings of the 11th Military Operations Research Symposium (MORS), Naval Analysis Group, Office of Naval Research, Washington, D. C., Spring 1963, 217-230 (SECRET report).

APPENDIX I

OUTLINE PROCEDURE FOR DERIVING ESTIMATES OF ERROR RATES FOR SAFEGUARD TASKS

(Source: SC-R-67-1045 and references cited therein)

1. Have subject matter experts develop ordinal scales of HERCULES tasks in terms of error-likeliness.

   a. Subject matter experts are those with familiarity and experience with HERCULES tasks. These experts will also rank-order SAFEGUARD tasks.

   b. The method used will depend on the number of tasks to be compared. Ideally, paired comparison would be used, but from a practical standpoint, some kind of shortcut will be employed.

   c. There will be at least two separate scales: one for maintenance tasks and one for command and control (C&C) tasks. These two sets of tasks will be rated separately and this separation will be maintained throughout the procedure outlined in this paper. C&C tasks will be rated, assuming a nuclear conflict has started. Maintenance tasks will be rated, assuming peacetime conditions. It may be desirable either to rate them, assuming nuclear conflict, or apply a correction factor to the peacetime rating. This decision will be made later.

   d. A critical part of the rating method will be the development and standardization of the instructions to the raters. The intent of the instructions will be to get each rater to be making his judgments on the same dimension, error-likeliness, and with the same scenario in mind.

   e. Check inter-judge agreement. Interview outliers to see if they were using different criteria in spite of above detailed instructions. Discard results from outliers as appropriate. (This discarding is obviously fraught with methodological pitfalls, but is necessary from a practical viewpoint, and is consistent with the literature on psychological scaling.)

   f. Classify HERCULES tasks in terms of amount of inter-judge agreement as a means of establishing our confidence in task rankings. This classification can be used as an aid to the project personnel and help quantify "subjectivity" in some of the final decisions that will have to be made.

2. Convert the two sets of rankings to two interval scales of error-likeliness, one for maintenance tasks and one for C&C tasks. Procedures for the conversion are found in the cited references.

3. Obtain actual error rates from graded exercising of HERCULES. The stress level is high for tasks performed in HERCULES exercising and can be considered equivalent to the stress of military personnel in a Strategic Air Command Standardization Board (SAC Standboard) exercise where a man's military career can be made or broken, depending on his performance and the performance of the men under him.

4. Order the above error rates into two sets of ratio scales, one for maintenance tasks and one for C&C tasks.

5. Compare the two sets of interval scales obtained from expert judgments with the two sets of ratio scales based on HERCULES exercising.

   a. Try to resolve any major disagreements between the interval scaling based on the expert judgments and the interval scaling defined by the ratio scales. (Obviously, only the intervals in the scales can be compared because there is no zero point on the two scales derived from expert judgment.) It cannot be assumed that disagreement is the result of poor estimates of error-likeliness; it could be the result of inadequate error recording during the exercising. Fortunately, the White Sands project leader is very familiar with this exercising, having participated in several, and his judgments will be valuable.

   b. Discard those interval values from the judgment data or those ratio values from the system exercising records which are not consistent with the overall data.

   c. There will likely be some interval values based on expert judgment for which there are no corresponding error-rates obtained from system exercising. Retain these interval values unless it can be judged that they are clearly inconsistent with the overall data.

6. Convert the two HERCULES interval scales into ratio scales using the procedure outlined in SC-R-67-1045, "Field Calibrated Simulation." Basically, this conversion involves melding an interval scale into a ratio scale such that the same relative intervals between data points on the interval scale are approximately maintained while changing the absolute distances between these data points to be in accord with those data points on the existing ratio scale. The actual error rates on the existing ratio scale are known as key stimuli in this process. (In actuality, the conversion is done algebraically rather than graphically.)

7. Using the SAFEGUARD system and task analysis, meld the SAFEGUARD tasks into the two HERCULES ratio scales. There are different ways of accomplishing this melding. Ideally, we would do all of them as a study of methodology to compare the end results. At the very minimum, we should employ Procedure 1 (the graphic line procedure) and one of the other procedures.

   a. Procedure 1. Use a graphic error-likeliness scale for each set of HERCULES tasks (i.e., maintenance and C&C) and have the experts place the SAFEGUARD tasks on the appropriate scale. Neither of the graphic HERCULES scales would have any actual error rates on it, but would simply show the relative (i.e., interval) distance between the error-likeliness of all HERCULES tasks on that scale. Each judge would indicate relative position of each SAFEGUARD task on the appropriate scale, and his indication would constitute an interval scaling of the SAFEGUARD tasks which would also be a direct ratio scaling. The same procedure would be used to check inter-judge agreement as was used in rank-ordering HERCULES tasks. The ratio scales from the judges would be averaged to develop two combined ratio scales, one for maintenance and one for C&C tasks.

   b. Procedure 2. Using the above experts, develop two interval scales of SAFEGUARD tasks in the same manner as the interval scales of HERCULES tasks were developed. Then the project personnel as a committee would judge which SAFEGUARD tasks are behaviorally similar to which HERCULES tasks and meld the SAFEGUARD interval scales into the appropriate HERCULES ratio scales.

c. Procedure 3. Interval scaling of SAFEGUARD tasks would not be done. The project personnel as a committee would judge which SAFEGUARD tasks are behaviorally similar to which HERCULES tasks and place the SAFEGUARD tasks directly on the appropriate HERCULES ratio scales. (This approach would be the easiest to implement, but would obviously involve the most risk.)

d. Procedure 4. Using the above experts, throw all the HERCULES and SAFEGUARD tasks into two pots, one for maintenance tasks and one for C&C tasks, and have them develop two combined interval scales in the same manner as the two HERCULES interval scales were developed. These two combined interval scales would of course be directly translatable into ratio scales. (This approach would be the most difficult to implement, but might be considered to involve the least risk.)

8. Periodically re-scale the SAFEGUARD tasks as significant design changes are made.

9. Compare estimated error rates with "real" error rates as it becomes possible to exercise parts of the SAFEGUARD system. The use of quotes around "real" reflects the need to judge the validity of the Safeguard exercising and not blindly assume that because it is exercise data it is automatically better than the data derived from expert judgment.

APPENDIX II

July 1971

List of Relevant
Human Factors Publications
by Sandia Laboratories

(By Publication Date)
(Note: These reports have been released to the public.)

1. Rook, L. W., Reduction of Human Error in Industrial Production, SCTM-93-62(14), Sandia Labs., Albuquerque, N. Mex., June 1962.

2. Swain, A. D., Altman, J. W., and Rook, L. W. Human Error Quantification, A Symposium, SCR-610, Sandia Labs., Albuquerque, New Mexico, April 1963.

3. Swain, A. D., A Method for Performing a Human Factors Reliability Analysis, SCR-685, Sandia Labs., Albuquerque, N. Mex., August 1963.

4. Swain, A. D. "Human Factors in Design of Reliable Systems," Proceedings of the Tenth National Symposium on Reliability and Quality Control, Institute of Electrical and Electronic Engineers, New York, January 1964, 250-9. (Also SCR-748)

5. Swain, A. D. THERP, SC-R-64-1338, Sandia Labs, Albuquerque, N. Mex., August 1964.

6. Rook, L. W., "Evaluation of System Performance from Rank-Order Data," Human Factors, 1964, 6, 533-536. (Also SC-DC-64-1119)

7. Swain, A. D. "Some Problems in the Measurement of Human Performance in Man-Machine Systems," Human Factors, 1964, 6, 687-700. (Also SC-R-66-906)

8. Rook, L. W. Motivation and Human Error, SC-TM-64-135, Sandia Labs, Albuquerque, N. Mex., September 1965.

9. Swain, A. D., "Field Calibrated Simulation" Proceedings of the Symposium on Human Performance Quantification in Systems Effectiveness, Naval Materiel Command and the National Academy of Engineering, Washington, D. C., January 1967, IV-A-1 - IV-A-21. (Also SC-R-67-1045)

10. Swain, A. D. "Some Limitations in Using the Simple Multiplicative Model in Behavior Quantification," W. B. Askren (Ed.) Symposium on Reliability of Human Performance in Work, AMRL-TR-67-88, Aerospace Medical Research Labs., Wright-Patterson AFB, Ohio, May 1967, 17-31. (Also SC-R-68-1697)

11. Rigby, L. V. "The Sandia Human Error Rate Bank (SHERB)," R. E. Blanchard and D. H. Harris (Eds.), Man-Machine Effectiveness Analysis, A Symposium of the Human Factors Society, Los Angeles Chapter, 15 June 1967, pp. 5-1 to 5-13. (Also SC-R-67-1150)

12. Rigby, L. V. and Edelman, D. A., An Analysis of Human Variability in Mechanical Inspection: Summary, SC-DC-68-2173, Sandia Labs., Albuquerque, N. Mex., May 1968.

13. Rigby, L. V. and Edelman, D. A. An Analysis of Human Variability in Mechanical Inspection, SC-RR-68-282, Sandia Labs., Albuquerque, N. Mex., May 1968.

14. Rigby, L. V. and Swain, A. D. "Effects of Assembly Error on Product Acceptability and Reliability," Proceeding of the 7th Annual Reliability and Maintainability Conference, American Society of Mechanical Engineers, New York, July 1968, pp. 3+12 to 3-19. (Also SC-R-68-1875)

15. Swain, A. D. Human Reliability Assessment in Nuclear Reactor Plants. SC-R-69-1236, Sandia Labs., Albuquerque, N. Mex., April 1969.

16. Swain, A. D. "Overview and Status of Human Factors Reliability Analysis," Proceeding of the 8th Annual Reliability and Maintainability Conference, Amer. Inst. of Aeronautics and Astronautics, New York, July 1969, 251-254. (Also SC-R-69-1248)

17. Webster, R. G. and Swain, A. D. "Human Factors Inputs to Large-Scale Field Tests," Synder, M. T., Kincaid, J. P., and Potempa, K. W. (Eds.), Human Factors Testing Conference 1-2 October 1968, AFHRL-TR-69-6, Air Force Human Resources Lab., Wright Patterson AFB, Ohio, October 1969, 35-59. (Also SC-R-70-4220)

18. Rigby, L. V. "The Nature of Human Error," Annual Technical Conference Transactions, American Society for Quality Control, Milwaukee, Wisc., May 1970, pp. 457-466. (Also SC-R-70-4118)

19. Swain, A. D. "Development of a Human Error Rate Data Bank," J. P. Jenkins (Ed.) Proceeding of U. S. Navy Human Reliability Workshop 22-23 July 1970, Naval Ship Systems Command, Office of Naval Research and Naval Air Development Center, Dept. of the Navy, Washington, D. C. February 1971, pp. 113-148. (Also SC-R-70-4286)

See Also

20. Swain, A. D. Design Techniques for Improving Human Performance in Production, Industrial and Commercial Techniques, Ltd., London (in press).

III. TEPPS - TECHNIQUE FOR ESTABLISHING PERSONNEL PERFORMANCE STANDARDS

INTRODUCTION

TEPPS is a set of procedures developed by Dr. R. B. Blanchard and his co-workers for using analytic and probabilistic techniques which organize and employ (1) system operational requirements data, (2) system descriptive data and (3) human capability data or estimates of these data. The system to be evaluated is described by the Graphic State Sequence Model (GSSM) which is a function flow-oriented diagram identifying the various ways in which system requirements may be accomplished. The basic behavioral unit of the GSSM is the PEF (personnel-equipment functional) unit which is analogous to the task. The GSSM describes system input, output and intervening states (note the resemblance to the S-O-R framework employed by THERP). The GSSM is transformed into the Mathematical State Sequence Model (MSSM) which uses conventional probability equations to describe the mathematical relationships among the units comprising the GSSM.

The steps involved in applying TEPPS (much simplified of course) are:

1. Describe the system to be evaluated.
2. Develop the GSSM on the basis of (1).
3. Determine predictive data (probability and time estimates) for the GSSM units.
4. Apply the predictive data to GSSM.
5. Develop the MSSM.
6. Perform quantitative analyses to derive a reliability effectiveness value describing the system.

GOALS

This model has several goals, which are more or less clear depending on the particular report that one reads. Basically TEPPS is

(1) A method for "deriving specific personnel performance standards with definite relations to system effectiveness requirements" (p. 10, Ref. 5). In this application the model is supposed to allocate (or apportion, as reliability engineers would say) pre-established system effectiveness requirements (SER's) among the personnel (and tasks) engaged in

performing system tasks. These permit the development of personnel performance standards which can then be used as indices for comparison with actual performance requirements (SER's).

In other words, if the SER for a system calls for a personnel performance reliability of .9874, and there are N personnel and tasks whose performance in interaction are required to make up .9874, then the technique assigns the individual performance requirements for those personnel and tasks so that .9874 can be accomplished. Depending on the interactive loops involved in the system, task 1 might have a reliability requirement of .9999, task 2, .9965, etc., all of which performance requirements when combined will provide .9874.

(2) The model is supposed to "yield a measure of system effectiveness" which is essentially the same as that produced by THERP. If one assigns predictions of performance reliability to the individual tasks in the system, what one derives is a measure of system effectiveness (i.e., the human part of it).

In the case of the "derivative" model (case 1 above), one takes the SER and breaks it up among various tasks: the probabilities associated with these tasks become performance standards. In the second use of the model (its so-called "integrative" use) the SER is unnecessary; one takes the tasks making up the system and, assuming a store of predictive data, assigns predictions to the individual tasks. It is this application which is of particular interest to this study and the one described in this section; however, the developers of the method suggest that this application is dormant because of the lack of man-machine data. From the standpoint of this report, however, the lack of data is no reason for not considering the model, since it is always possible that suitable data will become available, and in any event the model should suggest the kind of data needed.

On page 2 of Reference 6 a number of potential outputs of the model are listed which are of interest to us because these outputs imply model goals. Besides the two goals already specified, these outputs imply as goals of the technique:

(1) determining the quantitative effects of system design changes on system performance or on personnel performance requirements;

(2) determining system components most and least likely to cause errors;

(4) establishing standards for corrective maintenance.

The reader should compare these goals with those of THERP. The reference to system design changes and the determination of system components most likely to cause errors reminds us very much of THERP. The concept of establishing performance standards is something we have not so far encountered (and is in fact unique among the models reviewed).

In the course of its development, the use of the technique has been expanded. The developers indicate that "the technique provides a general tool for system analysis" (p. 11, Ref. 5).

The model goals, therefore, have specific design and maintenance as well as operator performance implications. The model should therefore have general applicability to most systems, although the developers concede that it cannot handle continuous and decision-making tasks.

"At present TEPPS most closely resembles an operability submodel", that is, it attempts to "predict the degree to which a man-machine system is capable of being operated. . . assuming that the system is ready and available" (p. 17, Ref. 5). It does not consider states resulting from system failure, hence it is not an availability model.

ASSUMPTIONS

To start off with some definitions, what TEPPS is either allocating or predicting performance to is a unit of behavior called a PEF (personnel-equipment functional) unit. As was the case with THERP also, the basic unit of behavior (PEF unit) is not well defined, but appears to correspond to what is generally called a task. Certainly it is not at the level of the task element but is more detailed than a function. Examples of PEF units (taken from p. 104, Ref. 6) are: VP observes target track (two or three points) and computes heading (within ± 7 degrees); RCO observes tote board and determines accuracy of target heading shown there, based on previously received, mentally recorded heading message.

Note that we are dealing here with behaviors which are relatively system-specific, which might make it more difficult for TEPPS to make use of a "general" data bank of the AIR type. Blanchard points out, however, that this would depend on the degree of behavioral generalization possible from such a data bank.

Note also that the tasks described in PEF units are not well described in terms of equipment characteristics, which bears negatively on the capability of the model to handle design aspects (i.e., to be able to predict differences in performance dependent on differences in equipment design). However, there appears no reason why the PEF unit could not be more rigorously defined. The reliance on expert estimates for the collection of predictive data (see later discussion) to be associated with PEF units makes it difficult, however, to define the PEF unit more precisely in terms of equipment characteristics, since it is unlikely that judges could discriminate adequately in terms of molecular equipment characteristics.

The System Effectiveness Requirement (SER) already referred to describes the required system output, generally in the form of a probability, a time or both. The term reflects only the system requirement imposed on system personnel.

The Index of Task Accomplishment (IOTA) is the probability of successful accomplishment of man-machine activities (probability associated with the PEF unit). IOTA's may be derived by observation or by subjective techniques (although TEPPS developers have made primary use of the expert estimate technique to derive these data, since direct observation or simulation is very difficult, according to Blanchard).

A basic assumption concerning time is that a maximum value exists for time (T_{max}) beyond which the probability of accomplishing a task successfully does not change. TEPPS' maximum probability of successful accomplishment is therefore assumed to occur at T_{max}. This assumption appears to be quite reasonable on the basis of logic and anecdotal evidence, although it is unlikely that one will find substantive confirmation in the experimental literature (because of the sparseness of that literature). Similarly, below some minimum time (T_{min}) probability of task accomplishment would be expected to be zero (activity cannot be performed). Presumably between T_{min} and T_{max} probability of successful task accomplishment increases as we go from one to the other.

The relationship between probability and time to accomplish a given activity is assumed to be lognormal. Probability is normally or lognormally distributed as a function of time. It is also assumed that the number of people correctly performing a task is normally or lognormally distributed over time; therefore it is also assumed that the function would be similar for a single individual and for combinations

of tasks. It follows also that failures to accomplish a task would be lognormally distributed over time. Within the area $T_{max} - T_{min}$ probability of task accomplishment is normally or lognormally distributed. This last assumption seems to imply that probability of a correct response within an individual is distributed normally or lognormally as function of time; however, it is reasonable to expect that as time approaches infinity performance probability reaches as asymptote and remains constant.

The assumption of a relationship between time and probability is extremely important, because it defines the distribution of P_j and thus permits a valid application of probability statistics for predicting performance, something not found in the AIR and THERP methodology. The question is, however, whether the actual distribution of probability over time is the lognormal distribution postulated by the following equation:

$$f(\log t_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp. -\frac{(\tau_j - \mu_j)^2}{2\sigma_j^2}$$

where j = a subscript designating one of n PEF Units

τ = log t, i.e., the independent variable of the normal distribution derived from the lognormal one

μ = mean of the normal distribution of τ

σ^2 = variance of the normal distribution of τ

The probability that the jth activity will be performed in time t_j, therefore is the integral of the above equation from $-\infty$ to log t_j. If IOTA is considered to be directly related to probability of accomplishment, and if failures to accomplish an activity are assumed to be lognormally distributed, then IOTA varies with time according to the following equation:

$$i_j = I_j \int_{-\infty}^{\text{Log } t_j} \frac{1}{\sigma_j \sqrt{2\pi}} \exp. \frac{(\tau_j - \mu_j^2)}{2\sigma_j^2} \, d\tau_j$$

where I_j = the highest value IOTA can be expected to achieve

i_j = variable IOTA, i.e., ranging from 0 to I_j.

Availability of estimates of the maximum and minimum times (T_{max} and T_{min}, respectively) within which an activity can be performed establishes the values μ and σ.

TEPPS also distinguishes between logical and functional dependencies. Logically dependent activities are those whose accomplishment necessarily implies the successful occurrence of some prior or concurrent activity. A logically dependent relationship assumes that the prior activity will be performed with a reliability of 1.0, since feedback from its performance will be complete and the subsequent activity will not be performed until or unless the prior activity is performed (which means a reliability of 1.0). A functional dependency is an interaction such that the degree of effectiveness of one activity is related to the degree of effectiveness of another. "For example, an activity may be performed poorly, thus affecting the performance (but not the occurrence) of another" (p. 8, Ref. 5). Failure to include functional dependencies will result in a system effectiveness estimate which is highly conservative (lower than actual).

There is an implication that, in contrast to what THERP says it can do, TEPPS cannot handle functional dependencies. "Unfortunately, functional dependencies cannot be treated in reliability or allocation models because their occurrence and degree of effectiveness cannot be determined without empirical investigations".... In general, therefore, the determination of system reliability will actually produce an underestimate..." (p. 8, Ref. 5). As we shall see later, this underestimation has posed severe difficulties for TEPPS (as well as other models) in its operational use.

One might wonder, however, whether it might not be possible to include functional dependency data in TEPPS as part of the IOTA information provided by the expert estimates used to secure TEPPS predictive data. Would it not be possible to refine the expert estimate technique (to be discussed later) so that some indication of the functional dependency of tasks could be provided by expert judges?

TEPPS also assumes the existence of redundancy, which occurs when two or more activities or activity sequences give rise to essentially identical outputs, the occurrence of only one of which is necessary for mission success. Two types of redundancy are postulated, which are termed "true" and "apparent" redundancy. True redundant events are independently conducted activities which yield the same output state,

like applying the emergency brake and placing the automatic gear shift in PARK while parking a car. Apparent redundancy is illustrated by two operators monitoring the same airspace via two scopes.

According to probability theory, true redundant events can be handled multiplicatively, as

$$P = 1 - (1-P_1)(1-P_2)\ldots(1-P_N)$$

For example, assume four redundant events having probabilities of .9, .8, .7, and .6. According to the equation above,

$$\begin{aligned} P &= 1 - (1-.9)(1-.8)(1-.7)(1-.6) \\ &= 1 - (.1)(.2)(.3)(.4) \\ &= 1 - .0024 \\ &= .9976 \end{aligned}$$

Apparent redundant events are not really redundant, but their effects approximate redundancy and can be treated as such. The equation above that TEPPS uses to account for redundancy implies independence, which is probably not correct if the activities are performed by the same individual or two individuals who interact with each other. TEPPS argues that the data collected for its use takes that interaction effect into account and that consequently use of the equation above will result in little error. This assumes that the data collected have sufficient fidelity to the modelled situation to be representative of second-order effects. If this is the case, there would seem to be no reason why the functional dependencies discussed previously (which are important in terms of developing an accurate system reliability value) could not be treated in the same manner.

The assumptions made by TEPPS could well be applied to THERP. In fact, there appears to be a high degree of similarity between the TEPPS assumptions and those of THERP, except that the conceptual structure of TEPPS is more rigorously elaborated than is that of THERP. We note, for example, that the PEF unit is essentially the same as the behavioral unit employed by THERP; and the need to account for logical and functional dependencies and redundancies must also be be accounted for by THERP. The relationship between probability and time (which we will also find in Siegel's digital simulation models) represents, however, a significant theoretical advance over THERP. We shall see also that the development of the GSSM and its

transformation into the MSSM is very similar to THERP's graphic probability branching tree and its use in probability equations.

In contrast to THERP, however, TEPPS makes no use of stress as a significant parameter. This is perhaps because the "integrative" (system effectiveness) model was derived from the allocation model which would have no use for a concept like stress (since it deals solely with the apportionment of reliability requirements). This is an inadequacy in TEPPS which could, however, be remedied without undue difficulty. [1]

METHODOLOGICAL SCOPE

In analyzing the types of systems, functions and tasks to which TEPPS can be applied it is necessary to distinguish between the qualitative analyses possible with the GSSM and the quantitative analyses for which the MSSM is necessary. Since GSSM is purely descriptive, it can handle any type of system, function or task, including maintenance operations. On the other hand, in a mathematical sense TEPPS cannot deal with the complex feedback loops involved in continuous tracking-type tasks and in decision-making tasks where there are "complex and unpredictable interactions among individuals" (p. 17, Ref. 6) because the state of the art in mathematical modeling (upon which the MSSM depends) is deficient.

TEPPS is therefore restricted largely to predicting the performance of discrete activities. In this respect the methodological scope of TEPPS is similar to that of THERP.

PARAMETERS

In the course of the preceding discussion we have covered essentially all the parameters that TEPPS includes in its conceptual structure. These include feedback, dependency relationships, redundancies, the inter-relationships between probability and time, and the behavioral

---

(1) Blanchard states: "This is actually not an "inadequacy" in that "theoretically" such effects would be reflected in input data if those data had been obtained under actual conditions or under high fidelity simulation. Otherwise, one would do as in THERP, and introduce an estimated "adjustment" factor. In short, the technique is thoroughly amenable to the consideration of the effects of stress and anxiety."

unit being predicted (PEF). All of these have been implied in previous models (and are in fact required by any modeling process).

TEPPS is more distinctive, however, when one considers the parameters it does not include in its conceptual structure. As indicated earlier, it does not deal with stress or any of the "performance shaping factors" that THERP considers necessary, or with molecular equipment characteristics such as those found in the AIR Data Store.

Blanchard points out (personal communication to author) that this is true primarily of the derivative use of the model, but not so of its integrative use. Presumably, if the input data used for prediction in the integrative model were to be affected by stress or other "performance shaping" factors, then the model would implicitly make use of these parameters. However, our comment refers to the fact that the model does not explicitly call out these parameters and relate them formally to predicted performance.

Whether this reluctance to include more complex parameters represents an inadequacy in the TEPPS model or a realistic appraisal of what can presently be accomplished in behavioral modeling is a value judgment we would not care to make. However, a defensible model of behavioral processes must ultimately take more complex parameters into account.

DATA

1. Measures Employed

There are two general measures: probability of task accomplishment and performance completion time. These can, however, be related to any kind of system-relevant measure, such as number of targets destroyed, radar detection range, etc. which means in effect that the model can handle many specific system measures. At the present time the TEPPS approach is based on the use of probabilistic statements in which the effectiveness measure is dichotomized into a pass/fail criterion. However, in anticipation of the future development of the model, TEPPS has explored distributed measures as well.

Probabilities are associated with particular effectiveness dimensions. For example, if the dimension of interest is radar detection range at 100 miles, a probability of .90 of performing detections at that range might be determined. In this situation the radar detection range has been dichotomized into 100 miles and beyond 100 miles.

Or if the dimension of interest were the number of types of targets to be destroyed (distributed effectiveness dimension), a probability could be associated with each number and type. Since the effectiveness of a specific man-machine activity is defined formally and logically in the same manner in which system effectiveness is defined, the appropriate dimension for some activities might be of the dichotomous type, whereas other might be distributed. In any event, the measure employed is highly flexible and poses no limitations for application of TEPPS.

2. Data Sources

Like THERP, TEPPS is highly data-dependent, perhaps more so TEPPS is conceptually more rigorous and elaborated than THERP, bu perhaps for that very reason, is more data-limited (or apparently so in the minds of its developers). Because of the pragmatic nature of THERP, already discussed, THERP will accept data from any source as long as it will provide a "reasonable" answer. This is not true of TEPPS.

TEPPS would accept empirical data (i.e., data based on observation of performance), but such data are comparatively rare. It rejects data available from the experimental literature, as one finds it abstracted in the AIR Data Store, and sees little possibility that such "universal" data banks will prove of use, at least in the near future (10 years). "It soon became apparent to us that the various laboratory data were not amenable to meaningful manipulations and that little could be gained by further attempts to extract data... We made such a conclusion... with the realization that it left us without a means of exercising our model... With regard to human performance data collected in the field, it is essentially non-existent in any generally useable form. The amount that is or may be available is probably insignificant for use in general man-machine models...." (Ref. 2).

As a consequence, TEPPS has made use of a fairly rigorous but complex paired-comparison technique to derive "expert" estimates of performance probabilities and time. We shall not endeavor to describe the process in detail; readers who are interested should read references 1, 5 and 6 (particularly 1). The essence of the technique is to provide judges with individual PEF descriptions and ask them to compare each description against all others to determine which has the higher probability of accomplishment.

The data resulting from the paired-comparison technique (IOTA) form a scale which provides meaningful distances or intervals between scale values. However, it is desirable to transform these scale values (which vary from around 3.0 to zero) into a more conventional probability scale with a range from 0.00 to 1.0.

If the transformation is accurate, the resultant probabilities can be used as estimates of the probability of accuracy with which tasks can be performed. (This is necessary for use of the model as a system measurement; much less so for the allocation use of the model.) The resultant probabilities assume performance under optimum conditions and no time constraints.

Transformation of IOTA to probabilities requires an absolute determination of the probability of task accomplishment for the two activities having the highest and lowest scale values. That is, these two probabilities are determined on an absolute (.9999, .9996) rather than on a comparative basis (Task 1 has a higher probability of being accomplished than Task 2). If at all possible, estimates of these two tasks would be determined empirically (i.e., through observation or from test data). However, it may be necessary to secure the absolute values by means of judgments.

In any event, having these two "anchor" estimates permits solution of the two simultaneous equations:

$$P_1 = 1\text{-}Ae^{-BS_1} \qquad (1)$$

$$P_2 = 1\text{-}Ae^{-BS_2} \qquad (2)$$

where P = obtained probability of one PEF unit
S = scale value of that PEF unit
A, B = constants.

The median of the probability estimates for the two tasks having highest and lowest probabilities of performance are inserted into equations (1) and (2) in place of P_1 and P_2. The equations are then solved.

The reader might note some similarity between this procedure and the scaling method adopted by THERP. Swain acknowledges a debt to Blanchard in this connection.

The transformation assumes that performance reliability is an exponential function of the scale values derived from the paired-comparisons.

The basis for this is the best fit regression analysis curve developed by Irwin et al (Ref. 4) using Data Store values and judges' ratings. This curve is exponential because the best fit was produced by plotting the scattergram on log paper.

Using the Irwin results as the basis for the equation relating IOTA to probability values appears to be somewhat flimsy. There are several reasons for this: (1) The Irwin study itself has certain deficiencies based on the lack of data for comparison purposes with the ratings made by the TEPPS judges; (2) the tasks judged were Data Store tasks which are at a more molecular level than the PEF units; (3) the judges' ratings were not paired-comparisons but involved having the Irwin judges rate each task in one of 10 categories, ranging from most to least error; consequently we are talking about rating rather than paired-companions. In connection with this last point, however, Blanchard points out that rating and paired-comparison methods can both result in interval-property scales.

Despite the inadequacy of the supporting evidence, we are prepared to accept the exponential relationship (tentatively). If there were no supporting data at all, it would still be possible to accept the exponential as an hypothesis. However, there are more basic problems with the entire subjective data methodology.

What we have reference to is the relatively low between-judge agreement. Internal (within-judge) consistency was high (.909) but between-judge agreement was only .683 and decreased in subsequent studies (.416 in the user test, Ref. 7). There is a value of .687 reported for linear consistency (p. 27, Ref. 5, but see also Ref. 1) which is defined as the extent to which the paired-comparison model fits the observed data. It is not known, however, what the observed data were.

The low level of between-judge agreement would tend to negate the use of the subjective technique; if judges cannot agree on what they report, how much confidence can be placed in their judgments?

The developers of the technique answer the problem by saying that individuals vary in their judgments. They indicate that the probability of correct performance varies in a normal manner as a function of the number of individuals performing. Any probability estimate therefore represents mean performance. In any event, there is large variability in probability of correct performance, and the dispersion in the judges' comparisons reflect that variability.

If we interpret the argument correctly, (pp. 38-39, Ref. 5) the developers are saying that each judge's comparisons represents a subset of the population who actually performed and whose performance was observed by the judges. Judge 1 saw subset population 1 with its performance, judge 2 saw subset population 2 with its performance, etc. If then judge 1 says that that probability of task accomplishment with task 1 is higher than that of task 2, and judge 2 says that probability of task accomplishment with task 2 is higher than that of task 1, this merely describes what they actually saw. In that event the variability in judgments merely reflects the normal variability of observed performance in real life. No single judge would see all performance, so that inevitably it would be biased.

Under these premises a low between-judge reliability is no bar to using these judgments as long as internal consistency is high enough. Indeed, one would expect to find low between-judge consistency where real performance variability is high.

This is an ingenious argument, but not very satisfying, if only because it obscures rather than enlightens. All judgments by this reasoning, no matter how discrepant (from other judgments), are acceptable because they reflect normal variability and that portion of reality (and only that portion) that the judge observed. Thus, if in a given situation high interjudge reliability is secured; this is fine; if in the same situation low interjudge reliability is secured, this is also understandable. How then is one to know when a judgment should or should not be accepted?

Moreover, the whole premise underlying judgments is that a single judge's evaluation describes the total picture of the phenomena being judged, and therefore should correspond to the mean of the real variability (assuming that he has had the opportunity to observe a range of performances). If one rejects a judge because he is inconsistent among his own ratings, why should this be so? since his sample of behavior may have in actuality been distorted.

The developers indicate that the transformed scale values should not be used in the integrative model (p. 18, Ref. 7). They see the integrative model as dormant because it requires actual man-machine performance data.

If we have spent as much time as we have in examining the bases of the expert estimation method, it is because the method

requires further investigation. We say this not only because TEPPS uses subjective estimates, but because (to anticipate our ultimate conclusions), almost every other predictive model makes use of subjective estimates, usually in a far less rigorous manner than does TEPPS. Consequently, if this is to be a general procedure, the method must be explored systematically, its advantages and disadvantages noted, and it must be standardized. If modelers were to exclude subjective data from their techniques, one would not have to consider the problem further; it is apparent, however, that in order to exercise their models they will make use of such data, regardless of any inadequacies it may have.

Moreover, although one cannot accept Blanchard's statement "that the subjective scaling of man-machine activities (is) undoubtedly a valid procedure", on the other hand, one should not reject the method completely.[2]

So far we have said nothing about time data, which are also secured by subjective estimates; here, however, absolute judgments are required. The results of two studies produced a high degree of variability, so that the TEPPS developers recommend discontinuance of this method. However, see the study by Burger et al. (Ref. 3) and Smith et al. (Ref. 8) which suggest that the situation is more promising than had previously been considered.

---

(2) Blanchard remarks: "I feel that the viewpoint taken of the utility of subjective scaling methods is much to narrow. We must acknowledge that the dimension of probability of accomplishment or error is a complex one, and not held rigidly to traditional interpretations of values designed to assess uni-dimensionality. Unfortunately, my attempts to come to grips with a problem with rather unique theoretical as well as methodological implications had to be made secondarily, with the primary objective being to develop and apply a man-machine model. As a result, I am afraid my "shirt-sleeve" efforts were rather feeble compared to the scope of the problem. I sincerely feel that subjective techniques are applicable to the problem, and perhaps we have not yet evolved the proper conceptual framework or model for their use. We must surely tackle the problem though, since we will need some "interim" technique for obtaining useful input data for modeling purposes until empirical data become available."

The preceding discussion of data sources has been relative to the use of TEPPS as a system effectiveness prediction model. It is important to note that when TEPPS is used as an allocation (i.e., apportionment) methodology, it "does not require actual or objective measures of performance...; relative measures, far easier to obtain, will do just as well. That is because allocation of SER's is accomplished through mathematical consideration of the relative "weightings" of system PEF units" (p. 20, Ref. 5).

3. Output Metric

As in the case of the Data Store and THERP, the output measures employed by TEPPS are: (1) probability of successful task accomplishment; and (2) completion time. Of the two, probably somewhat greater emphasis has been placed on the accomplishment measure. Certainly workers in the field have found this measure of greater interest in every model they have considered. The reason for emphasizing probability of task accomplishment measures rather than time is that for many systems[3] time is not really relevant; as long as the task is accomplished satisfactorally, the length of time (within reasonable limits, of course) does not matter. This is true of troubleshooting operations, for example; research has indicated that technicians concentrate on making the proper diagnosis, rather than getting through in a specified time. It is indicative that in the user test performed with TEPPS, many of the time estimates given by operators were of "infinity", meaning that the time dimension was meaningless. Of course, an "infinite" completion time for a task would be intolerable; such a response means merely that, given that the task is completed successfully within a certain liberal time maximum, time variations prior to that maximum would have no significant effect on mission performance.

---

(3) On the other hand, a number of the simulation models, e.g., HOS, ORACLE, place major emphasis on performance time as a criterion of effectiveness. This difference in orientation toward system modeling is something that needs further exploration and resolution. Must we have different models for systems in which time and error are more or less important?

The developers of TEPPS repeatedly warn us that the probability of successful accomplishment measure produced by TEPPS will produce a highly conservative estimate of system effectiveness, because the facilitating effects of subsystem interaction, partial task success, task repetition and feedback looping cannot be handled by TEPPS very precisely.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

In order to develop the GSSM (Figure 5) which, like THERP's probability branching tree, is critical to use of the method, it is necessary to break the total system operation into more elemental units. This may require several developmental iterations before a satisfactory product is achieved, because great emphasis is placed on the accuracy of the GSSM. The analytic process required to develop the GSSM is conventional system/task analysis going from gross system functions to tasks. The task level for TEPPS is similar to that of THERP, that is, the task, but not the task element. Examples are: adjust cursor and read range counter; transmit bearing data; mark fade chart. It should be noted that there are no equipment details included in the task description, which, again like THERP, bears on the adequacy of the model for design purposes. However, Blanchard points out (personal communication) that "there is absolutely no reason why equipment details could not be included in the descriptions if such details were of interest or considered to be potential sources of performance variability."

The analytic method requires the description of all alternative modes of operation and contingency events, which leads to a fairly complex presentation. This again is comparable to what THERP does. Included in the modes of operation considered are redundancies, sequential and parallel activities, alternative pathways (depending on external situational events) and feedback loops. According to the developers it is possible to construct a GSSM for continuous and decision-making tasks, even though it is not possible to model them mathematically for the MSSM.

With regard to feedback loops, it is claimed that TEPPS has a built-in capability to treat some (although not all) of these loops (p. 20, Ref. 6), thereby reducing somewhat the need for the assumption of independence. However, the fact that TEPPS still employs the independence assumption means first that complete feedback is not modeled and secondly that the TEPPS output tends to underestimate the effectiveness of the system.

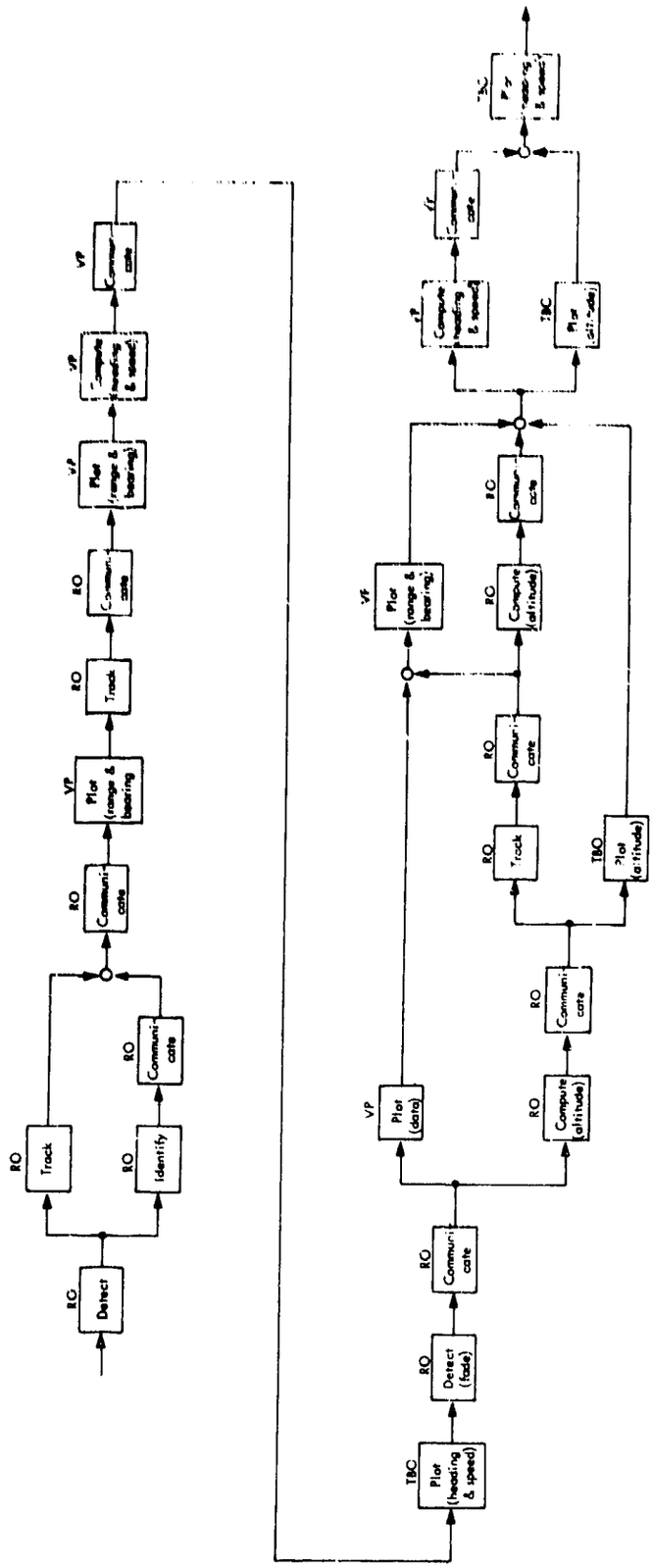


Figure 5. Functional GSSM of the CIC System. (Taken from Ref. 6)

The output of the task analysis is represented in the GSSM, which has in contrast to the absence of detailed procedures for the development of THERP's probability tree, fairly rigid construction rules. Ref. 6, p. 10-14 provides a complete guide to the construction of the GSSM.

Whereas the GSSM is essentially a flow diagram describing behavioral operations, the MSSM "is essentially identical (in principle) to a reliability equation. It can be used for determining system reliability or ... system probability of success ... (p. 6, Ref. 6). A computer program is available as part of the TEPPS package which eliminates the necessity for actually constructing a MSSM.

In summary, there appears to be no essential difference between the analytic method used in TEPPS and that used by THERP, although the former is more rigorous than the latter (or at least described in more detail). The GSSM and THERP's probability tree are also essentially alike, although it appears to this writer that it is easier to use the former. (This may be simply a bias on his part, based on his greater familiarity with functional flow diagrams, which is what the GSSM is essentially.) The PEF unit is essentially the same as the THERP task; both do not, as does the Data Store, deal on a detail level with the equipment description. Both TEPPS and THERP have great difficulty dealing with functional dependencies, although THERP says it can handle these. THERP deals explicitly with error, i.e., it achieves P_i by error measurement, but TEPPS simply utilizes probability of successful task accomplishment without evidence of much interest in error. Like THERP, TEPPS does not include in its GSSM system operations that do not have a significant impact on the system output; i.e., PEF alternatives that are possible but extremely unlikely.

2. Method of Synthesis

Synthesis of the system after it has been broken down into its component PEF units is accomplished through the MSSM. This is essentially "an equation which expresses the relation between the required probability of achieving the system output state and the probabilities (to be derived) of accomplishing system PEF units", (p. 51, Ref. 6).

In its general form the equation is

$$P = (P_1)(P_2)(P_3) \ldots (P_j)$$

where

P = system effectiveness requirement (SER)
P_j = required probability of accomplishing the jth PEF unit.

The above equation is therefore identical to a conventional reliability equation except that P would be defined as system reliability, rather than a pre-established required SER, and P_j would be defined as the reliability of the jth PEF Unit, rather than the required probability of accomplishment.

What this means simply is that although the above equation is for allocation of required probabilities (performance standards), it can be used for the determination of system effectiveness, once a weighting factor (IOTA) is applied to each P_j. Note that the equation is multiplicative (i. e., assumes independence). As a consequence the MSSM is extremely sensitive to the number of PEF units included, so that each additional PEF unit will result in a lower system reliability estimate.

A method is presented for handling redundancies. The general equation for redundancies is P = 1- (1-A)(1-B) ... (1-N), where A, B... N are alternative pathways. This gives us the probability of either/or redundant pathways.

All the examples given in Reference 6, however, are one way (⟶); they do not show feedback loops. Obviously the lack of such feedback loops will produce an overly conservative estimate of human performance, as was shown in the user test (Ref. 7). As far as TEPPS is concerned, (but this applies also to other predictive models), one problem requiring research is how to identify, model and include in the MSSM feedback effects.

3. Data Application

The basic probability data (derived from IOTA's) and time estimates relate to the individual PEF unit and are applied by the computer program to that unit. Again this procedure is identical with what is done on THERP. Once the probability/time values are assigned to the PEF unit, the MSSM proceeds to combine in accordance with the product rule and redundancy equations noted earlier.

ANTICIPATED MODEL USES

1. Prediction of System Effectiveness

Under this heading we can include two aspects:

a. estimation of system reliability
b. allocation of performance requirements.

As far as estimation of system reliability is concerned, TEPPS will presumably supply

(1) An absolute estimate of the system reliability to be anticipated when the system becomes operational (e.g., the system will eventually perform with a reliability of .9978).

(2) A comparison of estimated system reliability with that required to achieve mission success (e.g., estimated system reliability is. 9978, but mission success requirements (SER) call for .9987; or, the mission must be accomplished in 16 minutes, whereas required system performance time is 18 minutes).

Both of these assumed capabilities are identical with those assumed by the AIR Data Store and THERP.

It should be noted that both the above predictive functions are data-limited, that is, the adequacy with which they can be performed depends on having available to the model a store of predictive data and this adequacy is only as effective as the accuracy of those data. In addition, their adequacy depends on the ability of the technique to model feedback loops which compensate for the innate conservatism of using multiplicative procedures to synthesize system reliability.

As a consequence of their experiences (see Ref. 7) the developers recommend allowing the predictive function to be dormant until such time as an appropriate data bank and more effective modeling techniques become available. The author is not that pessimistic. In any event, considering the high degree of similarity between the TEPPS and THERP models, and the fact that THERP (if one believes the published reports) can provide valid estimates of system reliability, it is difficult to see why TEPPS cannot do the same.

Perhaps a systematic comparison of the two techniques (based on a comparison of data inputs and data processing operations) would be useful to determine what modifications to TEPPS might be made to improve its predictive capability.

TEPPS should be far more successful with regard to the allocation of performance requirements (i.e., given an SER of .9995, how this requirement should be apportioned among the various system tasks), since this function does not depend on an empirical data base. Note that TEPPS was initially developed for this purpose and that no other man-machine model (as far as the author knows presently) pretends to apportion SER's quantitatively. This function is also well worth exploiting for the very early system planning design stages. However, it calls for having an SER applied to the personnel operations of a system, which is something that is not too often included in system planning.

2. Design Analysis

Under this heading we include several aspects:

a. The comparison of alternative system configurations to determine which should be selected for implementation.

b. The determination of redesign requirements for a system which cannot satisfy system requirements.

c. Recommendations for initial design, design changes, selection of components, equipment layouts, etc.

With regard to the comparison of system configurations, it should be noted that, as long as the same data inputs are applied to both configurations, the absolute accuracy of those inputs is of relatively little importance. One can compare two or more of anything as long as the basis of comparison is the same. (Obviously the basis of comparison must also be relevant, but this is assumed.) From that standpoint TEPPS, along with either of the two models previously considered, can be used to make such a comparison. The same qualifying statements must be made for TEPPS as for THERP, in this regard; since both do not deal with equipment details, the comparison will be meaningful only as long as the behavioral units (i.e., tasks or PEF units) differ between the two configurations; where the same behavioral units are found in both configurations, it is likely that no real differences between the two configurations will be shown by the comparison.

It is important to consider, however, that system developers could have much more confidence in this type of comparison because it is not data-limited. The ability to model interactive and feedback behaviors would be more important. In any event, the system developer would be wise to consider the use of any technique to perform such a comparison in early system design.

With regard to more specific design/redesign recommendations, TEPPS, like THERP, may indicate where a design/redesign analysis is necessary (in other words, which task merits design attention). Presumably that PEF unit with the lowest predicted reliability becomes a candidate for redesign. Again, this procedure is akin to THERP's ranking of Q_i values. However, TEPPS, like THERP, will not suggest what the nature of that design change (e.g., a different component, a different layout) should be. This is because, like THERP, TEPPS functions largely with behavioral and not with equipment dimensions.

3. Manpower Selection and Training

TEPPS provides comparatively little information concerning manpower selection and training. A low predicted performance reliability for a PEF unit may suggest that personnel performing that unit must be better qualified or trained more intensively, but does not suggest what the basis of the selection and training should be.

Although the user test critique (Ref. 7) denigrated the value of TEPPS for training purposes because the high probability values achieved did not provide much discriminability among tasks, the author's objections are not based on this point. Even if probabilities for various PEF units differed markedly, this information would not suggest what is to be done in training. It is therefore only half an answer. That half (answer) would be useful, but it is our feeling that much the same information could be achieved by other means.

VALIDATION/APPLICATION STUDIES

TEPPS remains unvalidated in the sense of having its predictions compared with criterion performance data. The Navy user test (Ref. 7) cannot be considered as a validation because no real comparison with

criterion data was possible (although see below). Essentially the study involved a test of the usability of the allocation methodology. Moreover, as the model developers point out, the test was severely flawed because there were only 3 Navy analysts, two of whom did not participate fully in the test program. Nevertheless, the concept of a feasibility test to examine the usefulness of a predictive technique under relatively controlled conditions is valuable and should be extended to other methods.

Of particular interest to this discussion is the fact that system effectiveness was predicted for the test system (NQS-6) and this prediction could be compared with the "real" (although unknown) system effectiveness, since the NQS-6 has been used for a number of years with at least reasonable success. The predicted system reliability was .01, which might suggest serious inadequacies in the predictive use of TEPPS except for the fact that, as Blanchard correctly points out (personal communication), "the test was indeed seriously flawed".[4]

EVALUATIVE SUMMARY

Validity - Although a formal "test" of the validity of the integrative part of the model has been performed, the test was so flawed that one must conclude that no validity data exist, nor has the method been applied to system development problems.

Reliability - The user test (Ref. 7) suggests that the personnel "had no difficulty in following the procedures and in applying the TEPPS techniques..." (p. 42). This suggests that when users are properly trained, their development of the GSSM and MSSM should be reasonably consistent among them.

---

(4) Blanchard notes: "Considerable rescoping of the project during its lifetime resulted in it being only a "test" of mechanical feasibility or its "exercisability". The GSSM was oversimplified and actually partially incorrect. Also, there were serious reservations about the data obtained for transforming the subjective estimates to p values (see pages 18-24 of Reference 7). In brief, the result was that there was in fact no test at all of TEPPS ability to provide quantitative output data relative to the test system."

System Development Applicability

A. Comprehensiveness: The model is somewhat limited in its application to discrete behaviors.

B. Applicability: The method in its integrative application outputs a prediction of system effectiveness but can be used only with difficulty for design analysis, manpower selection and training.

C. Timing: The method appears applicable to early as well as later stages of design, provided that sufficient information is available concerning system operations.

Model Characteristics

A. Objectivity: The data base to be applied to the model requires many expert estimates; however, the procedures for securing these judgments is highly explicit.

B. Assumptions and parameters underlying the model are explicitly defined and described in detail.

REFERENCES

1. Blanchard, R. E., Mitchell, M. B. and Smith, R. L. Likelihood of Accomplishment Scale for a Sample of Man-Machine Activities. Dunlap and Associates, Inc., Santa Monica, California, June 1966.

2. Blanchard, R. E. and Smith, R. L. Man-Machine Modeling: Some Current Deficiencies and Future Needs. Paper presented to Navy Human Reliability Workshop, Washington, D. C. July 23-24, 1970.

3. Burger, W. J. et al. Validity of Expert Judgments of Performance Time. Human Factors, 1970, 12, (5), 503-510.

4. Irwin, I. A., et al. See Pontecorvo, A. G. Method of Predicting Human Reliability. Ann. Rel. and Maint., 1965, 4, 337-342.

5. Smith, R. L., Westland, R. A., Blanchard, R. E. Technique for Establishing Personnel Performance Standards (TEPPS), Technical Manual, Report PTB-70-5, Vol. I, Personnel Research Division, Bureau of Naval Personnel, Dec. 1969.

6. Smith, R. L., Westland, R. A., Blanchard, R. E. Technique for Establishing Personnel Performance Standards (TEPPS), Procedural Guide, Report PTB-70-5, Vol. II, Personnel Research Division, Bureau of Naval Personnel, Dec. 1969.

7. Smith, R. L., Westland, R. A. and Blanchard, R. E. Technique for Establishing Personnel Performance Standards (TEPPS), Results of Navy User Test, Report PTB-70-5, Vol. III, Personnel Research Division, Bureau of Naval Personnel, Dec. 1969.

8. Smith, R. L. et al. Subjective Judgment as a Means for Generating Corrective Maintenance Time Data. Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio (in press).

OTHER RELIABILITY-ORIENTED PREDICTIVE MODELS

INTRODUCTION

A major class of man-machine predictive models may be termed "reliability-oriented". These models apply estimates of performance based on error rate data to tasks and subtasks; and these estimates are then combined using probability theory to output a prediction of total system performance.

The most widely known examples of such reliability-oriented models are the AIR Data Store, THERP and TEPPS. However, from the beginning of interest in human reliability prediction, which we can arbitarily date from the publication of the Williams (1958) paper, a number of model variations have been developed which should be considered. Among these we include methods described by Pickrel and McDonald, Berry and Wulff, and Navy Electronics Laboratory Center (Mr. Richard Coburn). In addition we should consider reliability-oriented research which bears upon model development; in particular, the work of Drs. Askren and Regulinski.

Because the techniques we shall discuss in this section are either not complete models, or based upon others already described, the author does not propose to describe them in as much detail as previous techniques.

IV. THE PICKREL/MCDONALD TECHNIQUE

It will become obvious to the reader that this technique bears a strong resemblance to THERP, upon which it is in large part based. The purpose of the technique is to identify and eliminate sources of critical human-induced failures. It is unclear from the authors' presentation (Ref. 3) whether they use the term "failure" in the sense of an equipment malfunction resulting from human error, or failure in the more general sense of any degradation (i. e., failure to accomplish the mission, delayed mission accomplishment, or degraded mission accomplishment) resulting from human error. The first definition would deal only with a subset of the total number of effects possibly resulting from human error.

In any event, although the method will permit an estimate of system performance, its intent appears to be more design oriented than merely predictive. For example, the developers indicate that "the method is designed as an aid for allocation of...personnel to tasks... (it) provides a framework for ordering man's inputs to the system...and...an objective measure upon which to base decisions which must be made in terms of costs to reduce or eliminate causes of these errors..."(p. 647, Ref. 3).

The methodology is, in fact, only a part of a larger plan for applying human factors inputs to design. Thus, the assignment of an error probability to a task is not the end of the operation; "each task is analyzed for the purpose of identifying probable sources of human errors... Various alternatives... are considered..." (p. 659-660, Ref. 3). In terms of making the predictive process part of the overall design process, the method is again similar to THERP.

The assumptions inherent in the technique are more pragmatic than conceptual. The basic assumption is "that efforts to eliminate sources of human errors depend upon the expected frequency that a system failure will follow this error, and the probable consequence of the system failure condition" (p. 647, Ref. 3). A subsidiary assumption is that "most critical performance occurs during system operations or during the processing of items whose failures may result in loss of the system" (p. 647, Ref. 3). Obviously, it follows from these assumptions that it is necessary to determine the probability (frequency) of error and error consequences. However, other assumptions of a more behavioral nature are not described in the model, which suggests a certain lack of sophistication in the methodology.

The methodology involves the following major steps:

(1) Task identification and description;
(2) Estimation of crew activity time and workload;
(3) Estimation of probability of error and error effects;
(4) Elimination of error sources.

Tasks are defined as major functions to be performed by a crew which may consist of one or more persons. It is obvious from the manner in which crew activity time and workload are determined that the tasks one begins with must be broken down into task elements at a very molecular level. It is unclear, however, whether this is also required for estimation of error probabilities. In any event, the procedure for performing this breakdown involves the development of flow diagrams for tasks and subtasks. Task descriptions involve, in addition to specific statements of work, "statements of: what must be perceived by the operator, intellectual functions.... (e.g., recalling information or interpreting display information) and the related responses..." (p. 651, Ref. 3). Contingency events (e.g., emergencies) must also be analyzed.

It should be noted that there is nothing unusual about the analytic procedure described above. Presumably if the system developer has performed an appropriate task analysis, all the items noted above will be available to him.

The estimation of crew activity time and workload (a procedure called "discontinuous analysis" and based on Ref. 1) involves assumptions that we will see in Siegel's digital simulation model. This assumption is that where the time required to perform crew actions is more than the time prescribed or available for these actions, a condition of overload exists which increases the probability of error. "Overload... will indicate portions of the mission... which may require modification if... equipment or crew malfunctions and failures" (p. 655, Ref. 3) (are to be avoided). In contrast to the systematic manner in which this assumption is used to influence error probability in Siegel's digital simulation model, estimates of overload in Pickrel/McDonald are used in a purely qualitative manner to suggest where human engineering modifications are desirable. In other words, workload analysis here is entirely distinct from error probability estimation and does not enter into the prediction of system effectiveness. As a matter of fact, the original paper describing the methodology (Ref. 2) does not include the estimation of crew activity times. However, we will describe the process in

brief because it is considered by the developers in Ref. 3 as an integral part of their technique.

Time estimates are developed for each task element of each subtask. "Each element is essentially a single perceptual-motor operation" (p. 652, Ref. 3). The items of information to be gathered for this analysis are:

(1) Number of bits of information which must be processed by the crew member for each task element.

(2) Information process rate (bits/second), or the time rate at which it is assumed the crew members can mentally "process" the information in the task element.

(3) Information process time (decision or reaction time) required for each task element. This is derived from the following formula:

$$I = a + H/R, \quad \text{where}$$

I = information process time;

a = simple reaction time;

R = information processing rate (item 2 above);

H = number of hits to be processed (item 1 above).

The reader will note the resemblance to similar parameters in the DEI.

(4) Visual transition time (time required for eye movements and focussing) in each task element. Note the similarity to "perceptual shift" in the AIR Data Store.

(5) Reach time (to reach the control involved in the task element).

(6) Manipulation time (time required to manipulate the control in the task element).

Times for items 4, 5 and 6 are secured from Methods Time Measurement (MTM) data.

The information processing, visual transition, reach and manipulation times are then summed for each task element. The individual task element times are then summed to secure subtask time. "System wait time" (or system lag time) is added in to secure a total cumulative time.

The above required task times are then compared with available task times taken from mission profiles, time line analyses, the mission requirement, etc. An overload condition exists when the total time secured above is longer than available time. This situation then calls for a system modification, because manifestly the crew cannot perform its mission under these circumstances.

Note that the above procedure is very similar to what the AIR Data Store provides in its analysis of performance time, although the Data Store times are tied directly to equipment characteristics; this is not the case here.

"Once it is established that there is sufficient time for a given task to be accomplished, that task can be analyzed in terms of the probability of errors occurring and the potential severity of their effects... human error is defined as the failure to perform a task within a designated time..." (p. 656, Ref. 3).

We note that this methodology, like others we have reviewed, deals with error only on a binary basis; its occurrence or non-occurrence; and cannot take into account qualitative differences in performance.

The following items must be estimated:

1. The probability that a specified task will result in a human error of class i; this is E_i. These estimates are made on the basis of the AIR Data Store, as well as "expert" judgments by human factors personnel, information from system simulation studies, experiments and design mockups. It can be presumed that every source of potential data will be exploited, just as it is by the techniques previously reviewed (except, of course, TEPPS). As in the case of THERP, no systematic manner of securing required data is suggested.

It is somewhat unclear whether E_i refers to the probability of occurrence of a specified error or to the probability that any error (of any type) will occur. Note that P_i (in THERP) refers to the probability that a given error type will occur, based on a previous analysis to determine the most significant errors that might occur. Because the Pickrel/McDonald methodology builds on THERP, we will assume it follows the same pattern.

2. The probability that a degrading effect will occur, given the occurrence of an error of class i; this is F_i. This probability, which is identical with F_i in THERP, is a judgmental one, based on error and failure data secured from system tests. When an error is likely to have more than one effect,

$$F_i = 1 - (1-F_1)(1-F_2)\ldots(1-F_n)$$

3. The probability that a human error of class i will occur and will have a degrading effect is given by

$$Q_i = E_i \cdot F_i$$

Note the identity with a similar formulation in THERP (F_iP_i).

4. The probability that one or more human errors will degrade the system is given by

$$Q_T = 1 - (1-Q_i)(1-Q_j)\ldots(1-Q_k), \quad \text{where}$$

Q_i, Q_j, and Q_k are the separate probabilities of system degradation resulting from human errors of classes i, j and k respectively. Alternatively the formula

$$Q_T = 1 - \prod_{i=1}^{n} (1-Q_i)$$

which is identical with the formulation in THERP, can be used.

5. The probability that the system will not be degraded, or task success, is given by

$r_t = 1 - Q_T$ or 1 minus the probability of human error degradation. This too is identical with THERP.

Up to this point, the Pickrel/McDonald method has done nothing more for us than THERP could do, and with somewhat less sophistication, because they have not included any assumptions such as the probability of operator detection/correction of an error. The question arises, however, of, given that a failure effect will occur as a result of human error, how significant is that effect. The further development of the technique provides a quantitative (although subjectively determined) way of assessing the significance of the error effect. The purpose of this rating is to

permit errors and/or tasks to be rank-ordered in terms of their error effect criticality and thus weighting the cost of a remedial modification against expected returns.

The criticality rating for a human error of class i is given as

$$C_{E_i} = E_i \cdot F_i \cdot S_i$$

We are familiar with all except S_i. This is determined by judging the potential effect of an error on the basis of the following scale:

Safe---	0-0.1
Marginal---	0.1-0.3
Critical---	0.3-0.8
Catastrophic---	0.8-1.0

The criticality of a given task is given as

$$C_T = 1 - (1-C_{Ei})(1-C_{E_j}) \ldots . (1-C_{E_k})$$

and of a mission as

$$C_M = 1 - (1-C_{T_1})(1-C_{T_2}) \ldots . (1-CT_n) \text{ where}$$

C_{T_1} is the criticality of Task 1, C_{T_2} is the criticality of Task 2, etc.

The relative rank of the criticalities assigned to tasks provides a basis for determining the amount of effort that should go into providing a fix for the problem. The remainder of the process is traditional human engineering.

It is apparent that, except for the determination of criticality effects, the methodology is almost identical with that of THERP. The one significant feature is S_i, which does appear to represent an advance over F_i alone.

Like THERP, the methodology is very pragmatic. Any source of applicable data will be utilized; the accuracy of the probability statements is less important than the use made of them. "Thus, even when the objection of questionable accuracy is allowed, approximation still enables the tasks to be rank-ordered by the criticality ratings" (p. 661, Ref. 3). "While the quantitative techniques of this method may be criticized,... this alone does not rule out the usefulness of the method" (p. 661, Ref. 3).

EVALUATIVE SUMMARY

Validity - No validity data exist or are available; method has never been applied to system development problems.

Reliability - No information. Method has never been utilized by anyone other than its developers.

System Development Applicability

A. Comprehensiveness: No limitation. Method can be applied to all types of equipments/systems/tasks.

B. Applicability: Model outputs a prediction of system effectiveness. Although it is intended to be used for design purposes, this application depends on supplementary human engineering procedures.

C. Timing: Method can be applied to systems in early as well as later stages of design.

Model Characteristics

A. Objectivity: Many subjective judgments required, the basis for which is not specified.

B. Structure: Assumptions and parameters underlying the model are not specified.

REFERENCES

1. Gross, R. L. et al. The Application, Validation and Automation of a Method for Delineating and Quantifying Aerospace Flight Crew Performance. Paper presented at the Workshop on Quantification of Human Performance, University of New Mexico, Albuquerque, New Mexico, August 1964.

2. Parker, D. B., Pickrel, E. W. and McDonald, T. A. Elimination of Potential Sources of Critical Human-Induced Failures in Space Systems. Presented at Symposium on Quantification of Human Performance, University of New Mexico, Albuquerque, New Mexico, August 1964.

3. Pickrel, E. W. and McDonald, T. A. Quantification of Human Performance in Large Complex Systems. Human Factors, 1964, 6 (6), 647-662.

V. THE BERRY-WULFF METHOD

One of the early (1959) studies attempting to develop a method of predicting the human reliability of the man-machine system was that of Berry and Wulff (Ref. 1). Their goal was to develop two basic techniques: (1) one for describing component reliability which would be appropriate both for hardware and personnel components; and (2) a technique for combining the reliabilities.

To solve the first problem they defined overall system reliability as the proportion of time that system output is in tolerance. This permits one to describe the reliability of a component not in terms of the component itself but in terms of the output of the component. The advantage of this particular concept is that both hardware and human components can be treated identically; in other words, the output classification is essentially neutral.

A reliability of an output of .90 means that 9/10th of the time when a correct input is provided to the operator, his output will be in tolerance. It is important to note that this is not quite the same thing as a .90 probability of task accomplishment, the usual way in which human reliability is conceptualized. Nevertheless, it comes close enough to answering what we wish to know about operator performance to satisfy us.

Berry-Wulff propounds certain basic theorems of probability statistics which might have been novel to the behavioral scientist then but which we are now quite conditioned to accept. The reliabilities of two related units (e.g., subsystems) may affect each other in multiplicative or additive fashion. A simple serial arrangement of components involves a multiplicative relationship. An additive relationship is used when components are redundant to another set of components. Figures 6 and 7 taken from Reference 1) indicates how an estimate for a man-machine system is derived.

Of somewhat greater interest to us is that in order to measure the reliability of operator performance, what is needed is an estimate of the proportion of times that the human performance will be acceptable for the purposes of the system. This involves a procedure analogous to "sampling by variables" in quality control. It is necessary to know what the minimum acceptable output of a component or operator performance must be. A sample of performance is secured. It is assumed that the distribution of successive performances will be normal (after the task

has been learned, of course) or this distribution can be normalized by appropriate transformation. The task is to estimate the proportion of the total distribution of performances falling within the specified tolerance bands. It is assumed that most human performances will be one-tailed, since excessively good performance does not impair system operation.

Even when no performance failure is observed, the proportion of failures may be estimated by calculating the mean and variance of the observed performance, using the following equation:

$$T = \frac{M - H}{\sigma} \quad \text{where}$$

T = the proportion of performance falling outside the tolerance band;
H = the minimal acceptable quality;
M = the mean quality observed, and σ is the standard deviation of the observed quality.

For example, suppose the time dimension were the critical quality to be measured. The response must be completed in no longer a time t than 2 minutes (H). Mean completion time is 1.2 seconds (M) and the σ of the completion time distribution is .6. Then

$$T = \frac{1.2 - 2.0}{.6} \quad \text{or} \quad \frac{.8}{.6.}$$

The reliability of the human performance is then found by computing the area from $-\infty$ to T under the unit normal curve, using a table of normal curve functions. If R is estimated reliability, then

$$R = 1 - I_B, \quad \text{where}$$

$$B = \frac{1 - \frac{M-H}{\sigma}\sqrt{\frac{N}{N-1}}}{2}$$

and I_B is the incomplete beta function, with parameters A and B equal to $\frac{N-2}{2}$. N is presumably the number of performance instances (e.g., 100) observed.

A special procedure is required where "a single man will produce two performances, or even that a single performance results in two different dimensions of output of a single performance" (p. 115, Ref. 1).

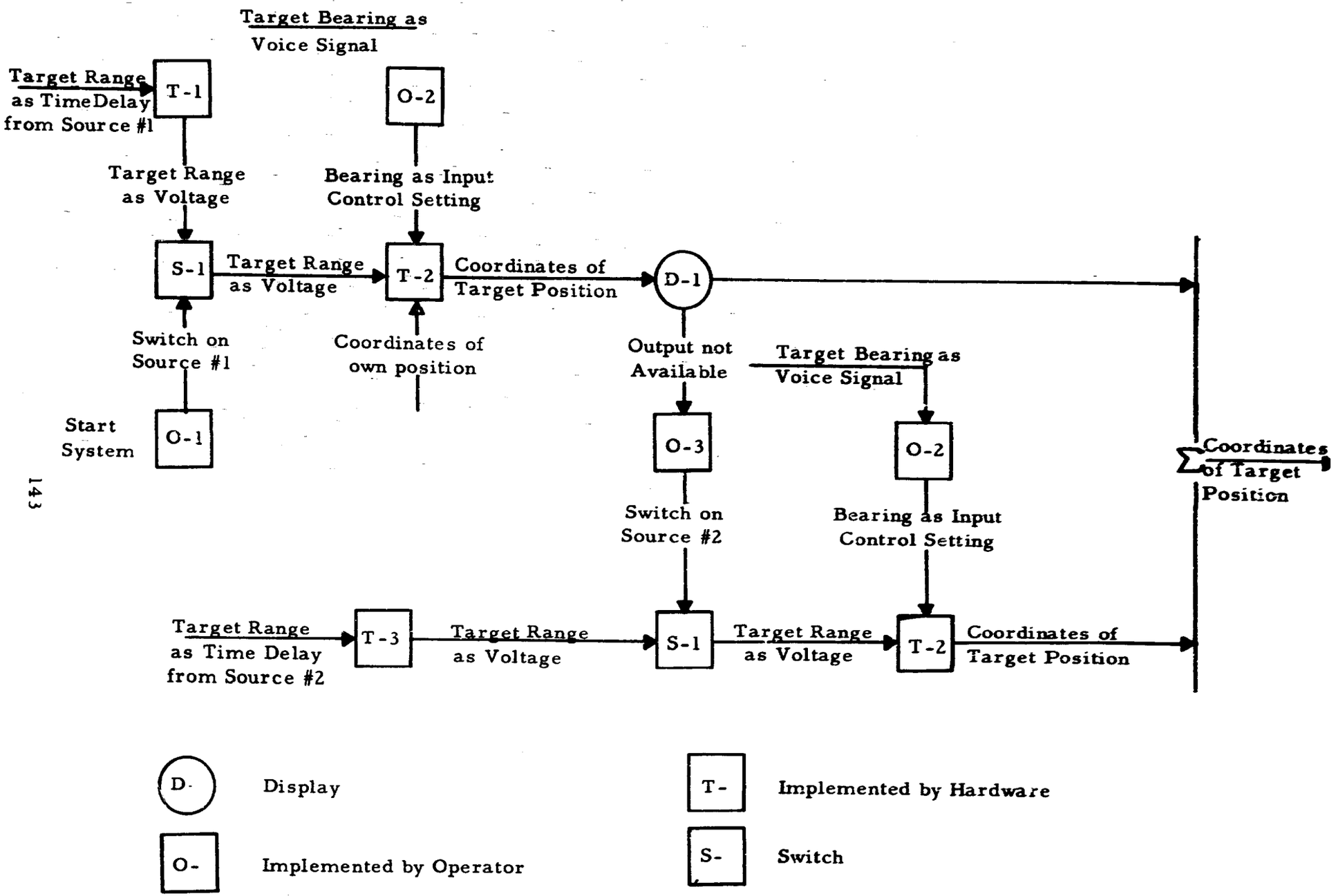
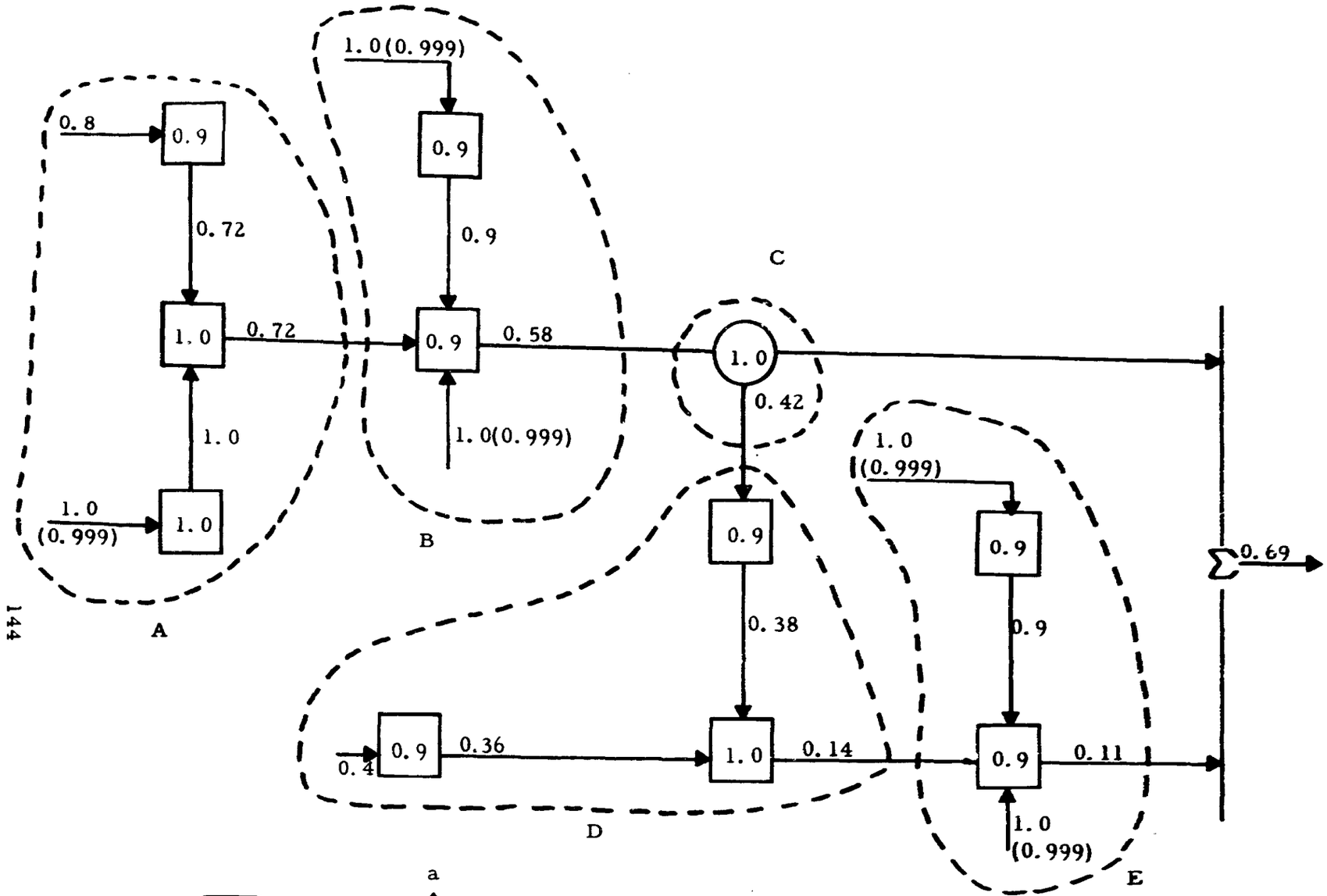


FIGURE 6. Example of an integrated man-machine system description in terms of performing units defined in terms of inputs and outputs.



$$r_{overall} = \overbrace{\underbrace{[(0.8)(0.9)(1.0)(1.0)]}_{A}\ \underbrace{(1.0)(0.9)(0.9)(1.0)}_{B}}^{a} + \underbrace{(1.00\text{-}a)}_{C}\ \underbrace{(0.4)(0.9)(1.0)(0.9)}_{D}\ \underbrace{[(1.0)(0.9)(1.0)(0.9)]}_{E} = 0.69$$

$$r_{hardware\ only} = [(0.8)(0.9)(1.0)]\,(1.0)(0.9) + (1.00\text{-}a)\,[(0.4)(0.9)(1.0\ [1.0)(0.9)]] = 0.76$$

FIGURE 7. Example of a procedure for calculating the over-all reliability of the system described in Figure 6.

Although this is a little obscure, the author interprets this as being analogous to a sonar operator using the same output to derive first the detection of a target and then its classification as submarine/non-submarine. Alternatively, the same performance could have two output dimensions, response time and error. "In such cases the two performance dimensions will clearly be correlated, and the usual multiplicative rule for combining probabilities will not be appropriate" (p. 115, Ref. 1). The procedure then is to correlate the two performances; the joint reliabilities of the two performances is given by the formula

$$R = \frac{W \pm \sqrt{W^2 - 4(Z-1)R_1R_2}}{2(Z-1)} \qquad \text{where}$$

$$Z = \left(\frac{\pi}{\cos^{-1}(\sqrt{1.2})} - \right)^2 \qquad \text{and}$$

$$W = 1 + (R_1 + R_2)(Z-1)$$

All of this looks extremely simple and it is, particularly that feature which does not require actual observation of failure. The limiting factor is that the output whose reliability is being estimated must have a tolerance band (H). There are many outputs which do not have such a precisely defined minimal acceptable quality. It is difficult to apply the procedure for discrete binary outputs, such as throwing a switch or making a binary classification (submarine/non-submarine). The procedure is, however, particularly applicable to continuous type outputs whose quality may vary on a continuum of more or less. For example, if one were estimating the reliability of tracking task, and H were an error of not more than X units (whatever these might be) and M were the mean tracking error, then the metric could be applied. Conceivably the measure could be applied to discrete tasks, if a continuous quality could be assigned to the output of the task. For example, if the task were one of detection, instead of conceptualizing detection reliability in terms of number of detections, one might think of that reliability in terms of a minimum time to detect.

One thing that becomes immediately apparent from reading the paper is that no special assumptions are made about the nature of operator behavior (except the notion of a normal distribution of responses (which itself is suspect)). Indeed, the application of industrial sampling procedures to the problem of operator reliability estimation suggests that

the developers think of instances of operator behavior like industrial component samples being inspected. In other words, the "black box" concept of operator behavior. This method is therefore at the other end of the continuum from the more behaviorally oriented Siegel, Blanchard and Swain models.

The reliability prediction made using this procedure describes the proportion of the time the system will be in tolerance. The prediction is not based on error or error rate, and in fact, as we saw previously, does not require any observation of error to secure the reliability estimate.[1] It assumes that any out of tolerance performance degrades the system; as a consequence, one does not have to determine the criticality of an error as was necessary in THERP or the Pickrel/McDonald methods. Obviously, any output of a component is essential to the system; errors can be considered more or less essential only if they merely lead up to that output.

Moreover, the method does not require building up an estimate based on molecular or intermediate control-display reliabilities (as we saw in the AIR Data Store), because it does not deal with the reliability of the component actions themselves, but only with the output of these actions. Nor is it concerned with the size of the unit for which reliability is predicted. An output is completely neutral with regard to the size of the component unit from which it was derived.

Theoretically a reliability estimate is an estimate of performance as a function of time, but because the procedure used here is one based on quality control sampling, it ignores time. It is related to the number of instances of occurrence of a given performance. Although the reliability estimate is translated into a quasi-time measure (proportion of time), the use of the sampling concept makes it difficult to attack the meaning of the measure because it does not take time into consideration.

The procedure for determining inputs and outputs is in all likelihood a rather simple task analysis organized to reflect not behavioral operations but inputs and outputs, the same sort of analysis needed to develop a flow diagram which in fact is extremely helpful (even if not absolutely required) in this methodology (see Figure 6). The flow diagram is broken out into those units which are implemented by the operator and those implemented by hardware. The size of the unit and the components involved are at the discretion of the user of the technique. Since one is concerned only with inputs and outputs, the size of the intervening components is unimportant. "Within a given system description it may be appropriate to show a simple

1. However, an out of tolerance output is automatically erroneous.

switch as one performing unit... and at the same time to show a large computer as another..." (p. 114-115, Ref. 1).

The procedure for deriving an estimate of the reliability of the total system involves use of conventional probability statistics. No suggestions are made by the authors concerning sources of data for T. Obviously one could measure performance on the component whose output reliability is to be estimated, but this would defeat the idea of predicting early in system development. The sources of data utilized by other predictive models (e.g., experimental literature, test results) could presumably be used for this model, but the data would have to be phrased in terms of minimal acceptable quality. It is unlikely that the experimental literature would be phrased in such a way. It would also be extremely difficult to take advantage of subjective estimates, moreover, because judges would have extreme difficulty in estimating variance.

The purpose of the methodology is to derive an estimate of system effectiveness. Because it is concerned only with component outputs, it provides no information about the design of the component itself. Presumably if the output of a particular unit were unacceptably low, one could go back to the component itself to try to determine why that output reliability were low, but nothing in the methodology would suggest the cause. Again, however, this is the case with all the other methods examined. Similarly, no information is provided which bears on manpower selection or training.

There are obviously advantages in avoiding behavioral assumptions. If the Berry-Wulff method could be applied to a variety of human performances, it might be unnecessary to make use of more complex predictive methods which rely on behavioral assumptions. Unfortunately, the concept of a minimal acceptable quality (derived from sampling inspection operations) is a severe restriction on the use of the method. This may be why the author has encountered no data describing the validity of the method or its application to specific operational problems. Again, one is left in the air with a predictive method that might be of some use but which has never been carried far enough to determine just how useful it could be.

EVALUATIVE SUMMARY

Validity - No data available.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Model can presumably be applied to all equipments, systems and tasks.

B. Applicability: Method outputs a prediction of system effectiveness, but has no applicability for design analysis, manpower selection or training.

C. Timing: Because of the nature of the data required to apply the method, it would seem applicable primarily to systems that are already operational. However, the absence of a more complete description of the method renders this judgment tentative.

Model Characteristics

A. Objectivity: Model is highly objective, requiring few if any judgments.

B. Structure: Conceptual structure is ill defined. Many questions exist.

REFERENCES

1. Berry, P. C. and Wulff, J. J. A Procedure for Predicting Reliability of Man-Machine Systems. IRE National Convention Record, 112-120, March 1960.

VI. THE THROUGHPUT RATIO

We class the throughput ratio developed by the Navy Electronics Laboratory Center (Ref. 1). as a reliability-oriented predictive technique because it makes use of probabilistic indices reflecting task performance success. When we describe the ratio in detail, we shall point out these indices.

The ratio is designed to estimate the operability of man-machine interfaces or stations (the most common example of which is, of course, the control panel). Operability is defined "as the extent to which the performance of a man-machine station satisfies the design expectation for that station" (Ref. 1). The design expectation is presumably the design requirement. Notethat the ratio defines design adequacy (or operability) in terms of the operator performance that can be secured with that design. For example, if an operator generates only one acceptable output item during an interval when 5 are needed, the operability of the man-machine station is 20%. This puts the ratio in the same class with the AIR Data Store (which can be used for the same purpose), but not in the same class with the DEI, whose index is not related to performance.

Emphasis should be given to the term "output". The throughput ratio emphasizes the responses of the operator. The term throughput reminds one somewhat of the Berry-Wulff method, which also evaluates component reliability in terms of outputs.

The throughput ratio is defined as

$$M_p = 100\left(\frac{n_h}{n_d} - C_e\right) \quad \text{where}$$

M_p is the predicted man-machine operability in percent.
n_h is the number of acceptable throughput items generated per unit time.
n_d is the number of throughput items which must be generated per unit time to meet design expectation.
C_e is the correction for error or out-of-tolerance output.

The 100 in the ratio is designed merely to secure a percentage figure. The ratio n_h/n_d is essentially the same as achieved operator or equipment reliability and is completely analogous to reliability estimates produced by THERP, TEPPS and Siegel's digital simulation model. Indeed, it is a percentage itself.

The definition of n_h and n_d implies that the throughput ratio includes the performance reliability of both the man and the machine, so that theoretically it would be unnecessary to derive a separate equipment reliability estimate. Although operator performance at a man-machine interface cannot be implemented without machine processing, we do not feel that the throughput ratio can substitute for a value specifically describing equipment reliability (e.g., wearout). Hence we interpret the operability value to reflect only operator functioning as implemented by the machine.

Several points should be noted about this measure:

(1) The index is in terms of the number of items or responses emitted by the operator (hence the term "throughput", which implies transmission). The time factor is included in the index but only as a constant, i.e., as so many items per unit time. Hence the ratio is applicable primarily to time-dependent tasks. The suggestion is that the measure is most suited to tasks requiring repetitive responses that must be completed in a specified time period. The question that arises is whether the measure would be equally applicable to tasks not requiring repetitive responses or for equipment which is to be operated only once in a given situation, e.g., a missile launch panel. Would the ratio be effective where the number of possible responses is not under operator control but is determined solely by the equipment or the situation; or where the quality of the response rather than its quantity is the critical dimension in performance? Examples of the types of tasks for which the ratio is considered satisfactory by its developers are:

(1) speed and accuracy of entering data by means of keysets;
(2) speed and accuracy of updating positions on a CRT by means of joystick;
(3) percentage intelligibility in transmission and receipt of voice messages of specified redundancy and transmission channel characteristics;
(4) accuracy in combining component threat probabilities;
(5) probability of correctly setting a series of switches;
(6) degradation of detection performance as a function of time on watch;
(7) memory span for aurally presented data items.

(2) Note that the ratio includes a correction factor (C_e), although n_h is already defined as acceptable items. We interpret this definition

as a mistake; obviously, if the items are acceptable (i.e., correct), there is no need for a correction factor involving error.[1] Hence we consider n_h as simply the number of responses output.

Along this line the developers see the ratio n_h/n_d as being unacceptable without some consideration of the consequences of delay and error. Two factors are included in C_e: the probability of detecting/correcting an erroneous output and the system consequence of the error, factors which will be familiar to the reader from his review of THERP and the Pickrel/McDonald methods. It is accepted that C_e must at the moment be "somewhat subjective" (Ref. 1), because a foolproof way of scaling these factors is not yet possible. However, something like the scale of system consequences used in Pickrel/McDonald might well be employed here; as far as the probability of detecting and correcting an error is concerned, the assumption used in THERP might also apply here (although the THERP assumption involved the concept of stress, which is absent here). There is some mention of using the AIR Data Store to help generate C_e, but no details are provided, and the nature of the Data Store seems unsuited to C_e.

Before proceeding to a detailed examination of the ratio, it may be helpful to review the purposes for which the ratio is to be used. These are:

(1) To compare the operability of alternative designs (note the similarity to DEI);

(2) To help establish system feasibility, which presumably involves comparison of predicted performance with that required of the design;

---

1. Coburn comments that "You suggest... that since the numerator of the throughput ratio is in terms of acceptable items, there is no need for a correction factor. You could make this interpretation; however, we were trying to use the correction factor in a different sense. That is we wanted some way to reckon the system impact of error. By this concept, acceptable throughput is a positive operability factor, whereas uncorrected detrimental error would be a negative factor."

(3) To help correct human engineering discrepancies, which suggests that the ratio can be used in redesign of designs already evaluated;

(4) To serve as a measurement tool or metric to demonstrate system acceptability.

The ability of the ratio to perform these functions will be discussed later.

The ratio applies to something called a man-machine station, which is not too well defined. The simplest way to think of that station is in terms of any set of controls or displays required to produce the output to be measured, e.g., a control console for a computer or a sonar console. In any event, since the ratio reflects an output, it cannot refer to any individual control or display, but rather to all the controls and displays required to produce a specified output. Where a given man-machine station includes several banks of controls/displays, each of which is used to produce different output, presumably a separate ratio value would be required for each output-distinctive set of controls/displays. A problem may arise when the same man-machine interface is used to produce two different outputs, e.g., a sonar console requiring both detection and tracking. Because of the different behavioral functions involved in two separate activities, two distinct ratios might be derived; would these then be averaged, or displayed jointly?

The correction factor C_e is a composite of error rate, probability of not detecting an error, and probability of function failure resulting from the error. C_e is defined as follows:

$$C_e = \frac{\frac{r_s}{r_r} \times n_h}{n_d} \times f_1 \times \frac{r_s n_h f_1 f_2}{r_r n_d} \quad \text{where}$$

r_s = number of trials in which the control-display operation is performed incorrectly;

r_r = total number of trials in which the control-display operation is performed;

f_1 = probability of the operator not detecting an error;

f_2 = probability of function failure arising from the error.

Since n_h/n_d represents the ratio of all outputs expected to be performed to all outputs required, without regard to their correctness, C_e must be subtracted from the n_h/n_d ratio.

f_1 is the probability of not detecting an error because one wishes to subtract this value from the overall ratio.

Note the difference between r_s/r_r and n_h/n_d. The former refers to erroneous trials; the latter to the number of items or responses output. A trial is presumably equivalent to the unit time in which the responses must be emitted, so that if unit time were an hour, 10 trials would last 10 hours.

The C_e factor interrelates error rate with detection probability with the probability of function failure. It is not clear whether there is any particular conceptual basis for the way in which C_e is defined, but it seems a bit over-elaborate.

An example will indicate how an operability value is derived.

Assume that 10 responses are required per unit time (n_d), but that only 5 are emitted (n_h).

r_s = 2 incorrect trials;
r_r = 10 trials performed;
f_1 = .20 (probability of not detecting an error);
f_2 = .70 (probability of function failure).
M_p would then be calculated as follows:

$$100\left[\frac{5}{10} - \left(\frac{\frac{2}{10} \times 5}{10} \times .20 \times \frac{2 \times 5 \times .2 \times .7}{10 \times 10}\right)\right]$$

$$100\left[\frac{5}{10} - \left(\frac{1}{10} \times .20 \times \frac{1.4}{100}\right)\right] \quad \text{or}$$

$$100\left[.5 - .0003\right] = .4997 \times 100 = 49.97\%$$

The question may be asked, how sensitive is the correction factor? In the example prepared above, its impact on the n_h/n_d ratio was minimal. Validation data (if such were available) might provide some indication of how discriminating C_e is.

An essential preliminary step in deriving the index is the performance of a task analysis, as indeed it is for all the other methods reviewed. This probably should include the development of function flow diagrams to trace an output back to the man-machine interface responsible for the output.

Since the throughput ratio deals, like the Berry-Wulff method, only with the outputs of the man-machine interface, the data needed would seem to refer only to the type of response output. For example, if the items output were messages, the data needed would be the number of messages that could on the average be output by operators. If the items output were switch settings, one would need data on the number of switches that could be set in a given time.

We note, however, that the above description of the required data items is insufficient. For example, the number of messages output depends on many variables. e.g., type of message (code or plain English), length of message, message composition device (typewriter, chord keyboard, lightpen, etc.). Hence it would appear as if the task analysis would have to specify those variables that had to be taken into consideration in selecting from the available data. (Of course, if one uses actual system test results, this problem does not exist, but there are many situations in which actual test data are not available.) What we are trying to say is that not only must the task analysis methodology specify the variables to be included in the applicable data, but the data source itself must be described by a number of pertinent variables. A data item which would merely say, for example, that on the average 27 messages can be transmitted per hour would be completely unusable because anyone who wished to make use of these data would not know whether the 27 messages were 5 or 10 or 15 characters long, in code or plain text, etc.

In other words, the task analysis- and this applies to any method which applies data for predictive purposes, other than test data generated on the very same system being evaluated, or expert judgments based on experience with that system- must indicate the variables affecting the output performance to be estimated. Unfortunately the procedure for the throughput ratio, as well as all the other procedures reviewed, never make this point clear; nor do they specify for the data they require the variables describing those data. Undoubtedly this matter is taken into account when the predictive procedure is actually utilized, but failure to indicate this consideration in the description of the technique leaves the technique rather up in the air.

The creation of any data bank requires the development of a taxonomy of variables describing the data included in the bank. In many cases the pertinent descriptive variables can be easily inferred by the data bank developer from the nature of the data; however, the data bank user does not have the same advantage and will require a descriptive classification

of the data. The process of selecting the appropriate data for predictive purposes is then one of matching the task-analytically derived variables against the taxonomic variables describing the data bank. The process reminds one of information-retrieval procedures and is illustrated in reference 2.

The index describes only the individual man-machine station. The question may be asked whether one can combine the percentages for two or more stations in series or parallel, either by averaging them or applying probability statistics. The answer is unfortunately no; one cannot use probability statistics because the output measure is not a probability but a percentage. However, if one were to ignore the "100" in the equation (which is purely for purposes of deriving a percentage), the output values of the equation could be interpreted as probabilities, because the elements of the equation (i.e., n_h/n_d, r_s/r_r, f_1, f_2) are essentially probabilistic. Used in this way it should be possible to combine individual station values to secure a measure of system output.

With regard to the comparison of alternative man-machine interface configurations, once an operability index has been calculated for each configuration, the comparison is simply a matter of checking one index value against another. The higher the index, the better the configuration.

Except for the throughput concept, the index is fairly traditional: n_h/n_d, r_s/r_r, f_1, f_2 have all been encountered previously. The equation takes the general form of the traditional reliability equation: 1 minus a failure rate, except that here 1 is replaced by n_h/n_d. Because of the probabilistic nature of the elements forming the equation, it should be possible to predict the effectiveness of the single man-machine interface by disregarding the percentage transformation. 80% operability for example, is really a task success probability, although the index measure is ostensibly oriented around number of items output. In other words, the measure is actually the probability of successfully outputting a required number of items.

So far we have talked about the assumptions underlying the throughput ratio. No assumptions of a behavioral nature appear to be required, nor can time/error relationships be inferred from the methodology. This may or may not be a good thing. f_1 and f_2 obviously require certain assumptions if valid measures of these are to be derived, but the description of the technique does not suggest any. Mention has already been made of the possibility of using Swain's assumptions regarding F_i to derive f_1 and Pickrel/McDonald's scale of system effects to derive f_2.

As with the other techniques reviewed, one wonders how sensitive the technique is for design purposes. A low operability index would suggest the need for some redesign, but the nature of the metric would not suggest what changes would be desirable. Indeed, the index is not at all related to any design parameters. This is in contradistinction to DEI, which is conceptually grounded in certain design relationships (formulated in information terms). It is almost as if, if one wished to use a predictive technique for design suggestions, it would have to be accompanied by a technique specifically descriptive of equipment characteristics, something like a checklist, perhaps, with the checklist items being calibrated in terms of the values of the predictive index. This is true of course of all the techniques reviewed, except possibly DEI.

The effect of any redesign to improve operability could be evaluated with the ratio by determining a ratio value for the redesigned configuration and comparing that with the ratio value for the original design.

Again, as with most of the other techniques reviewed, no validation or indeed application data describing the technique are available. As a consequence, the technique is only hypothetical, which is somewhat frustrating to anyone interested in evaluating its feasibility and utility.

EVALUATIVE SUMMARY

Validity - No formal data available; has never been applied to system development situations.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Reasonably applicable to all types of equipment, systems, task and behaviors.

B. Applicability: Outputs a prediction of effectiveness but cannot be used (except by inference) for other purposes.

C. Timing: Applicable to early as well as later stages of design, assuming availability of data.

Model Characteristics

A. Objectivity: Few judgments required.

B. Structure: Few assumptions required.

REFERENCES

1. Bunker Ramo, Final report on contract N00123-69-C-0132 for Navy Electronics Laboratory Center (undated).

2. Meister, D. and Mills, R. G. Development of a Human Performance Reliability Data System, Phase I. Final Report, Contract F33615-70-C-1518, July 1971.

VII. THE ASKREN/REGULINSKI MODEL

The man-machine predictive model most directly derived from reliability (and probability) theory is that of Askren and Regulinski (Refs. 1, 2, 5). As they point out, "Classical reliability analysis uses statistical inference to translate time-of-failure observations to a relevant model or models, and the prediction of reliability is obtained from the model via probability theory. This requires knowledge of some stochastic functions; e.g., probability density function (PDF) of the failures of the equipment with respect to time for the operations involved. Also, classical reliability modeling employs the first moment of the random variable which for the continuous case is time, denoted by "t", and known as mean-time-to-failure (MTTF), mean-time-to-first-failure (MTTFF) and mean-time-between-failures (MTBF)" (Ref. 1, p. 1).

The model derived is

$$R(t) = e^{-\int_0^t e(t)\,dt}$$

where R(t) is the reliability of human performance for any point in time of task operation and e(t) is the error rate for the specific task. Askren/Regulinski hold that the above equation "is completely general in that it holds whether the error rate e(t) is time variant or time invariant" (Ref. 1, p. 3).

The similarity of the equation above to the frequently used exponential equation for equipment reliability $(R = e^{-\lambda t})$ should be immediately obvious.

It may be asked whether the above formulation can actually be considered a model. Nevertheless, it is a model by the definition used in this report because it contains assumptions (e.g., that the tasks to which the formulation is applicable are performed in a time-space continuous domain), parameters (error rate), interrelationships among parameters (PDF), measures (MTTFF), etc., and a set of operations for deriving R(t).

The overall thrust of the work performed by Askren/Regulinski is that it is necessary to determine the distribution of error as a function of time (i.e., PDF) before one can adequately derive a prediction in probability terms. It has been pointed out that one of the objections raised to the reliability-oriented THERP and TEPPS models is that

these models utilize point-estimates based on 1 minus the error rate as input data to derive system effectiveness predictions. Such a formulation would be appropriate only if the shape of the distribution were exponential (see comments by Regulinski on page 159, Ref. 5). In any event, lacking knowledge of the nature of the error rate distribution, the result is an unknown amount of imprecision in the derived predictions. The use of a known error rate distribution to derive a prediction will result in a significantly more accurate prediction.

Two laboratory-type continuous tasks were developed to generate error data for testing the Askren/Regulinski model. The first involved a vigilance task, the second a manual control task (see Refs. 1 and 2 for a more complete description of the experimental tasks). A number of distribution functions (i.e., normal, exponential, Weibull, Gamma and log-normal) were tested against the empirical data (which function fitted the observed data best). Depending on the particular error measure employed (see following discussion), either the Weibull or log-normal distributions best fitted the data. Again it must be emphasized that these distributions were for continuous type laboratory tasks.

Note that the nature of the error rate distribution might well change depending on the experimental conditions modifying the task (e.g., if one increased accuracy or response time requirements).[1] Moreover, what the distribution would be for discrete tasks cannot be ascertained presently. Nevertheless, the basic principle remains valid: in order to make precise (relatively, that is) predictions, the nature of the distribution must be known. Any distribution is, however, valid only for a certain type of task, and, if there were N task conditions, it is conceivable that one would need a distribution function for each of them. At the very least, the parametric values in the distribution of function would vary.

---

(1) Askren points out "The thought is that the basic distribution of the data (log normal, Weibull, etc.) may remain the same for a particular task or a family of tasks, and the equation parameter values would change. Thus, we might have parameter value changes for various amounts of training, for various performance standards, various equipment features, etc., with the PDF remaining the same."

Hopefully, however, one would find that a given distribution function (e.g., Weibull) was applicable to a class of tasks, so that one would not be faced with the necessity of deriving a function for each task. Even here, however, the specific parametric values for the generalized function (e.g., the a and b values in the Weibull) would probably vary from task to task. [2]

Once one knows the distribution applicable to a particular task, or class of tasks, the probability values derived for these tasks using the function can be combined according to probability theory to secure a prediction of system performance involving these tasks. For example, if a system were composed of the two experimental tasks studied by Askren/Regulinski, the probabilities for each of these two tasks could be combined using conventional probability theory.

It is apparent therefore that there is nothing in the Askren/Regulinski formulation which conflicts with or supersedes the other reliability-oriented models (e.g., THERP, TEPPS) discussed in this report. These other models are not superseded because they contain variables not found in the Askren/Regulinski model, variables like the effect of stress, which describe the manner in which behavioral factors the basic error-rate function. For example, the distribution derived for the two experimental Askren/Regulinski tasks does not take into account the effects of stress conditions. If the system analyst considered that stress was an important factor in system operation, it would be necessary to include this additional factor in the prediction of system performance. [3] Of course, if one had an error-rate distribution for the two experimental tasks

(2) Askren notes that "Our thesis is that classes or families of human tasks exist for which underlying functions can be determined---variations in the conditions of the task (stress, training, equipment features) can be accounted for by changes in function parameter values."

(3) Regulinski points out that "Human stress manifests itself in human response time. The greater the stress the greater the hazard rate (error rate). Hence no additional factor in the prediction of system performance is necessary."

with stress included as part of the experimental conditions, then it would be unnecessary to apply the THERP or TEPPS formulations to secure a system performance prediction. However, the stress factor would have to be specifically denoted as one of the conditions under which the data distribution applies. Insofar as the combinatorial process in the other models discussed utilize conventional probability theory, then special operations would not be needed to supplement the Askren/Regulinski model, because the latter also applies conventional probability statistics to combine task predictions into a system prediction.

I consider therefore that the Askren/Regulinski model is essentially a way of securing more precise human performance probability estimates which could then be applied in the other models in exactly the same way as they do presently. (Askren notes also that it provides a quantitative language which allows direct incorporation in system engineering models.) One problem that arises is that these other models deal with discrete tasks, for which an error rate distribution is not presently available. Eventually, however, distributions for the discrete time domain should become available.

The Askren/Regulinski model does not take into account the effects of a given error on system function, as expressed in the THERPian parameter F_i or the Pickrel/McDonald f_1. (Some such concept is needed because errors differ in terms of their system consequences.) However, this matter can be taken care of by defining the measure for which the Askren/Regulinski probability prediction is derived as "errors leading to system malfunction" or "errors leading to catastrophic failure" or what ever is desired. But one will get a different distribution depending on how one defines one's measures.

In this connection Askren/Regulinski have conceptualized a series of potential measures closely related to traditional equipment reliability measures. To quote from reference 2 (pp. 3-4) "In reliability engineering, the term mean-time-to-failure (MTTF) is applied to components that are not repairable..., whereas mean-time-to-first-failure (MTTFF) and mean-time-between-failures (MTBF) are applied to equipment subject to repair. The three terms are useful in dealing with human performance reliability. MTTF translates into mean-time-to-human-initiated-failure (MTTHIF) and describes when a system function could be expected to fail as a result of an error or an accumulation of errors by one or more persons performing tasks in that function...

"MTTFF and MTBF translate into terms which describe errors whose effects are correctable. Thus, MTTFF transforms into mean-time-to-first-human error (MTTFHE). This is useful in treating errors that are highly critical, such that the first occurrence of an error would be costly.. The term MTBF converts to mean-time-between human errors (MTBHE). This is useful in treating errors of a less critical nature..."

Additional measures were also considered necessary to account for error correction. Building on the concept of mean-time-to-restore (MTTR), two additional terms were developed: (1) mean-time-to-first human error correction (MTTFHEC- an atrocious acronym, but necessary, one supposes) "which indicates the time on the average for man to correct his first error. However, man, during the course of a workperiod may commit a number of errors, yet recover from them. Thus a second term is necessary. This is mean-time-to-human errors correction (MTTHEC) and indicates the time, on the average, for man to correct all of his errors" (Ref. 2, pp. 4-5).

An objection was raised in the discussion reported in reference 5 that the Askren/Regulinski model is organized solely around time, particularly in terms of the measures cited above, when what the system analyst may want is a prediction of absolute performance, e.g., how well--- .87, .95-- will the system perform its mission? The various measures cited in the previous paragraphs are, however, only mirror images of an absolute prediction. If one knows, for example, that MTTFHE is 20 minutes, this is translatable into a reliability value of the .87, .95 type.

In the original version of this section the writer indicated that "The Askren/Regulinski model is not a universal panacea, however. It will not, for example, solve the problem of conditional probability. By this we mean the determination of the effect of one task or task condition on another concurrent or subsequent task. Probability theory can handle the problem of combining two or more task probabilities, which have already incorporated in each of them the effect of the other probabilities. The Askren/Regulinski model does not, however, suggest how to determine quantitatively these conditional effects."

In his comments Regulinski noted that the model will in fact handle conditional probabilities mathematically. The problem is actually an experimental one, not a mathematical one. If experiments are performed which describe quantitatively the relationships between two parameters

or two tasks performed over time, then the resultant data can be modeled to show their conditional probability. The difficulty is getting the data into the model (which is not necessarily the responsibility of the model developers, although it does bear on the feasibility with which one can utilize a model).

The problem above is one of securing appropriate data. This leads us to a consideration of the implications of Askren/Regulinski for data collection. Manifestly it is more difficult to develop a data distribution than a point estimate.[4] The difficulty is compounded if one seeks to secure such data from operational test sources. The reason is that the situation is usually insufficiently controlled in operational testing, and the opportunity to gather large amounts of data is quite restricted. Certainly this is the writer's own experience, reflected in reference 4. In consequence, if one must make use of laboratory studies to secure the requisite data, two difficulties arise:

(1) the process of data collection is slowed down;

(2) it is necessary to validate the laboratory results in comparison with results from operational testing to ensure that distribution A (secured from the laboratory) is much the same as distribution A' (secured from

---

(4) Regulinski notes that "Obtaining data is not an easy task whether modeling is to be done in time continuous or time discrete domain. I respect your opinion regarding development of data distribution vs point estimate, but reject your implied assertion that it is easier to obtain point estimates from operational test sources. The same data (obtained from operational test sources) which yield point estimates generally yield also data for the densities!!!!

"Case in point is the vigilance task reported in my paper,... The very same data can lead to point estimates of human reliability..... but this leads to the preposterous result of 0.99927536 (two errors in 2760 trials). In short, if the difficulty is as monumental as you suggest, does this justify such approximation as 0.99927536 vs 0.70 as reported? I assert that this is not an approximation gross or otherwise. It would appear to be more a case of looking in the middle of the night for a lost watch under a street lamp, and nowhere else, simply because under the street lamp one can see."

operational tasks). We know from much prior experience (see Ref. 9) that prediction form laboratory results to operational tasks is hazardous.

This does not mean that the situation is hopeless by any means, particularly if a generalized distribution function can be found to apply to classes of tasks.[5] Under these circumstances, even if one's operational data are limited, it may be possible to extrapolate the distribution function from limited data.

In any event, data collection to fit the rather rigid requirements of the Askren/Regulinski model will require the following steps: (1) First it is necessary to determine whether the Weibull/log-normal distributions will fit many more continuous tasks in the laboratory. (2) Then one must determine what experimental (e.g., task, environment, equipment) variables will do to change the distribution function and to what extent; (3) One must then validate the efficiency of a distribution based on laboratory tasks by comparing its predictions with empirically derived operational results; (4) Finally one must expand the methodology to account for discrete types of tasks and repeat the preceding process.

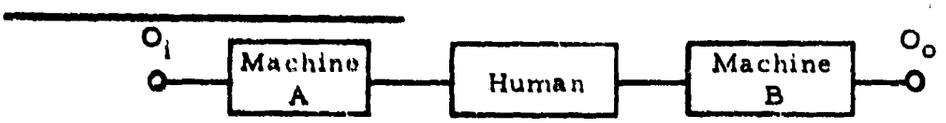
Just as we did for the other models, we must ask what the Askren/Regulinski model will do for us.[6] Manifestly it will provide an absolute prediction of task and performance reliability. Consequently, we can use its results to compare two or more alternative configurations by predicting the task probability of each configuration and taking the highest value. Presumably the Askren/Regulinski metric is sensitive to design in the sense that different equipment features will produce different task probabilities. How sensitive it is, we have no idea; laboratory studies should be done to determine this; it is, however, probably as sensitive as any equipment reliability predictive method is to design differences. This

---

(5) Regulinski comments: "I believe that you hit the nail squarely on the head. So far, our research leads us to believe that some classes of tasks are governed by the same density (pdf), and that stress changes only the parameters of the distribution. This much has been demonstrated and is reasonably clear. But the bounds of the parameters is one of many research aspect's to be investigated."

(6) Regulinski comments: "What will it do for us, . . . ? It is more likely to bridge a gap between the behavioral scientist (human factors. . . variety) and the systems engineer (reliability. . . variety) who has given the problem of modeling a man-machine system such as. . . . . "

may well not be sensitive enough to provide meaningful initial and redesign guidance to the system developer. It is unlikely that we will get more out of this model in this respect than we can from the others. The relevance of this metric to the solution of manpower selection problems, is dubious. The authors in reference 2 makes some claims for the methodology relative to determining how much training should be provided, but until relevant empirical data can be provided, it is difficult to take the claim seriously.[7]



will need

$$R_A(t) = e^{-\int_0^t h_A(t)\,dt}$$

and

$$R_B(t) = e^{-\int_0^t h_b(t)\,dt}$$

and

$$R_h(t) = e^{-\int_0^t e(t)\,dt}$$

in order to perform....

$$R_{system}(t) = R_A(t) \times R_h(t) \times R_b(t)$$

...and who is not likely to accept point estimate for $R_h(t)$ because by reason of simple logic and training he rejects the product of apples and pears to obtain bananas".

(7) Askren comments, however, that "The data in Tables 2 and 3 of reference 3 show that the mean-time values improve with training. For example, in Table 2, roll axis, the mean-time-to-first human error for 20 minutes of training (trial #1) is 23.4 seconds, and for 40 minutes of training (time accumulated for trials 1 and 2) is 214.9 seconds. Thus we have demonstrated that reliability of performance is related to amount of learning. This provides the empirical foundation for the assertion that a particular human performance goal (MTTFHE in this case) could be established, and the amount of learning time needed to provide the desired MTTFHE could be derived."

Obviously much more must be done before the methodology can be established without question. Several of the steps to be performed, including determination of generalized distribution functions, their application to different classes of tasks and in particular their validation, have been described previously. After all this is accomplished, it would be highly desirable to develop and test the adequacy of a prediction based on this model for an entire system, to determine whether the model can handle system prediction problems. Since the method is merely an extension of conventional reliability estimation procedures, there is no reason why it should not be used in the same way as the latter; but the point needs to be demonstrated empirically.

EVALUATIVE SUMMARY

Validity - Since the methodology derives from accepted reliability practice and probability statistics, there is no reason to question the validity of its application to system development problems.

Reliability - Since the methodology is a standard one, there is every reason to believe that different analysts will secure the same results when applying it.

System Development Applicability

A. Comprehensiveness: Although the methodology has as yet been applied only to a restricted set of tasks, there is nothing in the technique described which would prevent its being applied to the full range of tasks and behaviors.

B. Applicability: Like other models, the data derived from the method will predict task/behavior effectiveness, but does not in and of itself indicate design, manpower selection or training requirements.

C. Timing: Applicable to the entire system development life cycle.

Model Characteristics

A. Objectivity: Highly objective.

B. Structure: Assumptions and parameters well defined.

REFERENCES

1. Askren, W. B. and Regulinski, T. L. Mathematical Modeling of Human Performance Errors for Reliability Analysis of Systems. Report AMRL-TR-68-93, Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, January 1969.

2. Askren, W. B. and Regulinski, T. L. Quantifying Human Performance Reliability. Presented at the Second Annual Psychology in the Air Force Symposium, USAF Academy, Colorado, 20-22 April 1971.

3. Chapanis, A. The Relevance of Laboratory Studies to Practical Situations. Ergonomics, 1967, 10(5), 557-577.

4. Meister, D. et al. The Effect of Operator Performance Variables on Airborne Electronic Equipment Reliability. Report RADC-TR-70-140, Rome Air Development Center, Griffiss AFB, New York, July 1970.

5. Regulinski, T. L. Quantification of Human Performance Reliability: Research Method Rationale. In proceedings of the U. S. Navy Human Reliability Workshop, 22-23 July 1970, Report NAVSHIPS 0967-412-4010, Washington, D. C., February 1971.

VIII. THE DISPLAY EVALUATIVE INDEX (DEI)

INTRODUCTION

The DEI is a method for calculating a figure of merit of the effectiveness of equipment displays to transfer information to the operator and for the operator to perform control actions. Its primary purpose is to provide a quantitative method for comparing two or more alternative design variations of the same equipment without the necessity of constructing mockups and conducting operator performance studies.

GOALS

"The purpose of the DEI technique is to provide a quantitative method for comparing two or more design variations of the same equipment"... (p. 279, Ref. 2) (in terms, presumably, of their operability). "The technique also provides one basis for an impartial decision in the case of uncertainty about alternative equipment designs" (p. 279, Ref. 2). The technique is not intended for evaluating a single design. Conceivably one could use it to evaluate a single equipment, because the way in which the index is developed, its values range between 0 and 1, with 1 being ideal. Hence, the lower the index, the less desirable the design is. However, this is an ineffective way of utilizing DEI. On the other hand, the data, already developed by Applied Psychological Services, could represent a basis for some normative referencing.

In practice, because the technique concentrates on controls and displays, the technique is useful for evaluating control panels, display panels, and display/control panels, and not for evaluation of the entire system. On the other hand, since these panels are the focal point (usually) of all operator activities in the system, this limitation does not bulk too large. Although the technique emphasizes displays, it also includes control factors, and therefore need not be reserved for use with display panels only.

The technique also possesses some design diagnostic utility, although design diagnosis is not one of its enunciated goals. In the process of comparing alternative configurations, it is possible to determine which of the bases or links is responsible for the advantage of one configuration over another. Moreover, if one were interested in improving a particular configuration, he could modify the design in line with the principles used (e.g., fewer controls/displays, less mis-matching, etc.), measure the

results of the design change in terms of DEI and determine whether the change produces a higher DEI value (see examples of design variations in Ref. 1).

There has been some attempt to make use of the DEI as an instrument for determining training requirements (Ref. 3), but work on this is insufficiently advanced to assess the feasibility of doing so.. However, the preliminary indications suggest that the technique possesses some potential in this respect.

ASSUMPTIONS

The assumptions on which the DEI is based stem from information processing and communication theory, but are easily recognizable in common sense terms (not that the latter point should be considered any disadvantage).

"All else being equal, that system is best which:

"1. Requires the least operator information processing per subtask unit. For example, a predictive display is considered superior to a non-predictive display.

"2. Has the greatest directness between the information transmitters (displays) and the receivers (controls). For example, as the number of nodes in a communications network increases, the efficiency of the network decreases and its vulnerability increases.

"3. Has the least difference between the amount of information presented by an indicator and that required for a control action. For example, differential operator filtering and/or amplification may serve to increase error potential and decrease information transfer effectiveness.

"4. Provides for redundancy of information. For example, a message can be decoded quicker and with fewer errors when it contains redundant information.

"5. Requires the least intermediate data processing by the operator before he can perform the required control action. For example, unit conversions, transformations, multiple comparisons, integrations, differentiations and the like serve to delay information transfer and to introduce error potential.

"6. Has the least number of ... (indicators and control parts) ...

"7. Imposes the least amount of time stress on the operator as he performs the information processing...

"8. Has the least number of transfers which cannot be accomplished within a prescribed time. For example, if the operator must process more information than time permits, then certain transfers... will not be performed.

"9. Possesses the least number of critical transfers. For example, a system which contains a greater number of transfers which, if not accomplished correctly, requires task repetition or causes task failures is less efficient than the system with fewer critical transfers.

"10. Has displays and controls which are optimally encoded. For example, ... information transmission rate is lowered for highly symbolic stimuli. On the other hand, complete lack of encoding might create a large indicator or control matrix which would require hunting and searching...." (p. 279-380, Ref. 2).

Probability of correct response is tied directly to amount and manner of information transmission. Any design that requires the operator to work more, by being required to interpret more information or to make more decisions based on discontinuities between display and control information, is undesirable.

The necessity for processing information from displays and from display-control relationships may lead to operator stress which in turn could lead to increased probability of error. This is essentially the same orientation we found in the digital simulation model, except in that model the stress resulted from inability to perform within a prescribed time. In DEI the stress results from loading the operator's perceptual-motor channel capacity with more information than he can handle. The equipment is considered to provide information to the operator through its display and control design; the operator responds more or less effectively based on how efficiently he can organize and act on the information presented. Hence, for optimal response there must be a match between the operator's perceptual organization and the equipment's information organization. The conceptualization is quite elegant.

Each of the factors included in the index is essentially independent and hence can be combined multiplicatively. In view of the interactions generally found in factors influencing operator behavior, it is difficult to accept this assumption except as a pragmatic one, to simplify the calculation of the index. The effect of this multiplicative relationship is that any single factor which is deficient may exercise an excessive effect on the index value of the total configuration. Siegel notes, however, that the transfer chart which is an integral part of the DEI procedure, considers interactions between displays, between displays and the observer, and between the observer and his controls. In any event, since the technique merely ranks alternative designs, the absolute value of the DEI index is unimportant.

In the equation for the DEI the various factors are weighted by means of various exponents. The value of each exponent was not derived conceptually but as the result of an empirical fits to criterion data. As described in Reference 2... "appropriate transformations were then applied... so that a multiplicative combination of the base (factor) scores yielded close agreement between the merit ratings of the system design variations... and the DEI" (p. 30). There is no inherent objection to this procedure, but it must be recognized that the particular weights established have little conceptual basis.

Another assumption is that the display configurations being evaluated are properly human engineered (although see assumption 10 on the preceding page). In other words, the arrangement of stimuli on the display, for example, and the organization of the controls to be operated in relation to the display are not significant factors in the DEI. Consequently the index is primarily sensitive to the number of elements in the display configuration. This is not a criticism of the DEI as such, because the technique considers primarily the information transmission aspects of the design.

METHODOLOGICAL SCOPE

DEI is suitable for application to any control-display equipment, including those involving continuous tracking and decision-making

functions. The reason the index can include tracking and decision-making activities is that the index abstracts these activities and considers them only as nodes in the information transmission link. Consequently they are given standard link values (4). For example, "If intermediate data processing is required, this is represented by a box inserted in the link (creating a total of two links)" (p. 280, Ref. 2). The repetitive nature of tracking functions becomes unimportant, because no matter how many times a perceptual function is repeated as part of the same task, it receives only a single link value.

In consequence the technique considers behavioral functions only in a very abstract sense, which is entirely justifiable because DEI does not pretend to predict operator performance, simply differences in design quality.

The technique also considers differences between tasks. Tasks must be considered because different tasks may require operation of different controls and displays; where a task does not utilize certain displays/controls, these are ignored. As a consequence, a DEI value must be secured for a sample of tasks.

PARAMETERS

In its original development (see Ref. 1), DEI rested upon 5 bases which were later modified, expanded and retermed factors. These last are listed below and will be described in more detail when the individual measures are described.

1. Complexity factor: related to assumptions 1 and 4 (see section on Assumptions).
2. Directness factor: related to assumption 2.
3. Data transfer factor: related to assumptions 5 and 6.
4. Encoding factor: related to assumption 10.
5. Time factor: related to assumptions 7 and 8.
6. Match factor: related to assumption 3.
7. Critical link factor: related to assumption 9.

DATA

1. Measures Employed

a. The number of links between indicators and controls, the link representing the transfer of information between them. Each link has a weight according to the amount and complexity of this information transfer. These weights are assigned to classes of displays and controls. For example (see p. 18, Ref. 1) all cognitive links (regardless of type of cognitive function) have a weight of 4, all multi-state displays (4 or more states) have a value of 2, etc. The link weights were derived in accordance with the probability of successful performance of the link (see Ref. 1). This measure relates to the complexity factor which is represented by the formula

$$\frac{1}{1 + \Sigma w}, \text{ where}$$

w equals the sum of the link weights.

b. The total number of controls and displays $(n - m)_t$; the number of "used" displays and controls $(n - m)_u$; the total number of information links (N). These measures are used in the formula

$$\frac{(n + m)^2_u}{2N(n + m)_t}$$

which implies that the fewer unused controls/displays, the better. This measure is used in deriving the directness factor.

c. The number of gate, mixer or box symbols on the transfer chart (to be discussed later). A box represents intermediate data processing (e.g., computation); a gate (▷) indicates that information from two or more indicators is needed to set a control; a mixer (▶) represents a control activated on the basis of one of several displays. The data transfer principle uses the formula

$$\frac{2}{Q + n_o}, \text{ where}$$

Q = total number of displays/controls;

n_o = number of gate, mixer or box symbols.

d. The number of independent states presented or controlled by binary type units (e.g., on-off lights, pushbuttons). This measure, which is used for the encoding factor, applies only when the equipment contains at least one indicator or control with 12 or more independent states (not often found).

e. Time (Time factor), which is broken down into two types: (1) T, the total time required for subtask completion (what a given link actually takes in terms of time to be accomplished); and (2) T', the prescribed time for link completion. T is calculated from the formula, $T = 0.15 - 0.49I_d$, where I_d is the number of digits. The measure is used in the formula

$$\left[\frac{1}{16} I \left(\frac{T'}{T}\right)^3\right]$$

Note that the ratio between prescribed and required time is familar to us from the use made of this concept in the digital simulation model.

f. Amount of information mismatch (Match factor) between a control and the display(s) that provide the information to activate the control. This is determined by calculating the number of information based on the number of display states. This is contrasted with the amount of information based on the number of control positions. The amount of information from the control is subtracted from that of the displays (disregarding sign); this is done for all such display-control links and the resultant differences are summed (ΣM).

g. Number of critical links (N_c) (those that if not accomplished correctly, cannot be repeated and will cause task failure). The formula takes the form

$$\frac{\log N_c}{10}$$

and is related to the critical link factor. The fewer such critical links, the better.

Several thing should be noted about these measures: (1) They all stem directly from the assumptions underlying the conceptual structure of the methodology, although they can be summarized in terms of the common sense principle that that panel design is best which is simplest; (2) They are all more or less objective, requiring very little subjective

analysis on the part of the technique user; (4) Although the formula by means of which the individual measures are combined is fairly complex (see below), the derivation of the individual measures is quite simple, requiring only classification and counting.

2. Data Sources

The source of the data used to derive the above measures is to be found in the design of the individual equipment and the procedure for operating that equipment. External performance measures and data stores are not required.

3. Output Metric

The final formula for the DEI is:

$$\frac{(n+m)_u \sqrt[4]{(R)\exp\left\{-\frac{1}{4}\left[I\left(\frac{T}{T_1}-1\right)+\frac{1}{10}\,\xi I\left(\frac{T}{T}\right)^3+\frac{N_c}{10}+\xi/M/\right]\right\}}}{(1+\quad W\sqrt{4(n+m)_t\,(Q+n_o)}}$$

A somewhat simpler version of this formula was found in Reference 3:

$$\frac{(n-m)_u\left[\exp\left(-\frac{1}{4}\,\xi/M/\right)\right]}{(1+\xi W)\sqrt{N(n+m)_t\,(Q+N_o)}}\text{, where}$$

n = number of indicators
m = number of controls
N = number of forward links
$(n+m)_u$ = number of indicators and controls actually used in the console during a particular subtask
$(n+m)_t$ = total number of indicators and controls on the console
ξW = sum of weights applied to links
$\xi/M/$ = sum of absolute values of mismatches
Q = total number of display/control elements for used controls and displays
n_o = number of boxes and triangles representing intervening processes

In this simpler formula no use is made of the time, encoding and critical link factors.

Close examination of the elements of the above equations will show how the formula is composed of the individual measures.

The various factors are combined in these formulas according to weights that are derived "on the basis of the agreement of DEI applications to a number of systems with the opinions of the same systems of ... experts with whom the DEI was compared" (p. 284, Ref. 2). In essence an empirical fit procedure (see Ref. 1) was used to develop these weights and other expert ratings were used to cross validate these weights.

In other words, the conceptual structure underlying the technique was adequate to derive the individual measures, but not to integrate them. This is not an objection to the technique, merely an observation reflecting our lack of knowledge of how the behavioral elements of equipment operation interact.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Methods

A detailed task analysis is not required for this technique, since the transfer chart which is a major step in implementing the technique can be derived directly from the procedure for operating the equipment. The usual operating procedure is described in terms of direct control-display actions and hence is directly translatable into a transfer chart. Of course, if the DEI were to be applied to a stage of design before an operating procedure were available, it would be necessary to perform a detailed task analysis, down to the subtask (task element) level.

"To derive the DEI for any system, it is first necessary to select one or more representative tasks performed... A transfer chart for each task is then prepared. The transfer chart protrays the display and control elements and links them if they affect each other. Analysis of the transfer chart provides the basis for obtaining" (p. 8, Ref. 1) the various factors. "In preparing the transfer chart, the display and control elements involved are first listed. Then the symbols for the displays are drawn in a column near the left of the chart; the symbols for the controls are drawn in a column near the right... For a particular... chart, links are drawn between each indicator and the control(s) which it affects..." (p. 11 Ref. 1). Figure 8 presents a sample transfer chart (taken from

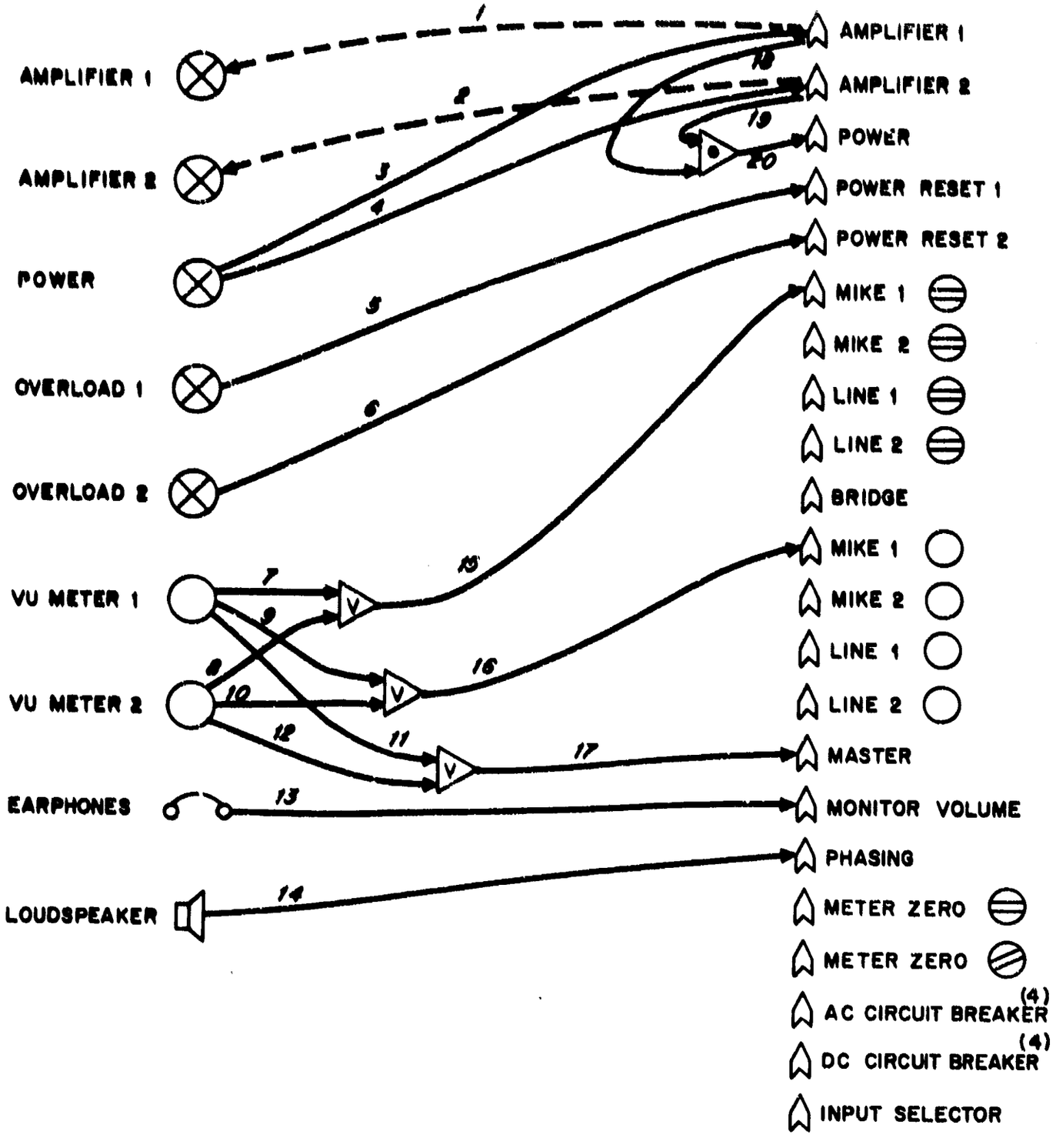


Figure 8. Transfer chart for variation O, Public Address Set

Ref. 1). The individual measures (described previously) are then derived directly from the transfer chart by counting and algebraic manipulation of the data.

The graphic presentation is therefore an essential part of the technique. This aspect is similar to that of TEPPS and THERP, except that the latter produce graphics at a task level rather than at the DEI subtask level; moreover, TEPPS/THERP graphics do not deal directly with equipment details.

2. Methods of Synthesis

These are inherent in the final DEI output metric and therefore need not concern us here.

3. Data Application

The measures for the individual factors are applied to the individual control-display link. The final output metric relates to the totality of the controls/displays used in a particular operating task. Note that the DEI measure relates solely to the individual task; where system operation involves several tasks, it is possible to derive a combined value for the various tasks by determining the relative importance of each task, multiplying the DEI value for each task by its weight and adding the resultant values. For example, assume that the DEI values for tasks 1, 2 and 3 are respectively .005, .004 and .006. If task 1 has a relative importance of .5, task 2, .2, and task 3, .3 (summing to 1.0, of course), then the summed weighted DEI value is represented by

$$(.005)(.5) + (.004)(.2) + (.006)(.3) = .0051.$$

ANTICIPATED MODEL USES

1. Prediction of System Effectiveness

Since the technique does not output an estimate of performance (either in probabilistic or other form), it cannot be used to predict the effectiveness of either an entire system or of a control panel. However, the technique was not designed for this purpose.

2. Design Analysis

In previous reviews we identified three aspects of this analysis:

a. The comparison of alternative configurations to select one to be implemented;

b. The determination of redesign requirements where the system cannot satisfy system requirements;

c. Recommendations for initial design.

DEI was developed specifically to permit comparison of alternative configurations and, judging from the validation data available, it does so quite well. One problem that may arise is knowing by how much one design is superior to another. Since the absolute difference between two designs is not related to an absolute difference in operator performance, it may be difficult to say whether a difference in DEI values is sufficiently great to warrant selection of the superior design. It stands to reason, however, that the larger the difference, the more significant the design differences would be, and the more compelling the rationale for selection of the design with the higher DEI. Other considerations, such as cost features (which the initial development of the technique attempted to incorporate, but later discarded) may negate a DEI recommendation, particularly if the DEI difference between two designs is not great.

The technique is not geared to indicate what changes should be made in a design to improve it, although one can use the principles on which the methodology is based to suggest redesign possibilities. If design A is inferior to design B, examination of design A in terms of DEI concepts may suggest that improvements could be made by eliminating intermediate data processing functions, reducing the number of unused displays, etc.

The technique does not pretend to provide any recommendations for initial design.

3. Selection

Not applicable to this technique.

4. Training

An attempt is being made to apply to DEI to the derivation of training requirements (see Ref. 3). While this effort is preliminary, initial indications are positive.

VALIDATION/APPLICATION STUDIES

As we indicated in previous reviews, one can consider validation in terms of either concepts (construct validity), cross-validation with external criteria like experts' judgments, or empirical operator performance. It has already been noted that the conceptual structure on which DEI is based is quite elegant. It is possible to consider DEI empirical validity in terms of

1. The ability of the technique to differentiate among alternative configurations;

2. Correlation of DEI values for alternative configurations with experts' rankings of the human engineering adequacy of these configurations;

3. Correlation of DEI values for alternative configurations with operator performance effectiveness on these configurations.

All three of the references noted at the conclusion of this review demonstrate DEI's ability to differentiate among alternative configurations and tasks. Reference 2 provides a table (p. 285) which indicates that the technique is highly sensitive to task differences (which presumably subsume equipment differences as well). Reference 3 is particularly interesting in this connection because the study was performed by researchers other than the developers (which would thus bolster the reliability of the technique). The DEI was applied to four equipments and four different subtasks and appeared to differentiate these quite satisfactorily. The only problem encountered was "determination of the number of states which controls, and in particular, displays could assume... To deal with this problem a number of conventions were adopted... In general, however, application of the DEI was straightforward. Values could be obtained fairly quickly, reliability did not appear to be a problem, and the index differentiated sub-tasks and devices The DEI possessed diagnostic value and was intuitively satisfying, varying in accordance with subjective impressions of sub-task difficulty" (p. 46-47, Ref. 3).

Similar results were achieved in comparing DEI rankings of configurations and experts' rankings of the same configurations. "The validational studies suggested that the technique empirically correlated in a strong positive manner, . . with the opinions of accepted human factors authorities." (p. 72, Ref. 1).

The author's own preference for validation would be a measurement of operator performance on design configurations (e.g., control panels) ranked in accordance with DEI evaluations. Presumably operators should perform more efficiently on configurations evaluated as being significantly superior in DEI terms. However, it is sometimes difficult to set up such a validation procedure and data on this kind of validation for DEI are not available.

Nonetheless, the validation studies that have been performed of DEI are quite impressive. Certainly the evaluative capability of the technique appears unquestionable.

Information concerning its actual application to the human engineering evaluation of control/display equipment is unfortunately not available. There would seem no reason, other than perhaps the somewhat laborious calculations involved, why this methodology could not be very successfully used in system development.

EVALUATION SUMMARY

Validity - Formal validation studies have been performed and show a reasonable degree of correspondence between DEI evaluations and other criteria of design adequacy.

Reliability - Controlled studies indicate that not only do various users get consistent results in applying the method, but that when the technique is applied to similar systems, consistent results are achieved.

System Development Applicability

A. Comprehensiveness: Method is limited to display configurations (including related controls, of course).

B. Applicability: Method does not predict or measure operator performance, but does evaluate display configurations.

C. Timing: Can be applied at all system development stages provided a display configuration design is available.

Model Characteristics

A. Objectivity: Highly objective; few or no judgments required.

B. Structure: Highly organized and well defined.

REFERENCES

1. Siegel, A. I., Miehle, W. and Federman, P. Information Transfer in Display-Control Systems. IV. Summary Review of the DEI Technique. Fourth Quarterly Progress Report, Contract DA36-039-SC-87230, U. S. Army Signal Research and Development Laboratory, Ft. Monmouth, N. J., 1962.

2. Siegel, A. I., Miehle, W. and Federman, P. The DEI Technique for Evaluating Equipment Systems from the Information Transfer Point of View. Human Factors, 279-286, June 1964.

3. Wheaton, G. R., Mirabella, A. and Farina, A. J. Trainee and Instructor Task Quantification: Development of Quantitative Indices and a Predictive Methodology. NAVTRADEVCEN 69-C-0278-1, Naval Training Device Center, Orlando, Florida, January 1971.

IX. THE PERSONNEL PERFORMANCE METRIC

INTRODUCTION

The methodology described in this section was developed by J. S. Brady as part of an attmpt to evaluate the contribution of personnel performance to the successful launch of the Atlas missile. As far as is known, the technique was never applied and consequently never validated. It has, however, some intrinsic interest because of the problems it attempted to solve.

GOALS

Before describing the methodology in more detail, it must be noted that it is purely an evaluational technique, not a predictive one. It is used to evaluate personnel performance effects in a system already built and in process of being exercised (tested or operationally utilized). Consequently it cannot be used to make any quantitative predictions of personnel performance. It is included among the methods reviewed in this report because evaluational methods are closely related to predictive ones, since, at the very least, they can supply data that might eventually be useful in prediction.

ASSUMPTIONS

The general criterion for system performance contains two parameters, quality and time. Since it is assumed that "the principal observable characteristic of the behavior leading to criterion performance is the exchange of information between man-and-machine... and between man-and-man... the "quality" of this intermediate behavior is determined by the efficiency of selection and utilization of the situational information available in the system. This, in turn, is defined by the procedure... Deviation from procedure, therefore, is considered as error" (p. 2,4, Ref. 1).

The assumption above, rigidly applied, means that all deviations from procedure are considered as errors, even when the deviation has no significant effect on system performance and even represents an improvement in the manner of operating the system. Theoretically, if the procedure has been correctly developed, no procedural step should

be unimportant or non-required, and to a large extent this is true of highly automated systems, since the need for performance flexibility (implying alternative procedures) does not exist.

Nevertheless, the assumption that every procedural deviation represents an error is bothersome, since we are all aware that every procedure is not as crucial as every other procedure. Hence one must account in some way for those procedural deviations that are minor. As we shall see, Brady attempts to do just this. [See also the distinction between essential and non-essential tasks in Siegel's digital simulation model, which deals with the same problem.

METHODOLOGICAL SCOPE

Because the methodology was developed in the context of a missile situation, its application appears to be restricted to those industrial and military systems in which execution of the system function is largely automated, few degrees of freedom are allowed in personnel performance (e.g., little decision-making because the situation is highly structured), and total or terminal system performance often cannot be observed or exercised. In such systems personnel functions are limited to "preparation, initiation, selection, surveillance, correction, preventive maintenance and repair. Hence evaluation of the contribution of personnel performance must be referred to some derived measure of system performance, e.g., amount of "out of commission" time, number of rejects, etc." (p. 1, Ref. 1). Personnel actions are rigidly specified and error-defined as deviation from procedure- is intrinsically undesirable. Since many system operations are automatic, "system performance must be inferred from fractional observation" (p. 1, Ref. 1).

The kinds of systems possessing the above characteristics are those of power station and telephone operation, automated and semi-automated factories and early warning radar systems.

PARAMETERS

The critical parameters in Brady's model are the errors that are made and the way they are classified. There are four classes of errors:

(1) Terminal error- deviations (from procedure) resulting in loss of missile quality.

(2) Risk errors- deviations resulting in decreased "confidence" in missile quality, e.g., omission or inadequate performance of tests.

(3) Delay errors- delay in the performance of requisite actions.

(4) Residual errors- all other procedural deviations.

The first thing to note about the above classification scheme is that errors are categorized in terms of their consequences to the system. The reader will recognize that other models have also included performance consequences in their methods, e.g., THERP's F_i, but have separated the occurrence of the errors, e.g., their probability of occurrence, from their effects. Ordinarily we distinguish between errors that may have significant and insignificant effects on performance. Brady's formulation deliberately amalgamates these two continua; this is possible in his methodology because he explicitly rejects probability of occurrence as an output metric, leaving only the effect-factor.

One practical consequence of the error classification is that in order to categorize errors in terms of their system effects one must have a detailed knowledge of system operations and functions; and even then the possibility of error in classifying actions in terms of system effects cannot be completely excluded.

Another thing to note is that the classification of these errors seems to fall into an inherent scale of their own, in which terminal errors have the most severe consequences and residual errors have the least severe consequences. Although the developer does not clearly specify the relationship, terminal errors are presumably akin to catastrophic errors, those leading to abort or actual destruction of the system. "Risk errors increase the probability of loss of quality. Delay errors may or may not affect overall performance time, depending upon the occurrence of masking by other activities" (p. 4, Ref. 1).

One of the difficulties that derive from the Brady methodology is that the definitions of the various error classes is not very clear cut in terms of effect. Residual errors are defined by exception, but this does not answer the question whether they have or do not have any effect on overall system performance. If one could speak for the developer, he would probably say that they do affect system performance, but so minimally that it may be impossible to discern that effect. Delay errors

presumably delay the mission, although the masking effect (something which also is not clearly defined) may obscure the delay. However, the definition of risk errors is quite tenuous, since any effect they have appears to be highly conditional on other mission activities.

What Brady appears to be suggesting is that every risk or every residual or every delay error has precisely the same effect as any other risk, residual or delay error. As we shall see later, this assumption poses difficulties for us when the individual classes of error are combined to evaluate personnel efficiency.

DATA

1. Input Data

Input data for the personnel performance metric are the performance events that the evaluator has observed or measured. Any externally imposed data bank is irrelevant, since the basis of the evaluation is the performance of the system itself.

2. Data Sources

The only data source required is measurement apparatus applied to the system.

3. Output Data

Brady considers the idea of determining the probabilities of the various kinds of errors but rejects the concept because of practical difficulties. This is especially so in the case of risk errors which are, as pointed out, highly conditional.

The kinds of problems one finds in the missile launch and missile-launch-related situations determine the kind of metric which was finally adopted. The problems can be summarized as follows:

"a. Only a portion of total system performance can be observed at one time.

b. Performance must satisfy several criteria- speed, quality and adherence to procedures.

c. Performance data consists of time and several categories of error, each of which relate differently to the criteria.

d. Raw time data contains periods which are not affected by personnel performance.

Therefore, any measure which is derived from such data:

a. Must be additive to permit accumulation of fractional observations.

b. Must bear a determinate relation to the performance criterion.

c. Must permit comparison between performances of different composition, preferably all performances should be scalable in a single dimension.

d. Must be selective, i.e., must score only relevant performance. Finally, the measure must be sensitive enough to discriminate between performances of varying degrees of quality.

In sum, the measurement requirements suggest a common additive metric which bears a determinate relationship to mission performance." (p. 10, Ref. 1)

The reader should note from the above that another criterion of system performance has been added- adherence to procedures. In general, adherence to procedures is not considered part of system performance criteria (at least in other models), but rather as a means of achieving those criteria. The inclusion of adherence as a criterion for successful system performance restricts the technique to those systems in which rigid procedural adherence is required. Systems in which procedures are flexible could be evaluated by this technique only with difficulty.

A number of possible ways exist to express the kind of metric Brady desires:

(1) Use of a metric expressing the probability of mission success, already rejected as being economically unfeasible.

(2) Abandon a common metric and express time and error separately. However it becomes impossible to compare two systems or missions which differ in terms of more than one measure.

(3) Define unsuccessful performance as occupying infinite time. That is, if a terminal error is made (i.e., one leading to missile destruct or abort), the time to successful completion of the operator's task is considered to be infinite. However, infinite time cannot be combined with more finite times for missiles that are successfully launched.

(4) Utilize a normative approach to time and errors, the resulting metric being the mean of the standard scores of time and errors. This is the approach taken.

(Note that the concern for terminal error presents a continuing problem for Brady. Such errors produce a distortion in the normal distribution of times to launch and therefore have to be treated separately, as we shall see.)

The model output supplies a centile score expressing the mean performance of personnel for the system under comparison. The centile value provides a score which is translatable into a relative measure of quality.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

In contrast to other methods, no formal or informal task analysis is required, since the formal operating procedure defines what an error is (but not the type of error). A task analysis may have been employed in the development of the operating procedure, but once that procedure has been formalized, no further task analysis is required.

Rather, a different type of analysis is needed, an error analysis, which focusses on the potential effect of the error. (We say "potential" because if the error effects actually occur, e.g., destruction of the missile, no error analysis is needed, since the effect is self-defining. This error self-definition applies of course only when the effect is sufficiently critical to be discernable or inferrable from physical changes in the system, which may sometimes be difficult.)

In any event, the error analysis involves considerable system knowledge and exploration and inevitably a considerable amount of judgment

is involved. Absolute certainty about the presumed consequences of anderror is not possible in certain cases.

2. Method of Synthesis

Since total performance has been decomposed into four classes of errors, and data have been secured for each type, to secure a system evaluation it is necessary to recombine them. This is done as follows:

"a. Given performance measurements:

t_i = performance time for Task i
r_i = the number of Risk Errors for Task i
d_i = the number of Delay Errors for Task i
m_i = the number of Residual (miscellaneous) Errors for Task i

b. In order to provide a single additive scale, normative data is accumulated and each performance quantified in relation to the performance population, i. e. , its standard score (z score) is computed, e. g. ,

$$Z_{t_i} = \frac{t_i - \overline{X}t_i}{\sigma t_i}, \text{ etc.}$$

where

$\overline{X}_{t_i}$ is the arithmetic mean of normative time performance for operation i,

t_i is the individual performance on operation i, and

σt_i is the standard deviation of normative performance on operation i.

c. Total error performance then, is expressed as the sum of error category performances:

$$\sum_{i=1}^{3} Z_{e_i} = Z_{r_i} + Z_{d_i} + Z_{m_i}$$ " (p. 12, Ref. 1).

The various categories of error are utilized instead of total error performance to enable the user to account for the differential relationships of each category to mission criteria. Obviously the different error categories vary in terms of importance (or effect on the system)

so that it is necessary to keep them separate and to weight them in accordance with that importance. Consequently Brady defines

$$\bar{Z}_{e_i} = \frac{aZ_{r_i} + bZ_{d_i} + cZ_{m_i}}{a + b + c}$$

where $\bar{Z}_{e_i}$ is the mean weighted error performance measure and a, b and c are weights.

One of the difficulties with this formulation, which the developer recognizes, is the derivation of the weights. "In the absence of a rigorour analytical basis for establishing weights, any differential weighting must be purely arbitrary" (p. 13, Ref. 1). It should be possible, however, to use a scale of the sort employed by Pickrel/McDonald to represent the presumed seriousness of the error class.

However, we have failed to deal with terminal errors. We can include terminal errors if we assume that a terminal error concludes the operation and hence can be replaced with the standard score value of the time spent up until the error was made. "The degree to which error commission affects the mission criteria is expressed in performance time" (pp. 13-14, Ref. 1).

Z_{t_i} (the standard score for performance time on task i) is combined with $\bar{Z}_{e_i}$ in accordance with the following formula:

$$Z_{E_i} = \frac{dZ_{t_i} + \bar{Z}_{ei}}{d + 1} \quad \text{where}$$

d represents an arbitrary weighting value.

The assumption that one can utilize the standard performance time score for an operation in place of number of terminal errors makes this author somewhat uneasy.

Summing over the various operations for which a personnel performance value is desired, we get

$$\bar{Z}_E = \frac{\sum_{i=1}^{n} Z_{E_i}}{n}$$

which expresses the mean performance of personnel for the system under consideration.

For convenience in interpretation, the distribution of Z_E is transformed to have a mean of 50, a standard deviation of 10, and to be high for good performance, i.e., E = 50-10.

Table 9 shows how E varies as a function of error and time performances. The values of E are based on the following normative data. Equal category weightings were used.

	Mean	S. D.
Risk errors	5	1
Delay errors	20	5
Residual errors	50	10
Time	30	5

Centile equivalents of E-scores are given from a normal distribution. The validity of this practice is, as the developer points out, open to some question because of the probable β-form of the time distribution.

The measure ultimately derived is an additive one which is appropriate because the sum of all errors can be considered equivalent to performance. However, the determination of the appropriate weighting factors for each type is quite obscure; in that sense it would appear that the model is not fully articulated. It is possible however, that a different set of weighting factors would be required for different systems.

ANTICIPATED MODEL USES

As pointed out previously, the method is purely evaluative; it does not predict. However, it is conceivable that the evaluation of one system could suggest the anticipated performance of a comparison system.

The method supplies no information relative to design, since it does not include design considerations in its data. The same applies to manpower selection and training.

However, the method does output an absolute measure of the adequacy of personnel performance, and presumably if a low centile score were achieved, this would stimulate some examination of the characteristics of the system. However, the model would not necessarily point to any specific aspects requiring corrective action. Note that the method supplies an output for personnel performance only; hence it is not a

TABLE 9

E-SCORE VARIATIONS AS A FUNCTION
OF ERROR AND TIME PERFORMANCES

Case	Condition	Time	Errors	Total Errors	E	Centile*
1	Average Performance	30	r = 5 d = 2 m = 50	75	50	50
2	20% fewer errors per category - Average time	30	r = 4 d = 16 m = 40	60	54.6	68
3	Singular errors avoided - total errors unchanged - Average time	30	r = 0 d = 20 m = 55	75	57.5	77
4	Average Errors - Improved time	20	r = 5 d = 20 m = 55	75	60	84
5	20% fewer errors - Improved time	20	r = 4 d = 16 m = 40	60	64.6	93

*Normal distribution

"true" system measure. Because E assumes a normalized distribution, it is questionable whether it could be combined with an equipment reliability value which assumes some other type of distribution.

VALIDATION/APPLICATION STUDIES

None.

In summary, the method is of interest primarily because it presents a somewhat different way of attacking system measurement problems. The use of a normative score based solely on error frequency and time would be highly attractive because it eliminates the necessity for data banks, etc., but only if certain of the problems discussed previously could be overcome.

There are a number of limitations on the technique:

"The necessity to infer total performance from fractional observations raises the question of sampling the population of crew tasks. A representative rather than a random sample is required to assume that specific tasks whose over-all importance to the mission and/or high frequency of occurrence are included in the sample. Further,

a. The tasks selected should adequately sample the repertoire of skills and knowledges of the crew members.

b. The samples should exercise each crew member on tasks of differing difficulty.

c. The tasks selected should require a degree of participation by each crew member that is representative of his participation in the total task population.

d. The work sample should be of such a character that the end product of performance can be evaluated." (p. 17, Ref. 1)

Another problem that must be considered is that when the sample is small or highly routinized, the opportunity to make errors will be quite limited. Consequently the distributions will be severely truncated. The same holds for performance times. Since the interpretation of E on population (centile) grounds is impossible without using the probability

density function of each distributions, the sample must be sufficiently comprehensive to make these distributions approximately normal.

Accumulation of work samples to achieve an overall system performance figure implies equal "importance" of the samples constituting the mean, i.e.,

$$E \propto \frac{\sum_{i=1}^{n} Z_{E_i}}{n}$$

If samples of disproportionate size are incorporated into this measure, the smaller samples exert unwarranted weight in the measure. A 5-minute work sample, for example, will be weighted equally with one of three hours or more duration.

A more rigorous exposition of the method described above is given in reference 2.

EVALUATIVE SUMMARY

Validity - No data available; method has never been applied.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Method is restricted to highly procedur-alized tasks.

B. Applicability: Method measures performance only; it does not predict that performance, nor is it related to design, selection or training requirements.

C. Timing: Useful only after system has become operational.

Model Characterisitcs

A. Objectivity: Ostensibly highly objective, but judgments of error effects are necessary.

B. Structure: Basis for making error effect judgments (critical to method) is unclear.

REFERENCES

1. Brady, J. S. and Daily, A. Evaluation of Personnel Performance in Complex Systems. Report GM 6300. 5-1431, ACPL-TM-60-1, Space Technology Laboratories, Inc., Los Angeles, California, 26 April 1961.

2. Brady, J. S. Application of a Personnel Performance Metric. In Majesty, M. S., Personnel Subsystem Reliability. H. Q. Ballistic Systems Division, USAF, 17 May 1962.

X. CRITICAL HUMAN PERFORMANCE AND EVALUATION PROGRAM (CHPAE)

INTRODUCTION

The model described below makes use of two rather novel concepts in human performance prediction (although individually they are familiar to human factors specialists): the use of checklists and ratings as a stage in the predictive process and the application of analysis of variance (ANOVA) techniques to develop the terminal reliability prediction. These concepts are the reasons why the technique is described in this report, since the methodology itself is ill defined and reflects significant deficiencies.

The methodology has three phases:

I. a. Analysis of the system;
   b. Establishing the rating "manual";
   c. Verifying the rating manual.

II. a. The actual rating (evaluation);
   b. Establishing performance criticality;
   c. Performing the overall systems criticality analysis.

III. Documentation, maintenance and follow-up phase.

GOALS

"The objectives of the applied CHPAE methodology are to:

a. Establish a criticality rank related to human performance.

b. Predict personnel effectivity or probability of human induced failure. . . .

c. Identify and eliminate sources of potential critical human induced failures.

d. Estimate required check redundancy for most probable success.

e. Evaluate designs from a human factors or man-compatability point of view.

f. Evaluate predesign concepts.

g. Provide inputs to training programs identifying and stressing areas of critical human performance.

h. Establish goals for optimum human reliability" (p. 117, Ref. 1).

We are familiar with most of those objectives from previously reviewed models. One interesting addition is the establishment of goals for "optimum" human reliability (whatever optimum in this context is supposed to mean). The author (M. A. Barone) tells us that "the technique is versatile and is applicable to a system, subsystem, event, activity or task. The evaluation or rating manuals can be tailor-made..." (p. 117, Ref. 1). "It provides a valuable tool for evaluating critical human performance, performing human error analysis, evaluating system and hardware designs, performing maintainability analysis, determining training requirements and evaluating human operational requirements..." (p. 122, Ref. 1).

The author has had occasion in the past to note the tendency of some model developers to characterize their techniques as being applicable to the entire range of human factors problems. There appears to be a pressing need on their part to establish a territorial claim, a sort of "territorial imperative". Such claims make it difficult for this reviewer (and for model users in general) to determine what the model is actually suitable for; it also generates a certain degree of skepticism (warranted or not) about these exaggerations.

ASSUMPTIONS

"The assumptions of the CHPAE methodology are:

a. Ideal potential human input interacting with ideal machine and/or environmental conditions facilitate the least probability of human error and the highest reliability.

b. Conversely, the worst possible human potential input with the worst possible machine and/or environmental conditions propagate the greatest probability of human error and the least reliability for success" (p. 117, Ref. 1).

No one could possibly disagree with the assumptions above, since they are underlying the basic concepts of Human Factors. Because of their generality, however, they throw very little light on some of the very complex processes involved in the methodology. For example, the rating process is the heart of the technique. The assumptions involved in the selection of the factors to be rated and the assumptions involved in the rating process itself are not enunciated. We have had occasion to note previously that some model developers ignore the need to clarify their assumptions, which makes it difficult to evaluate the adequacy of the model structure.

METHODOLOGICAL SCOPE

If the reader refers back to the goals of the model, it becomes apparent that the developer claims universality in terms of the systems, tasks and behaviors to which the methodology can be applied. One might be skeptical about this claim if it were not for the fact that almost everything can be rated. From that standpoint, if the methodology turns out to be valid, it would appear to be much more flexible than the other methods reviewed in this report.

PARAMETERS

"Development of the evaluation manual (the rating device) is preceded by project exploratory systems or task analysis. The purpose of the exploratory analysis is to select representative critical factors which best represent the man/machine and/or man/environment interfaces... The critical factors elected to be employed....are the factors which provide the optimum evaluation of the system... and have the highest correlation with prediction of potential human error." (p. 118, Ref. 1).

Typical candidate factors include motivation, personnel qualifications, auditory threshold criticality, visual demands, mechanical indicators, ease of maintenance, biological factors, physical demands, communications, test equipment and tools, work pressure, engineering change errors.

Note that such factors encompass the totality of the elements of a system and the influences operating upon these elements. They range from molecular equipment characteristics to rather gross personnel

qualities. Obviously each factor (particularly those of a molar nature) must be defined in terms of the elements making up the factor.

A more pressing problem is the basis upon which the analyst will select the factors to be considered. This basis can be entirely subjective, in terms of estimating the potential influence of the factor on error-production, or, as the developer indicates, by "correlation with prediction of potential human error". If the latter suggests some sort of statistical technique, the technique is not included in the model description. Alternatively, one could ignore the need to make a factor selection and attempt to rate all factors; this would, however, require very considerable effort on the part of the analyst.

If one is allowed to second guess the developer, the likelihood is that the selection will be made on very subjective bases. The inferred assumption is that the factors not selected will be those having a minimal impact on human performance; therefore, their elimination will not substantially affect the final reliability prediction.

DATA

1. Input Data

Despite the fact that ratings are involved, considerable input data are needed. The developer states "Following definition of the factors... data of known probabilities are slotted in their proper or estimated level in the defined rating scale. The slotted known or experimental probabilities in the rating scales are referred to as "inference marks"... The slotting of inference data of known or validated task probabilities is a continuous updating effort..." (pp. 118-119, Ref. 1).

Where these data are secured is not at all specified. Presumably it will come from the same sources used by other models reviewed. The input data consist of probabilities associated with fairly general tasks (not subtasks). For example, "read and identify simple electrical or mechanical instrumentation", simple leak checks, simple operational checks, install complex subassemblies, highly complex instrumentation readings, highly complex troubleshooting", etc. The probabilities associated with these tasks are apparently used to establish a weighting in terms of points (see Figure 9).

IDENTIFICATION		
The identification error factor appraises the probability that an error in identification will be committed when an object is identified incorrectly and then treated as if it were the correct object. Consideration shall be given to evidence that suggests the frequency of errors of identification is much higher than any probable human error.		
DEGREE	FACTOR LEVEL DESCRIPTION	PTS.
1	Little or no probability of an identicication error occurring Inference Marks ($\bar{R}$ = .975 ± 2.5%) o Read and identify simple electrical or mechanical instrumentation o Simple leak checks o General visual inspection o Simple operational checks o Identify items involving routine tasks o Remove and replace black boxes o Remove and replace standard piping (no complications)	7
2	20% confident that an identification error will occur at least once Inference Marks ($\bar{R}$ = .90 ± 5%) o Install complex subassemblies o Fault checks and isolation of detailed electronic instrumentaion o Read and follow complex instructions	18

FIGURE 9. Factor Definition (Rating Scale)
(Taken from Ref. 1)

DEGREE	FACTOR LEVEL DESCRIPTION	PTS.
3	25% confident that an identification error will occur at least once Inference Marks ($\bar{R}$ = .75 ± 5%) o Highly complex instrumentation readings involving highly complex systems o Highly complex troubleshooting	28
4	40% confident that an identification error will occur at least once. Inference Marks ($\bar{R}$ = .60 ± 5%) o Consideration for redundancy or redesign f of procedures required to maintain reliability	45

FIGURE 9 (continued). Factor Definition (Rating scale) (Taken from Ref. 1)

The input data selection and assignment procedure is essentially the same as that employed in other methods. However, the nature of some of the factors for which probabilities must be determined (e.g., human dynamics, occupational factors, work pressure, etc.) suggest that great difficulty may be encountered in securing appropriate data, since the experimental literature deals very inadequately, if at all, with such factors. The developer implies the use of expert estimates when empirical data are not available. Again this is a procedure employed by practically all models.

2. Data Sources

As indicated above, these are various, all sources being acceptable.

3. Output Data

Because the output measures derived from this methodology are somewhat unfamiliar, it would be best to combine their consideration with a description of the procedures for applying the model.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

As the developer indicates, the basic analytic method is task analysis. He provides no description, however, of what is involved in that task analysis, although it is obvious that certain unspecified processes are implied by the need to select the factors to be evaluated. In addition, some sort of analysis is required to differentiate the various factor degrees and the points allotted to each factor degree.

Once the critical factors are selected and defined verbally, a rating must be assigned to each factor. "The rating scale is by the degree of error confidence or probable error related to the degree of difficulty" (p. 118, Ref. 1). Presumably one of the elements in the task analysis is a determination of degree of task difficulty, but no definition of what difficulty means is supplied, nor is there any sample rating scale. As the developer suggested previously, the rating scales may be "tailor-made" to their application, which is fine but provides no guidance to the user who may wish to make use of this methodology.

"Values apportioned to the factors are based on the criteria of optimum prediction for criticality ranking. The preliminary values are slotted, known or estimated task probabilities.

"After validation of the preliminary manual, the final proportionate value weightings assigned to the factors are dependent on their correlation and contribution or ability to predict a valid critical rank. This can be accomplished by multiple or correlation coefficients in conjunction with trial re-runs of the sample validation procedure, until the rating manual is calibrated to provide a valid critical rank" (p. 119, Ref. 1).

The author must confess that much of this procedure is difficult for him to understand. He interprets the procedure as follows. Factors are selected and defined. A rating scale is developed for each factor

(see Figure 9) which is based on empirical or estimated task probabilities (of successful completion or error likelihood, but which it is impossible to determine). The intervals on the rating scale (degrees) are given a numerical weighting (points). The analyst examines each system factor and assignes a point value based on how the factor corresponds to each degree. In other words, harking back to Figure 9, if the task involves reading simple electrical instrumentation, he assigns a value of 7 points to the factor (identification) involved in that task. Any task may have several factors, each of which is assigned its appropriate value.

Presumably the validity of the rating scales is assured by running simulator or other tests and measuring operator performance on the task. That operator performance is transformed into actual task probabilities and the original assigned probabilities are compared with the actual ones, to be revised in accordance with actual performance. This is the only interpretation one can make of a validation process, unless the model developer implies that validation is a comparison with other experimental data in the literature. However, if actual performance validation of the rating scales is required, one questions the necessity for securing probability estimates from the literature. Moreover, the validation process would be a very strenuous effort.

Once the ratings are secured (from the validated scales) the rating values are included in a 2 way classification or split matrix as shown in Figure 10. This matrix displays tasks against factors. It is this matrix ("computerized as a 12 x 100 combined matrix with a subroutined split matrix 5 x 100 for the man affecting factors and 7 x 100 split matrix for the machine/environmental factors" (p. 119-120, Ref. 1) which is the source material for the variance analysis described below.

2. Method of Synthesis

The variance analysis model is represented as follows:

$$Y_{ij} = \bar{u} + a_i + \tau_j + e_{ij} \qquad \begin{matrix} i = 1, 2, 3, \ldots, r \\ j = 1, 2, 3, \ldots, c \end{matrix} \qquad (1)$$

where

Y_{ij} = The rating in the i^{th} task row and the j^{th} factor column.

Total Matrix

Activity	Man					Machine/Environment						
	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9	j_{10}	j_{11}	j_{12}
Task i_1	1	1	1	1	2	1	2	3	1	1	1	1
Task i_2	1	1	2	2	1	1	2	2	1	2	1	1
Task i_3	2	1	1	2	2	1	3	1	2	3	3	3
Task i_4	2	1	1	2	2	1	1	3	2	1	1	2
Task i_5	1	1	1	2	1	1	2	1	2	1	2	2
Task i_6	1	1	2	1	1	1	2	2	2	1	2	2
Task i_7	2	1	2	1	1	1	2	1	1	1	2	3
Task i_8	1	2	1	1	1	2	1	3	1	1	1	1
Task i_9	1	1	1	2	2	1	1	3	1	1	2	1
Task i_{10}												
i_n												

Split Matrix

FIGURE 10. Abstract Man/Machine Computer Matrix
(Taken from Ref. 1)

$\bar{u}$ = General mean.

a_i = Rating effect of the i^{th} row level of the factor treatments.

τ_j = Rating effect at the j^{th} column level of the factor treatments.

e_{ij} = Random rating error.

The following formula is applied to the matrix to analyze the variance of task probability ratings.

$$u = \frac{\sum_{ij} (\bar{Y}_i - \bar{Y}...)^2/(r-1)}{\sum_{ij} (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}...)^2/(r-1)(c-1)} \quad (2)$$

where:

$$S_1^2 = \sum_{ij} (\bar{Y}_i - \bar{Y}...)^2 = \text{Sum of the squares due to the factor treatments, } (r-1) \text{ degrees of freedom.} \quad (3)$$

$$S_3^2 = \sum_{ij} (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}...)^2 = \text{Sum of the squares for total variance.} \quad (4)$$

and:

$$u = \frac{S_1^2}{S_3^2} = \text{Variance distributed as F, if } H_o \text{ is true.} \quad (5)$$

The variance analysis related to factor treatments and the activity is expressed as:

$$u = \frac{\sum_{ij} (\bar{Y}_j - \bar{Y}...)^2/(c-1)}{\sum_{ij} (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}...)^2/(r-1)(c-1)} \quad (6)$$

where:

$$S_2^2 = \sum_{ij} (Y_j - Y \ldots)^2/(c-1) = \text{Sum of squares of task ratings (c-1) degrees of freedom.} \quad (7)$$

$$S_3^2 = \text{Sum of total variance as in (4).} \quad (8)$$

and:

$$u = \frac{S_2^2}{S_3^2} = \text{Distributed as F if } H_o \text{ is true.} \quad (9)$$

The computation of the CHPAE metric is accomplished by statistically comparing the rated distributions and statistical parameters of the total matrix and split matrices. Representative confidence curves are pre-positioned such that an ideal rating of the interface would reflect little or no overlap of the confidence curves and that when $X_2 - X_1 = 0$, the interface index would be highly critical.

By considering the variance of the scores, the inherent rated reliability of the interface P_{inh} is formulated below:

$$P_{inh} = \frac{(X_2 - X_1) - K}{\sqrt{\sigma_2^2 + \sigma_1^2}} \quad \frac{d_t}{\sigma_t} \quad (10)$$

where:

$(u_1 - u_2) = k$ or 0

X_2 = Average Environmental Rating

X_1 = Average Man Rating

σ_2 = Rating Variance - Environment

σ_1 = Rating Variance - Man

d_t = Activity Margin

σ_t = Total Variance

u_1 = Arbitrary Ideal Population - Man

u_2 = Arbi ary Ideal Populatio - Environment

The percent reliability, "R", is determined by referring calculated P_{inh} to t-Tables.

$$R = 100 - Q\,(t\text{-Table}) \qquad (11)$$

For assurance that (1) the probability of the specific points in the distribution curve actually overlap each other, a test of assurance can be applied to a probability of success, "P", and confidence level, "C", to yield a statistic of assurance, "T_A".

Then the test of assurance (T_A) is as follows:

$$T_A = \frac{P_{inh} - K}{1/N + \frac{K^2}{2N-1}}$$

where:

P_{inh} = from (1)

K = from Chi Square Table

The CHPAE criticality rank is obtained by ranking the significance level of the rated inherent probability P_{inh} - probability of successful human performance.

The criticality scale is as follows:

Criticality I - Above 0.1% level
Criticality II - 1% to 0.1% level
Criticality III - 5% to 1% level
Criticality IV - Below 5% level
Criticality V - Below 10% level

3. Output Data

The outputs of the methodology are, then, a reliability (R = 100-Q), a confidence level (T_A), and a set of criticality ratings based on the significance level of P_{inh}. Presumably those subsystems with high criticality levels are candidates for redesign.

ANTICIPATED MODEL USES

The potential model uses we are concerned with are:

(1) Prediction of system effectiveness;
(2) Comparison of alternative configurations;
(3) Suggestions for redesign;
(4) Suggestions for manpower selection and training.

With regard to the prediction of system effectiveness, it is somewhat unclear whether the model outputs such an estimate. The term "percent reliability" (R) may or may not be equivalent to the reliability metric employed by the other models reviewed, which can be defined in terms of probability of task completion or error likelihood. Assuming, however, that the metric is logically related to human performance, the model will predict that performance and thus can be considered a predictive tool.

As far as design analysis is concerned, the methodology appears to lend itself to a comparison of alternative configurations. R can be determined for these configurations and can then be compared.

We see the methodology as having even more potential for design diagnosis. Since the predicted effectiveness value is based directly on equipment/system characteristics, a low R or a high subsystem criticality rating can be related directly to the design feature which produced those values.

To the extent that one can validly rate manpower and traing factors, a low R or high subsystem criticality rating which resulted from these factors can be related directly to the responsible factors. One does not know, of course, what dimensions the manpower/training factors would possess (presumably this would be left to the discretion of the model user who would be required to make up his own rating manuals).

Theoretically an analysis of the rating scales for these factor-dimensions should suggest what aptitudes personnel should be selected for, what and how much training they should be given, etc.

This is the particular utility of the rating scale methodology: that if it works (in the sense of providing valid ratings that can be transformed into meaningful evaluations) it will supply a substantial amount of diagnostic information which other models might have difficulty supplying. Since the rating is directly applied to the factor of interest, it will be highly sensitive to the dimensions on which the factor is defined.

The "catch" of course is that rating scales and ratings are not known for providing very valid data because of the subjectivity inherent in the technique.

VALIDATION/APPLICATION STUDIES

None reported or known. In consequence the methodology can be considered only theoretical.

EVALUATIVE SUMMARY

Validity - No data available. Based on the model description (which is admittedly quite vague), there is reason to question the validity that could be achieved. The high degree of subjectivity inherent in ratings and the complexity of the factors to be rated are not conducive to confidence in the technique. Nonetheless, the concept of making use of factor ratings is an attractive one. The notion of evaluating the adequacy of the man-machine interface in terms of a significant discrepancy between the separate (split) ratings of the man and of the machine is also an interesting one. The assumption here is, one suspects, that when a significant difference in adequacy between these two subsystems is found, human performance can be anticipated to be poor. Data relative to the validity of this assumption is lacking, but it seems to be a reasonable one.

Reliability - No data available. The degree of reliability to be anticipated in factor ratings cannot be estimated because of idiosyncracies inherent in any rating.

System Development Applicability

A. Comprehensiveness: Very great.

B. Applicability: Can be employed for multiple purposes.

C. Timing: Can be applied to systems at all stages of development.

Model Characteristics

A. Objectivity: Highly subjective.

B. Structure: As presently described, very poorly defined.

REFERENCES

1. Barone, M. A. A Methodology to Analyze and Evaluate Critical Human Performance. 5th Reliability and Maintainability Conference, 18-20 July 1966, New York, N. Y., 116-122.

B. OPERABILITY PREDICTION MODELS

SIMULATION METHODS

I. DIGITAL SIMULATION TECHNIQUE (THE SIEGEL 1-2 MAN MODEL)

INTRODUCTION

In this section, we will consider only the models developed by Dr. Arthur Siegel and his co-workers. There are three such models: (1) the 1-2 man model; (2) the 4-20 man model; (3) the 20-99 man model. Of these three models, we shall concentrate only on the first. There are three reasons for doing so: (a) the basic model concepts and methodology are exemplified in the 1-2 man model, and therefore will serve as an introduction to the larger models, although Siegel indicates (personal communication) that "the latter consider different variables, possess different internal constructs, and provide different outputs"; (b) the two larger models encompass many psychosocial variables which may be of somewhat less interest to system developers; (c) a more practical reason: the two larger models are so elaborate that it would require a document the size of a book to discuss them fully. (A book, Ref. 6, has already been written.) We shall refer from time to time to these larger models, but only to compare them with the original.

We ought perhaps to begin by noting that until this point in the discussion we have been dealing not with full-fledged models, but rather with techniques which model man-machine processes only partially. The author defines a model in the Chapanis (Ref. 2) sense of a physical or symbolic representation of how the man-machine process functions and humans perform in the machine environment.

The probability branching tree in THERP and the GSSM in TEPPS are such representations, but these are only parts of techniques whose operations as a whole do not describe processes. The Siegel models are true models, since they simulate how the system being predicted functions; the essence of the models' predictive capability is the adequacy with which this simulation/representation is conducted.

Another thing that should be said about the digital simulation techniques is that they are dynamic, whereas the techniques we have considered previously are static. For the preceding techniques only a single probability /time value (or a mean and standard deviation) can be applied to the behavioral unit being predicted. In the digital simulation models, the computer samples from a distribution of values to derive the value applied at any one time to the behavioral unit. Moreover, it is the

interaction of model parameters during the course of the simulation which determines that sampling process. This interaction feature is far less evident in the non-simulation models.

Before proceeding to a detailed description of the Siegel model, it should be pointed out that the author does not intend to go into the details of the computer program used for this model. We accept the programming techniques used as given and merely look at them (when we do) in terms of their implications for inputs, outputs, data requirements, etc. We will examine the basic assumptions underlying relationships among parameters, but once these are accepted, it seems reasonable to assume that the manner in which they are included in the computer can be accepted as being correct.

The best capsule description of the 1-2 man model is given by Siegel himself in Reference 5 (page 557): "The model is used with a high-speed, general purpose digital computer. The system designer... makes an analysis of the man-machine system and the task under consideration. The performance of each operator is arranged into ordered, discrete actions called "subtasks," and for each of these certain source data are compiled. These data, together with selected parameter values (e.g., the time allotted for task performance), are put on punched cards and introduced into the digital computer... The computer sequentially simulated, according to the rules of the model, the "performance" of each subtask by each operator... A simulation is completed when the operators either use all allotted time or successfully complete the task. ... results are recorded indicating the areas of operator overload, failure, idle time, peak stress, etc., for the given set of selected parameters. Repetitions of the simulation, with different parameter values, yield a range of records.... If the results indicate modifications to the design of the system (are desirable), new designs may be similarly tested to determine the extent of improvements brought about by the modifications."

GOALS

The purpose of the 1-2 man model (described in Refs. 4 and 5) is to serve as a tool for system designers during the design stage and to indicate where the system (once conceptualized or formulated) may over or underload one or two operators. Because the model requires fairly molecular inputs (relating to something called a "subtask" which, from Table 1 in Reference 5, is at the task element level), it is less likely that it

would be used during the preliminary development of a system. We say "less likely"; this is not a bar to the use of the model in predesign planning, provided the designer can conceptually elaborate his plans to represent a full fledged system. [1]

On the other hand, assuming that the system concept has been elaborated, as would be required to exercise the model, the latter can be used to indicate (1) whether or not a given system design will perform to a desired level of effectiveness or to what extent; (2) where it does not, to indicate those stages of the system mission where the operator is not working efficiently (i.e., is over or under loaded); (3) to compare alternative system designs to determine which is best. The specific questions which the 1-2 man model should answer are:

(1) Can an average operator be expected to complete all required tasks successfully within time T (required) for a given procedure and a given design?

(2) How does success probability change with faster/slower operators and longer/shorter time periods?

(3) Where is the operator over or under loaded? (Note: The operator is overloaded if he cannot finish in time T; he is underloaded when he finishes much more quickly than the time requirement actually requires. Siegel notes that the model has recently been modified to allow stress to be a function of intermediate goals as well as the total time requirement.)

(4) What is the frequency distribution of failures as a function of stress and operator speeds?

Certain characteristics of the model are apparent from the questions it is designed to answer: (1) the model is highly responsive to time; indeed, as will be seen later, stress or work load is a function of the time available during the mission to complete remaining tasks; (2) the model is responsive to different types of personnel, which is not the case for other techniques. As a consequence, the model is highly dynamic because it permits one to examine system performance as a function of variations in stress and operator speed.

---

(1) Indeed, because of the "power" inherent in simulation techniques, we can see the Siegel model being used in early design to perform critical tradeoffs among parameters.

The same goals apply to the two larger models. The basic differences between them is the introduction of additional parameters. Thus, questions (1) and (3) above remain the same, but questions (2) and (4) are modified to include factors in addition to operator stress and speed, these factors representing psychologically oriented or psychosocial variables such as morale, goal aspiration, etc.

The primary use of the models as described in Ref. 6 is to test "the man-machine interactions in a proposed design." ""Testing" here probably means evaluation and in the case of a negative evaluation generation of ideas for a design fix and retest" (p. 141, Ref. 6). The implication here is that the model is not used to generate a new system design but to evaluate one that has already been designed. However, when one sees problem areas in his system design as displayed by the model, some insights may follow as to methods to compensate for these inadequacies.

There is a statement (Ref. 5, p. 557) that the purpose of the technique is to predict system effectiveness in early design and to enable comparative evaluation of alternative system designs. What "early design" means is, of course, subject to interpretation, but obviously design must have proceeded to a fair degree of system elaboration before the technique can be used. Because our own studies (Ref. 3) indicate that once a design has been elaborated, little or no attention is given by contractors to alternatives, the above statement may overemphasize this aspect of the model's use. If, however, we are talking about major system projects like the F-15 aircraft, for example, and particularly about specially funded efforts by governmental laboratories prior to system procurement, then the statement may well be correct.[2]

---

(2) Siegel contends "that the model is useful in the conceptual, definition, production, and test phases. It can be shown that such modeling can yield data useful in each of these phases. In the conceptual phase, the input data may be more gross, but certainly no more gross than the other engineering data employed at this point. The model doesn't care whether it works on molar or molecular subtasks. If one works with more molar subtasks during the conceptual phase, his output will probably be sufficiently precise for the requirements at this stage of an equipment development. In the production phase, the model may be employed to test the effects, for example, of design modification introduced by value engineers. In the test phase, it can be employed, for example, to perform tests which couldn't otherwise be performed. Our work on the effects of gamma-neutron radiation on F-106 performance (including both pilot and equipment) is an example here."

ASSUMPTIONS

The basic assumption in the Siegel models is that operator loading is the basic element in effective man-machine system performance (p. 10, Ref. 6). Although there may be a variety of reasons why the operator is loaded or unloaded, these reasons are compressed into a variable called "stress." The model makes the stress variable the key to operator performance in terms of both speed and quality of performance. This increase in stress may be caused by several factors: (1) falling behind in time on an assigned task sequence; (2) a realization that the operator's partner is not performing adequately; (3) inability to complete successfully a subtask on the first attempt and the need to repeat the subtask; (4) the need to wait for equipment reactions.

If one examines these factors in toto, it appears that stress is the consequence of failure or the expectation of failure. It is possible to think of other factors creating stress, particularly emergency conditions threatening the personal integrity of the operator (danger to life). (Emergency stress is included in the larger model.) An emergency condition would, however, cause the operator to abandon his programmed mission sequence and go into an emergency routine; the latter can then be handled as a separate sequence.

Stress resulting from a recognition of the urgency of the task being performed (e.g., a radar operator's determination that a signal represents an actual ballistic missile attack or a false alarm) does not appear to be handled by the model, because task urgency in Siegel's models results only from a lack of time to complete essential subtasks. However, this is a minor quibble.

It may appear as if the model, depending as heavily as it does on the stress parameter, would be inadequate in handling situations in which there are no time pressures and hence no stress. (However, there are probably few such tasks.) Moreover, the model can account for such situations by utilizing the average probability of completing the subtask ($\bar{p}_{ij}$) which assumes a non-stress condition.

METHODOLOGICAL SCOPE

The general methodology underlying the model (which see below) can be applied to any type of system or task. The model's features (e.g., stress or urgency) apply regardless of the specific nature of the system. The psychosocial variables also generally apply. The nature of the relationships among parameters also assume the status of general rules; for example, the effect of stress on performance, i.e., that up to some threshold point stress is organizing (and positively beneficial, at least in terms of speed of response and subtask success probability) whereas beyond that point it is disorganizing.

The application of the general model to represent a specific system does, however, require the collection of new data characteristics of that system. For example, subtask execution times will vary depending on the specific nature of the task. This forces the necessity of gathering new input data in the application of the model to a new system, and therefore makes it unlikely that a "universal" data bank will be useful; or, if it can be, it must be very extensive to cover the very many task idiosyncracies to be found in diverse systems.

We can say therefore that the overall model is generally applicable and is not constrained by limitations of decision-making and continuous tasks as found in TEPPS, for example. It is one of the advantages of the simulation process that it is not constrained (as are the methods previously considered) by the combinatorial limitations of probability statistics. The simulation takes interactive and feedback processes into account as it performs model operations.

The model has been applied to diverse systems such as landing an aircraft, firing a missile, searching out, detecting and classifying submarines, and re-entering the atmosphere in a space craft. The model has been used primarily with operator activities, but there seems to be no a priori reason why it could not be used for maintenance operations, although here a difficulty might be encountered in getting adequate data. Siegel notes that the model has been employed successfully for a number of maintenance simulations.

The model has generally been applied to missions lasting less than one hour, although the larger models involve 24 hour operations up to 90 days.

PARAMETERS

The following are the basic parameters of the model. Four principal parameter values are specified for each simulation run.

The parameter T_j, the mission time limit, specifies the total time allotted to each operator for performance of the task. This parameter is similar to the SER employed by TEPPS. For a two-man team, the task is considered to have been successfully completed only if both operators complete all required subtasks within their respective time limits.

The parameter F_j is an individuality factor for each operator; it accounts for variance among individuals operating the system. This parameter provides the ability to simulate an operator who usually performs faster or slower than the average operator for whom an F_j value of unity is assigned. The effects of faster, or more highly motivated operators ($F_j < 1$), and slower operators ($F_j > 1$) in the performance of the task are examined by performing several computer runs with different F_j values. The range of values for F_j from 0.7 to 1.3 has been found to be practically useful in simulations.

A third parameter which is central to the model is the stress threshold (M_j). Stress here is a central process (i.e., the "certainty" in the operator's mind that there is insufficient time to complete essential subtasks) defined as the operator's state of mind prior to his initiation of an essential subtask. However, it is operationally defined as the ratio of how much is left to do to the amount of time available in which to do it.

Initial stress build up is recognized in the model as having an organizing effect on operator performance as long as the value of stress remains less than M_j. When stress exceeds M_j, the effect is disorganizing. M_j can therefore be considered as the operator's breaking point. An M_j value of 2 indicates that the operator begins to work slower and less accurately at the point at which he has more than twice as much work to do (at average speed) as he has time in which to do it. Prior to this point, any backlog of essential subtasks creates a stress factor that makes his actions faster and more accurate.

The critical importance of stress is indicated by its relationship to probability of successful performance of the subtask ($\bar{p}_{ij}$). Thus the probability of success increases linearly with stress from a value of

$\bar{p}_{ij}$ until it assumes a value of unity at the stress threshold. Following this point, the probability assumes the average value $\bar{p}_{ij}$ after which it decreases linearly until, when stress has a value equal to $M_j + 1$, it levels off at a value which is decreased from $\bar{p}_{ij}$ by an amount equal to $\bar{p}_{ij}/2$.

Similarly, execution time for the subtask varies as a function of stress. The average operator requires $\bar{t}_i$ seconds to perform subtask i when stress is unity. It is assumed that actual subtask time is normally distributed with a mean dependent on $\bar{t}_i$ and $\bar{\sigma}_i$. $\bar{t}_i$ and $\bar{\sigma}_i$ are used unchanged when stress = unity; is decreased with increasing stress (via an empirically determined cubic equation) until M_j is reached; is used unchanged when stress equals M_j; and is increased linearly with increasing stress beyond M_j.

In his later work (Ref. 6), Siegel has added a fourth parameter, the time period P_j, which is applicable only to cyclic subtasks, those found, for example, in radar or sonar systems in which the equipment imposes a time, e.g., scan time, before which the operator cannot initiate his subtask. When such a subtask occurs in the task sequence, the operator must wait until the start of the next period before he can begin that subtask. This is the waiting period P_j.

The above represent the parameters to be found in all three models. In addition, in the larger models one finds such psychosocial parameters as leadership, group and crew size, equipment data, e.g., failure rate, repair time, etc., personnel data, such as areas of specialization, morale threshold, number of working hours per day, probability of emergency situation occurrence, etc. We will not endeavor to discuss these because we do feel that they would be included only during the development of the largest systems. It is our feeling that the model can be applied with or without any of these more complex variables, depending on the kind of question to be asked of the model and the availability of relevant data.

Several assumptions inherent in the above parameters need verification. The assumption that probability and execution time increase and decrease linearly with stress is probably an oversimplification, but acceptable until more precise information is available from experimentation. Some minor modifications in this formulation may be found in Reference 6.

DATA

1. Measures Employed

To exercise the model, 17 items of task analytic input data are needed for each subtask and each operator. These are punched on cards for input to the computer.

a. Operator number: j = 1 or 2, identifies the operator who is assigned to the subtask.

b. Subtask number: i, an integer that identifies the assigned subtask.

c. Type of subtask: a code indicating one of four special subtask types. Any type can appear without restriction wherever desired in the task sequence. A joint subtask (type = J) is one performed simultaneously by both operators; for example, a communication task is simulated simultaneously with one operator talking and one listening. An equipment subtask (type = E) is introduced to account for a delay in the task because of factors other than human performance (for example, to simulate an equipment warmup). No operator stress functions are calculated for this type of subtask. A decision subtask (type = D) is incorporated into the sequence to cause branching, skipping, or looping in the task sequence to simulate a choice made by an operator without the operator taking any action. A cyclic subtask (type = C) requires an operator to wait until the start of the next periodic time interval before he can initiate the subtask.

d. Indication of subtask essentiality: an indicator specifying whether or not the successful performance of the subtask is essential to successful completion of the task. This allows the computer to identify and ignore nonessential subtasks during "highly urgent" conditions. (E = essential; N = nonessential.) (In recent modifications of the model, this input has been expanded to provide for 10 levels of essentiality.)

e. Subtask precedence: d_{ij} (mnemonic delay): a number indicating a subtask that must be successfully completed by his partner before an operator can begin the current subtask. By proper selection of d_{ij} values, it is possible to cause either operator to "wait" until his partner has completed a stipulated subtask successfully.

f. Time precedence: I_{ij}: the point in time before which operator j is not permitted to begin subtask i.

g. Next subtask, success: $(i, j)_s$: the subtask to be performed next by operator j if he succeeds on subtask i or if he selects the first alternative course in a decision subtask.

h. Next subtask, failure: $(i, j)_f$: the subtask to be performed next by operator j if he fails at subtask i or if he chooses the second of two alternative courses in a decision subtask.

i. Average subtask execution time: $\bar{t}_{ij}$: the average time required by the jth operator to perform subtask i. This average value represents the case in which the operator is under no stress.

j. Average standard deviation: $\bar{\sigma}_{ij}$: taken around the mean $\bar{t}_{ij}$ for the average operator while not under stress.

k. Average subtask probability of success: $\bar{p}_{ij}$: the probability that the average operator j while not under stress can perform subtask i successfully or that he will select one or another course of action in a decision subtask. For most subtasks, probabilities of 0.97 and above have been found to be appropriate.

l. Time remaining, essential: T_{ij}^{E}: the time required to perform all remaining essential subtasks (including i) at average execution times, assuming no failures.

m. Time remaining, nonessential: T_{ij}^{N}: the time required to perform all remaining nonessential subtasks (including i) at average execution times, assuming no failures.

n. Indication of two special subtask types: the allowance for one operator to make a decision that will decide the sequence of future subtasks for both operators. The first enables each operator to jump to an individually specified subtask, depending on what the operator does. The second type of subtask provides a team decision capability to the model.

o., p. Next task numbers: like g. and h. above, for use on special subtasks.

q. Goal aspiration: $G_{i,j}$: the performance level at which operator j is satisfied with his performance on subtask i. Purely optional.

2. Data Sources

The level of data input, as represented by Table 1 in Reference 5, is fairly molecular; that is, it describes individual discrete perceptual and motor actions; indeed, it reminds one a bit of the AIR Data Store. This is understandable because the subtask is essentially a task element. From that standpoint, one would think that a "universal" data bank of the AIR type would be extremely useful. (In fact the AIR Data Store has been used by Siegel on some occasions.) On the other hand, since the model must describe in detail the functioning of a specific system, one might also think that more specific data (specific, that is, to the system being modeled) would be needed.

Sources of input data are varied. Data are secured from "task analysis, formal experiments, informal measurements, simulator measurements, literature search or personal interviews" (p. 13, Ref. 6). Again, Siegel, in referring to the assignment of subtask success probabilities says, "We have relied largely on logic, a knowledge of the characteristics of the subtasks under consideration (italics those of the author) informal observations and interviews with systems operators" (p. 15, Ref. 6).

The author gets the feeling that most of the input data are gathered by direct questioning of operators (to provide expert judgment) and that, in contrast to TEPPS, for example, the data gathering process is relatively informal. Although Siegel undoubtedly makes use of data banks (such as they are), it is likely that some new input data must be gathered for each new application of the model. Siegel notes that "on the other hand, experience with the model has indicated that it is relatively insensitive to input data vagaries" (personal communication).

In consequence, there is probably an element of error in both $\bar{p}_i$ and $\bar{t}_i$ data used to exercise the model. However, this error, particularly for $\bar{p}_i$, may be relatively less important than in other models. "For most subtasks probabilities of 0.97 and above have been found to be appropriate" (p. 15, Ref. 6). Because subtask probabilities are generally very high, and because of the multiplicative limitations of probability statistics are minimized in simulation models, the effect of these probabilities on the estimate of overall system success (in terms of reducing that estimate) is minimized. In addition, since the overriding parameter in the model is time, it is likely that success probability has less influence on model operations than the other techniques previously considered.

Like the previous techniques considered, the type of data input does not differentiate significantly between equipment characteristics (e.g., two different types of meters), although it does differentiate between types of equipment components (e.g., indicator lights and meters). This has negative implications for use of the model results to suggest design modifications, but it makes the model capable of using almost any kind of data source; an operationally gathered data bank of the type being developed by NELC [3] would therefore be useful. However, in view of the model's need for system-specific data, such a data bank would probably not supply all necessary data.

Things both positive and negative need be said about the Siegel model data inputs. For example, since the model makes use of the same type of data as the other techniques reviewed, its data are subject to the same qualifications: the probability estimates applied to the subtasks are not reliability estimates as one ordinarily thinks of reliability estimates (derived from error as a function of time), but rather 1 minus the percentage of error over a block of trials. This data inadequacy does not, however, have any apparent effect on model precision (probably because we have no way of testing for that effect).

The fact that additional data must be gathered for each new model application is disquieting because it takes time to gather the data, and the opportunity for error in gathering those data (particularly from "expert" judgments) is always present. However, this is more or less true of the models examined. Siegel notes (personal communication) that usually only a limited amount of new data must be acquired. Much of the data secured in past use of the model can be applied anew. Moreover, the model's data gathering requirements make it less necessary to establish a "universal" data bank, and thus relieves the model user of an onerous burden.

---

(3) See Coburn, R., "A Human Performance Data Bank for Command Control," in proceedings, U. S. Navy Human Reliability Workshop, Report NAVSHIPS 0967-412-4010, February 1971, pp. 276-282.

3. Output Metric

The model outputs a considerable amount of data for each operator which are listed on page 31 of Reference 6. We do not include them here because of the large number of these measures. These data are organized to provide the following dependent variables (per run):

a. average time expended;
b. average peak stress;
c. average final stress;
d. probability of task success;
e. average waiting time;
f. sum of subtasks ignored;
g. sum of subtasks failed.

These data can be plotted as a function of the following independent variables:

a. time available;
b. stress threshold;
c. speed factor.

It is apparent that the simulation model provides considerably more data than do the non-simulation techniques previously reviewed. Although from a system effectiveness standpoint we are interested primarily in time expended (completion time) and probability of task success (which the other techniques also provide), the other measures supplied by the Siegel model are of considerable interest in analyzing (diagnosing) the conditions that led to a given performance. The fact that one can plot the above variables as a function of each other is a capability which the previous techniques do not possess.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Methods

Like the previous models considered, not much is said about the manner in which the basic behavioral unit (the subtask) is abstracted from the overall task or mission. Presumably these units are determined

as the end product of a detailed task analysis, but the details of the analysis are assumed to be known by the user.[4]

The subtask is the lowest level behavioral operation possible, e.g., throws toggle switch, reads instruments. These subtasks can be accomplished in from several seconds to a few minutes (which suggests that some subtasks requiring the longer time are more molar than task elements). This molecular level of operation is required because the simulation must reproduce each individual operator action in real life.

No graphic method of organizing these subtasks in terms of something like a probability branching tree (THERP) or a GSSM (TEPPS) is required, because they are not needed to translate the model operations into mathematical form; the computer program does this directly.

2. Methods of Synthesis

The methods previous examined (the AIR Data Store, THERP, TEPPS) broke the mission or the task down into smaller behavioral units (i.e., task elements and tasks). It was therefore necessary to determine how these smaller units were recombined or synthesized to provide an estimate of system effectiveness for the larger unit.

This procedure is unnecessary for the digital simulation methods because, although analysis is required to prepare subtask inputs for the simulation, the exercise of the model itself serves as the combinatorial process; or perhaps one can say that combination is unnecessary.

Let us say, for example, that subtasks 1-30 must be performed in order to achieve the task output. In a non-simulation model, like THERP, it would be necessary to apply statistical formulae, e.g.,

$$Q_T = 1 - \left[\prod_{k=1}^{n} (1-Q_k)\right]$$

---

(4) Siegel notes: We are anticipating the preparation of a manual on how to do the input analysis this year. It isn't as hard to do as your text might imply. We have had about eight different people doing them, and no one has had any particular problem in understanding what to do or how to do it.

to indicate how the probability values applied to the 30 subtasks were to be combined to derive the estimate of system effectiveness. This combinatorial process brought us face to face with the problems of independence/dependence relationships.

In the simulation model the simulation itself, just like the operator in real life, operates through its Monte Carlo sampling process to arrive directly at the end result we are looking for. Note also that success or failure of the entire task or mission is not excessively dependent on the probability of accomplishment of any single subtask, but whether or not the operator completes all essential subtasks in the required time. Each individual subtask p_i has an effect on ultimate system success, but not necessarily a primary one.

As a consequence, all one has to do at the end of a series of computer simulation runs is to count the total number of run iterations and the number of successful iterations and divide the second by the first to arrive at the desired estimate of effectiveness. In fact, this estimate is provided by the computer.

Obviously, the fact that the simulation makes it unnecessary to model the combinatorial process in probability mathematics represents a significant advantage over the non-simulation models. As has been pointed out many times by others as well as by this writer, our understanding of combinatorial rules is highly limited.

In place of a discussion of the combinatorial process, we will describe (in very abbreviated form, of course) the simulation sequence. That sequence is graphically illustrated by Figure 11.

"Because the simulation of any individual task is based in part on a random process, it is necessary to repeat the simulation many times to obtain sufficiently representative performance data for each set of conditions" (p. 18, Ref. 6). A value of N, representing the number of times a given task is to be simulated, is selected prior to the simulation. N usually varies from 100-200.

"Another initial condition is R_O, the 9-digit number from which the computer generates subsequent pseudo-random numbers needed during the course of the simulation. The term "pseudo-random" is used because the last number generated in one run is used as the first value in the next run and thus any random number generated is not wholly independent of

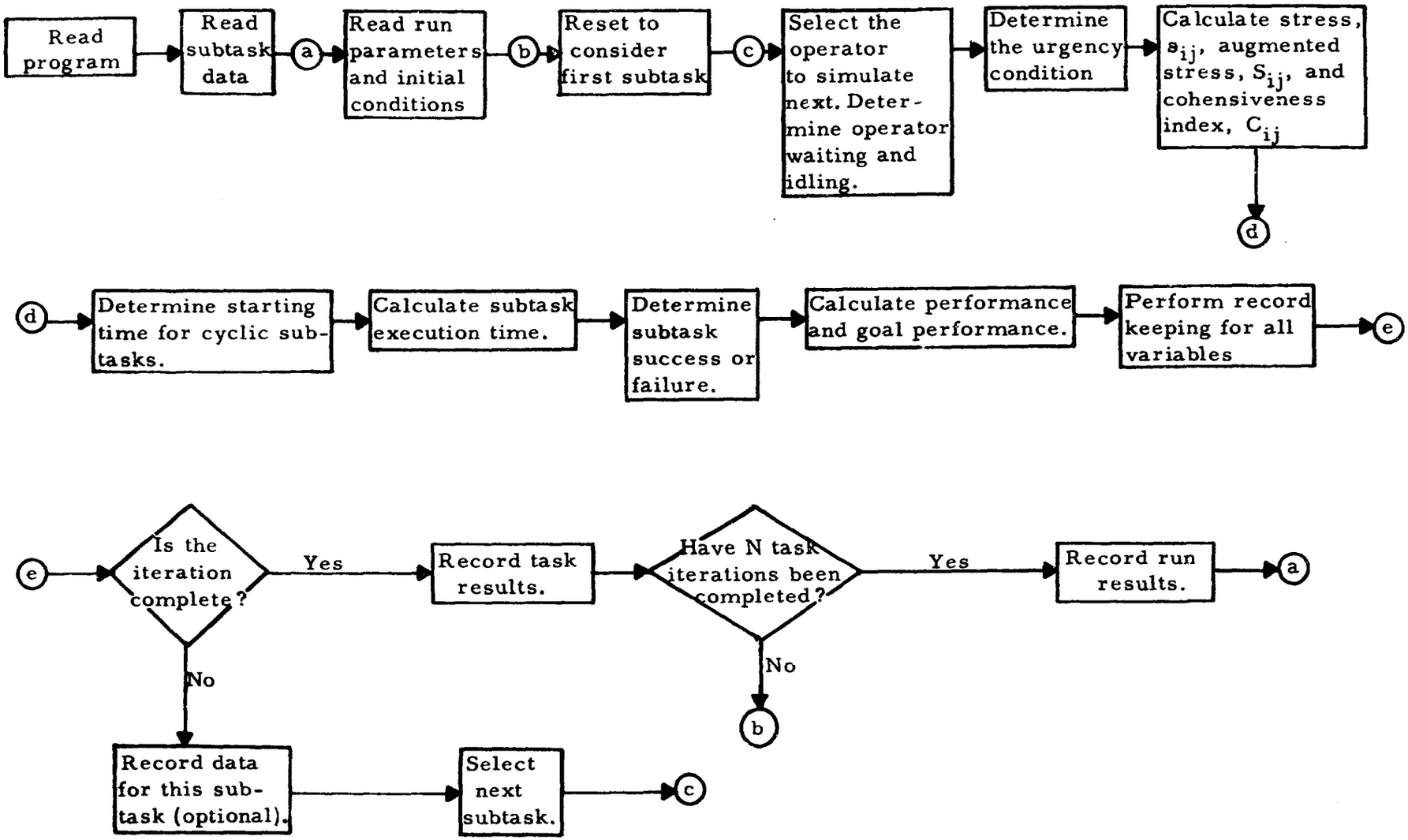


FIGURE 11. General flow chart of the one or two operator man-machine model. (Taken from Ref. 6)

the last. The distributions of pseudo-random numbers as generated are indistinguishable by reasonable statistical tests from numbers which result from a truly random or stochastic process" (p. 559, Ref. 5).

The pseudo-random numbers are used to determine alternative courses of action: given probabilities of the alternatives which sum to unity, "a pseudo-random number equi-probable in the range 0-1 will determine to which action it corresponds" (p. 19, Ref. 6). Where a value for a variable is known or assumed to have a particular statistical distribution (e.g., normal, Poisson, Weibull), the pseudo-random number is used to select a sample value from the distribution.

Once the program, parameters and initial conditions have been stored by the computer, it begins to process subtask data sequentially. The sequence of subtasks to be performed is determined in accordance with the operator's success or failure on a prior task and the total time expended by the operator on all previous subtasks. At any given time, the operator who has expended less total time is selected and his next subtask is simulated. If the selected operator must wait for his partner, the sequence continues using data for the other operator.

One of three stages of "urgency" is next determined, based on the remaining time available to the operator for completing the task. The situation is non-urgent if there is sufficient time to complete all remaining subtasks; it is urgent if time is available only for completing essential subtasks; it is highly urgent if there is insufficient time for completing even essential tasks. In the latter two conditions the computer ignores non-essential subtasks.

(There are systems in which all subtasks are essential in the sense that the task cannot be performed if any one of them is not completed. This does not void the model logic, however; it simply means that no subtasks can be ignored by the computer simulation.)

Following determination of the degree of urgency, the stress condition is calculated. During non-urgent and urgent conditions stress is defined as equal to unity. When the situation is highly urgent, stress is defined as the ratio of the sum of average execution (completion) times for remaining essential subtasks to the total time remaining.

Subtask execution time is next computed. For each subtask it is assumed that the actual subtask execution time is normally distributed.

Specific time values are selected by the Monte Carlo technique from a normal distribution limited by a fixed minimum, 0.75 second. The selected time values are used unchanged when stress equals unity, are decreased per third degree polynomial as a function of increasing stress until stress assumes the threshold value (M_j), used unchanged when s stress equals M_j, and increases linearly with increasing stress beyond threshold until stress equals $M_j + 1$ and assumes a value of $2\bar{t}_{ij}$ beyond that point.

The probability of subtask success and failure is then generated in much the same way as are the subtask execution times, with essentially the same time-stress relationships.

The above description of how the simulation process is accomplished is necessarily lacking in detail. In particular we have not considered team cohesiveness, the various types of subtasks, etc. Readers interested in specific details are referred to Refs. 5 and 6.

It is apparent that the "principal model variable is stress" (p. 29, Ref. 6). If one objects to the use of this variable in predicting system effectiveness, it is possible to exercise the model without time stress. In this case one sets an arbitrarily large value for required execution time so that even with a large number of subtask failures no stress build-up occurs. The developers report (p. 29, Ref. 6) that it is standard practice to make at least one such no-stress run in each series of trials as a reference for the stress condition. The results of such no-stress runs should be to provide estimates of task success which are overly optimistic, but pessimistic from the point of view of time; a comparison of no-stress run results with stress run results should provide a sort of "confidence level" for the system effectiveness estimate.

ANTICIPATED MODEL USES

1. Prediction of System Effectiveness

Under this heading we consider whether the model can supply

a. An absolute estimate of the system reliability to be anticipated when the system becomes operational (e.g., the system will eventually perform with a reliability of .99).

b. A comparison of estimated system reliability with that required to achieve mission success (e.g., estimated system reliability is .98, but mission success requirements call for a performance of .99; or, the mission must be accomplished in 32 minutes, whereas estimated system performance time is 34 minutes).

Both of these assumed model capabilities are identical with those claimed for the techniques previously considered (the Data Store, THERP, TEPPS).

It is obvious from what has been said earlier that the simulation model will provide estimates of system performance both for probability of success and task completion time. A comparison with system requirements is easy enough to provide, assuming that a mission requirement in terms of human performance has been specified.

In considering the above capabilities we have implicitly assumed that these predictive estimates would be made during early system development. One may question whether the necessary input data can be secured early enough in system development to make the model results useful. Since the input data needed are highly specific to the particular system being modeled, it will be necessary for the system developer to elaborate his design concept in detail in order to provide the necessary data. However, our studies (Ref. 3) indicate that the system developer, when pushed to do so, can provide these data quite early.

It is the author's feeling after reviewing the non-simulation models that the simulation model does not require significantly more detailed information than the other models. True, both THERP and TEPPS may require data at a slightly less detailed level, which should make it somewhat easier for the user of these models to get the necessary data; on the other hand, the amount of additional effort required to secure input data for the simulation model does not seem excessive to us, because the method employed to secure data for the Siegel model is essentially the same as that used by THERP and is not as rigorous (and hence presumably easier to apply) than that employed by TEPPS.

However, since the effort factor is likely to be a constraint on whether the system developer employs a particular technique, it would be very useful if one could get realistic time/manhour estimates from model developers as one basis for comparing the techniques. In an ideal sense, this effort factor should play a very minimal role in selection of a technique for use; but we suspect that it plays a greater role than one would like.

Since all the models so far considered purport to perform the same functions, a choice among them would seem to devolve upon the validity and precision of their outputs and the feasibility of employing the respective models in early design. Validity and precision will be considered later. With regard to feasibility of model use, there seems to be little difference between the efforts to develop a probability branching tree, a GSSM and the input data for the digital simulation model. It is assumed that the basic computer program (with perhaps slight modifications to tailor it to a particular system?) would be available for use with any specific system. The developers can perhaps give us a better idea of how long it takes to prepare the input data and program for a computer run.

2. Design Analysis

Under this heading we include several aspects:

a. The comparison of alternative system configurations to determine which should be selected for implementation.

b. The determination of redesign requirements for a system which cannot satisfy system requirements.

c. Recommendations for initial design, design changes, selection of components, equipment layouts, etc.

With regard to the comparison of alternative system configurations, it is apparent that if the simulation model can predict the system effectiveness of one configuration, it can also predict the effectiveness of another, and compare the two estimates. (As with the other techniques considered, differences in system configuration will reveal differences in performance only if the nature and organization of the tasks involved in the two configurations differ; if the two configurations differ only in terms of molecular equipment characteristics, it is unlikely that significant differences in performance will result because the input data are responsive more to task factors than to equipment factors.) The Siegel model is perhaps more sensitive to such a comparison between configurations than the previous models, especially as regards execution time requirements, the ratio of essential/non-essential subtasks, and decision and multiple action subtasks. We say "sensitive" because one can develop with the simulation model a function (i.e., a graphic plot) which relates

performance to any of a number of values assumed by these inputs. This can be done also to some extent by non-simulation models (e.g., ranking of Q_i values in THERP) but at greater cost because these models lack the flexibility of computer processing.

The digital simulation model, like most non-simulation models, lacks the capability to point out the equipment design implications of a system inadequacy. It may be easier with the simulation model to determine the point in time at which performance starts to degrade (e.g., from the graphic plot), but having determined that degradation or inadequacy exists, this does not indicate what should be done to remedy the situation. As with the other models, the input data for the simulation model do not appear to be overly sensitive to equipment characteristics; hence if the problem is one of equipment design rather than of task conditions, the model can only suggest the existence of a problem; further human engineering analysis of the situation is required.

This model, like the others, does not provide recommendations for initial design. It should be noted that the simulation model, again like the others, does not pretend to supply any such recommendations. It assumes a system configuration, rather than suggesting one. If this is an inadequacy, it is an inadequacy which all the predictive models reviewed contain.

3. Selection and Training

The author feels that almost all the models reviewed provide relatively little information about these factors. About his own model, however, Siegel reports that it "has been employed to derive training information and the outputs yield required levels of proficiency to be attained as a result of training". The discrepancy in viewpoint may be a difference in the interpretation of what constitutes meaningful information. Certainly inadequate performance in a system simulation (e.g., inability to complete certain subtasks in required time) may suggest the need for better selection or more personnel training. Since the simulation model is sensitive to variations in personnel capability (the F variable), it accordingly provides information relative to a need for a change in selection and training. The question for the author (which is unanswered) is, how detailed will that information be? Will it suggest the types of personnel to be selected, their required aptitudes, the amount of training they should receive, the subject matter which training should emphasize, etc.?

To anticipate our later argument, it would appear that no single model will perform every desired function. The models we have considered so far appear to predict system effectiveness and to compare alternative configurations to a greater or lesser extent. However, they are distinctly lacking in ability to satisfy specific design, selection and training requirements. One can therefore raise the question whether it is fair to ask any single model to handle all the varied problems that the system developer may encounter. Perhaps a number of different models or techniques (specifically developed to satisfy particular requirements) should be considered.

VALIDATION /APPLICATION STUDIES

Of the techniques so far considered only the simulation model has been exposed to any number of validation studies. We are talking here about empirical rather than construct validity. The validations performed on the previously reviewed techniques have been at best partial ones, in which comparison with operational performance has usually been lacking.

The validation studies performed on the digital simulation model have, on the contrary, been numerous and varied. (Why the developers of this model were more fortunate in having the opportunity to validate is something we cannot go into here.) Although a precise correspondence between the model predictions and operational performance has not been found (nor should one expect to find precise correspondence), the results of the validation studies (reviewed in Refs. 5 and 6) are highly promising.

Something that might be termed a "concurrent validity" study has recently come to the author's attention. This is a study to develop a digital simulation model for fighter pilot workload (Ref. 1), which was based on the Siegel model and used essentially the same input data but with certain advances (presumably) in computer programming. Presumably the new approach provides "additional distribution and simultaneous task information which have not yet been validated" (p. 87, Ref. 1). However, the new model can be considered only a variation of the Siegel model. The point of interest to us is, however, that extremely high correlations (better than .90) were found for those elements in common between the two model variations.

There exists then a fairly substantial body of evidence which attests to the validity of the 1-2 man digital simulation model, a body of evidence which does not exist for the other models/techniques so far reviewed.

EVALUATIVE SUMMARY

Validity - Formal studies have been performed which show a reasonable degree of correspondence between predicted and observed values. Of the various simulation techniques, this is the one which has been most completely validated and which in consequence serves as the pace-maker for other simulation models. However, much less is known about the use of the model in solving actual system development problems.

Reliability - Controlled studies have been performed that indicate that the method can be applied with acceptable consistency by various analysts.

System Development Applicability

A. Comprehensiveness: The method is not limited in its application to systems, tasks or behaviors (including both operation and maintenance applications).

B. Applicability: The method outputs a prediction of system effectiveness and is reported to provide data useful for other purposes (design, selection, training).

C. Timing: The model can be applied at all stages of system development.

Model Characteristics

A. Objectivity: Relatively few judgments are required.

B. Structure: Conceptually elegant and well described.

REFERENCES

1. Asiala, C. F. Digital Simulation Model for Fighter Pilot Workload, Report MDC A0058, McDonnell Aircraft Company, St. Louis, Mo., 19 September 1969.

2. Chapanis, A. Men, Machines and Models. American Psychologist, 1961, 16, 113-131.

3. Meister, D. and Sullivan, D. J. A Further Study of the Use of Human Factors Information by Designers. Final Report, Nonr-4974-00, Bunker Ramo, Canoga Park, California, March 16, 1967.

4. Siegel, A. I. and Wolf, J. J. A Technique for Evaluating Man-Machine System Designs. Human Factors, 1961, 3(1), 18-28.

5. Siegel, A. I. and Wolf, J. J. A Model for Digital Simulation of Two-Operator Man-Machine Systems. Ergonomics, 1962, 5(4), 557-572.

6. Siegel, A. I. and Wolf, J. J. Man-Machine Simulation Models: Psychosocial and Performance Interaction. Wiley, 1969.

APPENDIX

List of Reports on Man-Machine Simulations

Item	Title	Authors	Date	No. of Men Simulated	Task Simulated	Type of Report
1	Development of a Digital Computer Technique for Evaluating the Operator Loading in Man-Machine Systems	Arthur I. Siegel J. Jay Wolf	May 1958	1	--	Preliminary project report
2	*A Description of a Model and the Results of its First Application	Arthur I. Siegel J. Jay Wolf	Feb. 1959	1	Carrier landing	Formal project report to ONR
3	*Application of a Previously Derived Model to the Launching of an Air-to-Air Missle	Arthur I. Siegel J. Jay Wolf	June 1959	1	Air-to-air missile launching	Formal project report to ONR
4	*Description of a Model to Simulate a System Manned by Two Operators	Arthur I. Siegel J. Jay Wolf	Jan. 1960	2	--	Interim project report
5	A Technique for Evaluating Man-Machine System Designs	Arthur I. Siegel J. Jay Wolf	Jan. 1961	1	Two above	Journal article Human Factors, Vol. 3, No. 1
6	*A Modelfor Digital Simulation of One and Two-Operator Man-Machine Systems	Arthur I. Siegel J. Jay Wolf Kenneth Crain	March 1961	2	In-flight refueling	Formal project report to ONR

* Common major title: Techniques for Evaluating Operator Loading in Man-Machine Systems

Item	Title	Authors	Date	No. of Men Simulated	Task Simulated	Type of Report
7.	*A Further Application of a "Model" for Digital Simulation of One or Two-Operator Man-Machine Systems	Arthur I. Siegel J. Jay Wolf	June 1961	2	Air-to-air intercept	Formal project report to ONR
8	Computer Simulation of Man's Performance in Man-Machine Systems	Arthur I. Siegel	June 1961	2	--	Naval Research Reviews article
9	*Evaluation of a One or a Two-Operator System Evaluative Model Through a Controlled Laboratory Test	Arthur I. Siegel J. Jay Wolf R. T. Sorenson	July 1962	2	Synthetic man-machine task (model railway system)	Formal project report to ONR
10	*A Model for Digital Simulation of Two-Operator Man-Machine Systems	Arthur I. Siegel J. Jay Wolf	Oct. 1962	2	In-flight refueling and air-to-air intercept	Journal article Ergonomics, Vol. 5, No. 4
11	*Modification and Further Evaluation of a Digital Man-Machine Simulation Model	Arthur I. Siegel J. Jay Wolf	July 1963	1 (2 man model)	Air-to-air missile Launching	Formal project report to ONR
12	*Further Test and Evaluation of a Man-Machine Simulation Model	Arthur I. Siegel J. Jay Wolf R. S. Lantherman	Oct. 1963	2	Synthetic man-machine task (model railway system)	Formal project report to ONR
13	Computer Simulation of Man-Machine Systems	Arthur I. Siegel J. Jay Wolf	1963	2	--	Chapter in Unusual Environments and Human Behavior, Free Press of Glencoe
14	Application of a Digital Computer Simulation Technique to Maintenance of a Sonar System (CONFID)	Arthur I. Siegel M. A. Fischl	Apr. 1967	2	Sonar maintenance	Project PAIR report (Naval Ship Systems Command- ONR)

Item	Title	Authors	Date	No. of Men Simulated	Task Simulated	Type of Report
15	Verification of a Digital Technique for Sonar Operation Simulation	Arthur I. Siegel D. H. Macpherson	May 1967	1	Sonar operation	Project PAIR report (Naval Ship Systems Command-ONR)
16	Application of a Digital Machine Simulation Model during the Development of the PAIR System (CONFID)	Arthur I. Siegel M. A. Fischl	Aug. 1967	2	Sonar operation	Project PAIR report (Naval Ship Systems Command-ONR)
17	Recent Revisions to the Digital Simulation Model for Simulating Two-Operator Man-Machine Interaction	Arthur I. Siegel J. Jay Wolf	Nov. 1967	2	Sonar operation	Project PAIR report (Naval Ship Systems Command-ONR)
18	Prediction of Individual and Crew Performance by Computer Simulation	Arthur I. Siegel	Jan. 1967	2	Various	Symposium on Human Performance Quantification in Systems Effectiveness
19	Computer Simulation Savior: Sanctuary or Silliness?	Arthur I. Siegel J. Jay Wolf	Nov. 1968	2	Various	Symposium on Applied Models of Man-Machine Systems Effectiveness
20	Man-Machine Simulation Models	Arthur I. Siegel J. Jay Wolf	1969	2	Various	New York: Wiley
21	Operator Training Requirements through Digital Computer Simulation (CONFID)	M. A. Fischl J. Jay Wolf Arthur I. Siegel	Dec. 1969	2	Sonar operation	Project PAIR report (Naval Ship Systems Command)
22	Digital Model Modification on On-Line Simulation and Further Extensions of Operator Loading Treatment	Arthur I. Siegel J. Jay Wolf Wm. Miehle	July 1969	--	--	Project MISS report (Human Engineering Div. AMRL)

Item	Title	Authors	Date	No. of Men Simulated	Task Simulated	Type of Report
23	Vulnerability/Survivability Estimation through Computer Simulation: I. Program Logic, Task Analysis, and Initial Simulation for the F-106 Aircraft (CONFIDENTIAL)	W. R. Leahy J. Jay Wolf Arthur I. Siegel	Dec. 1969	1	1	Project BATH report (Human Engineering Division, AMRL)

II. THE TACDEN DIGITAL SIMULATION MODEL

INTRODUCTION

This section describes a predictive technique which has similarities to Siegel's 1-2 man digital simulation model. This model- developed by Dr. Gilbert Miller and his associates- makes use of a digital computer to simulate the behavioral processes involved in entering messages on a device known as the TACDEN (AN/MSQ-19). This device consists of a typewriter-type keyboard for data entry, a magnetic drum for format storage, and a CRT for visual display of the data.

A thumbnail description of the development and simulation process would include as highlights the following steps with which the reader is now familiar: first a task-equipment analysis (TEA) is developed to represent how message entry is performed; based on this analysis, lists of tasks, together with success probabilities and completion times, etc. are entered into the computer on punched cards; finally the computer simulates the entry of various messages by sampling randomly from the input data. A series of computer runs is made and performance is predicted in terms of number of correctly entered messages, numbers of errors in each message, etc.

This description of the above process will not be treated in complete detail for various reasons. First, we consider the model as only an application of the general digital simulation methodology, represented most significantly by Siegel's work; consequently we are interested primarily in the similarities and differences between the Miller and Siegel models. It is possible also to view the Miller model as a validation of the general digital simulation methodology because it attacks the prediction problem in a slightly different manner (than did Siegel) and also because a validation study was specifically included in the Miller study. To the extent that the Miller model produced results similar to those found by Siegel, we may have additional confidence in the general methodology.

SIMILARITIES BETWEEN THE MILLER AND SIEGEL MODELS

In addition to the fact that both models utilize a digital computer, a computer program, a TEA to derive lists of tasks and the sequence in which they are to be performed, a Monte Carlo method of sampling from distributions of input data and a comparison of sampled data with a performance requirement, more specific similarities between the two

models exist. (These similarities are perhaps more understandable because Miller et al. based their work on Siegel.)

(1) The inclusion of a stress factor. Much like the Siegel concept, "when stress is below a defined threshold value it is treated as an organizing agent on behavior, thus the individual's performance is assumed to improve. Above a defined threshold value, the effect of stress is treated as a disorganizing agent, with the assumption that performance of the human operator deteriorates. The stress factor... is an expression of the ratio of operator ability to perform to the workload imposed upon him." (p 22-23, ref. 1). A required corollary assumption is that "the individual will recognize when his capability to perform is exceeded by the requirements to perform" (p. 23, ref. 1). The similarity to the Siegel formulation should be obvious.

Miller's stress formulation involves the following equation:

$$S = \frac{(M_e - n)(T^u)}{T^R(n)} \quad \text{where}$$

T^u = Time used
n = Number of messages processed
$T - T^u = T^R$ time remaining
M_e = Predicted number of messages to be received
u = Urgency level indicator
S = Stress = $\frac{\text{workload}}{\text{predicted output}}$
R = Rate of message transmission = $\frac{T^u}{n}$
W = Number of messages left to enter = $M_e - n$
P = Predicted outputs $\frac{T^R}{R} = \frac{T^R(n)}{T^u}$

Note that stress is defined in terms of the ratio between workload and predicted output. This breaks down into the relationship between the amount of time remaining to process messages and the number of messages still remaining to be processed. Three conditions of urgency are defined by the various levels of stress:

If $S \leq M_1$, $u = -1$, non-urgent;
If $M_1 < S \leq M_3$, $u = 0$, urgent;
If $M_3 < S$, $u = +1$, highly urgent.

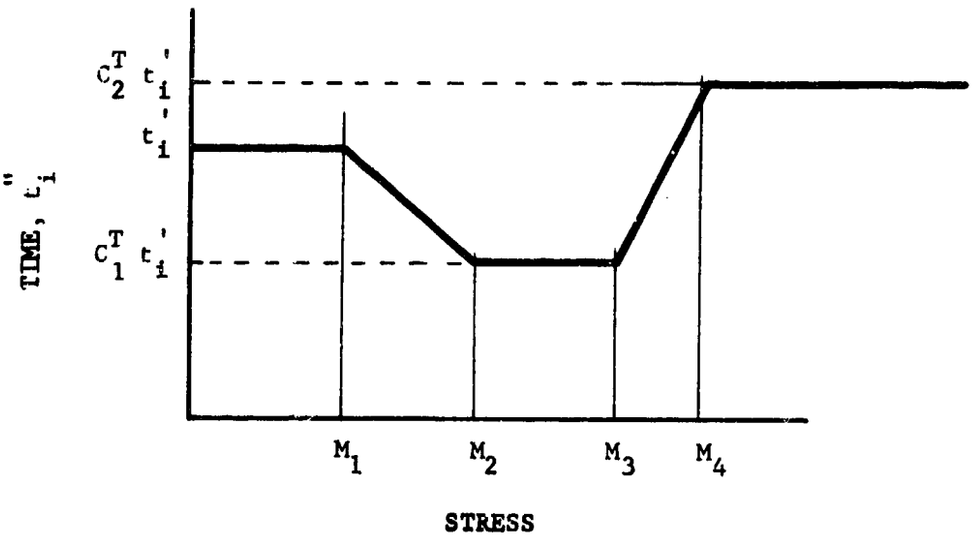


FIGURE 12. TIME VS. STRESS
(Taken from Ref. 1)

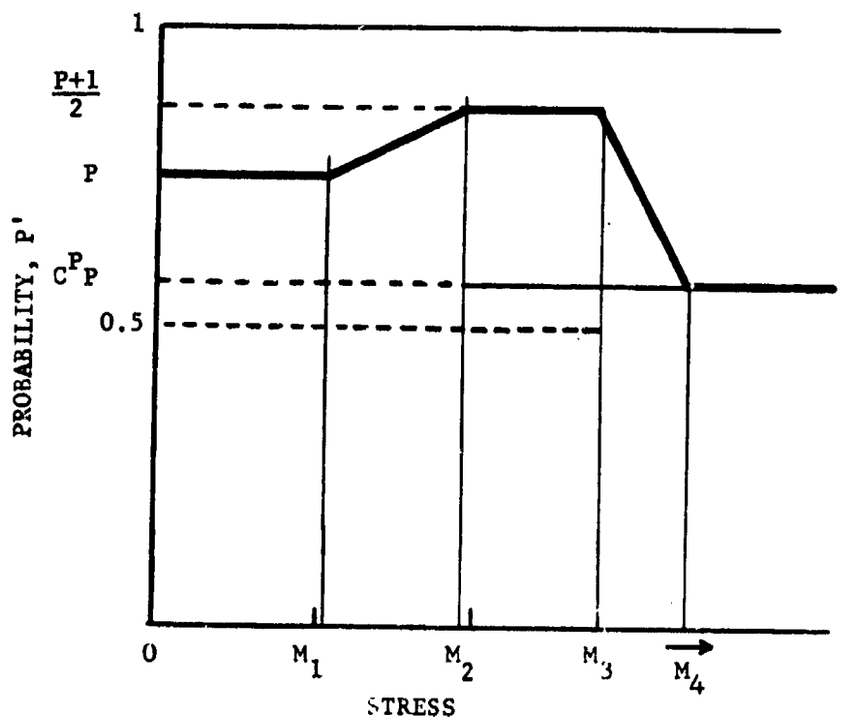


FIGURE 13. PROBABILITY VS. STRESS
(Taken from Ref. 1)

Both predicted performance time and the probability of successfully executing the subtask action also vary with stress, as indicated in Figures 12 and 13 (Ref. 1). The mathematical formulation of these relationships is shown below:

Performance time relationship with stress:

If $S \leq M_1$

$$\bar{t}_i'' = \bar{t}_i'$$

$$\bar{\sigma}_i'' = \bar{\sigma}_i'$$

If $M_1 < S \leq M_2$

$$\bar{t}_i'' = \bar{t}_i' - \frac{(\bar{t}_i' - C_i^T \bar{t}_i')(S - M_1)}{M_2 - M_1}$$

$$\bar{\sigma}_i'' = \bar{\sigma}_i' + \frac{(C_i^S \bar{\sigma}_i' - \bar{\sigma}_i')(S - M_1)}{M_2 - M_1}$$

If $M_2 < S \leq M_3$

$$t'' = C_1^T \bar{t}_i'$$

$$\bar{\sigma}_i'' = C_i^S \bar{\sigma}_i'$$

If $M_1 < S \leq M_4$

$$T_i' = C_1^T T_i + \frac{(C_2^T T_i - C_1^T T_i)(S - M_1)}{M_4 - M_1}$$

$$\bar{\sigma}_i' = C_1^S \bar{\sigma}_i + \frac{(C_2^S \bar{\sigma}_i - C_1^S \bar{\sigma}_i)(S - M_1)}{M_4 - M_1}$$

Probability and stress:

If $S \leq M_1$ $\qquad P_i' = P_i$

If $M_1 < S \leq M_2$ $\qquad P_i' = P_i + \frac{\left[1/2\,(P_i + 1) - P_i\right](S - M_1)}{M_2 - M_1}$

If $M_2 < S \leq M_3$ $\qquad P_i' = \frac{P_i + 1}{2}$

If $M_3 < S \leq M_4$ $\qquad P_i' = \frac{P_i + 1}{2} - \frac{(\frac{P_i + 1}{2} - C^P P_i)(S - M_3)}{M_4 - M_3}$

If $M_4 < S$ $\qquad P_i' = C^P P_i$

Symbols are defined in the section on inputs.

The four stress thresholds (M_1, M_2, M_3 and M_4) apparently represent increasing amounts of stress, but are not clearly defined except that M represents the "size of the queue (of messages) at the current time" (p. 71, ref 1) which suggests that the four stress thresholds are defined in terms of the number of messages remaining to be entered at any one time.

(3) The inclusion of an operator variability factor, F. Operators are classified as fast, average and slow. This rating is input at the start of a run and performance time is modified in accordance with this rating as shown below:

The classification of the operator, F, affects his performance as follows:

$\bar{t}_i$ = Mean of all mean times for a subtask

$\bar{\sigma}_i$ = Mean of the standard deviations for a subtask

$K_{TF} = K_{TS}$ = Multiplier of $\bar{\sigma}_i$ to give mean times for fast and slow

$K_{\sigma F} = K_{\sigma S}$ = Multiplier of $\bar{\sigma}_i$ to give $\bar{\sigma}_i$ for fast and slow operators

$K_{\sigma A}$ = Multiplier of $\bar{\sigma}_i$ to give $\bar{\sigma}_i'$ for average operators

If F = Fast

$$\bar{t}_i' = \bar{t}_i - K_{TF}\ \bar{\sigma}_i$$

$$\bar{\sigma}_i' = K_{\sigma F}\ \bar{\sigma}_i$$

If F = Average

$$\bar{t}_i' = \bar{t}_i$$

$$\bar{\sigma}_i' = K_{\sigma F}\ \bar{\sigma}_i$$

If F = Slow

$$\bar{t}_i' = \bar{t}_i + K_{TF}\ \bar{\sigma}_i$$

$$\bar{\sigma}_i' = K_{\sigma F}\ \bar{\sigma}_i$$

Performance as a function of operator characteristics can then be studied by performing a series of computer runs with different operator ratings included. This procedure too is identical with Siegel's.

(3) Similar inputs. The data required for each task whose performance is being simulated are:

a. subtask number
b. subtask type
c. subtask execution time
d. standard deviation of execution time
e. probability of subtask success
f. essentiality of the subtasks
g. waiting time before the subtask can begin
h. number of next subtask in the event either failure or success
i. operator rating (F)
j. initial queue size
k. rate of message arrival
l. stress thresholds
m. initial random number (R_0)
n. number of runs desired
o. message arrival time in seconds between messages
p. various constants.

Definition of model input parameters are given in Table 10 (Ref. 1.

When the above list is compared with the list of input parameters for the Siegel model, a considerable degree of similarity is observable.

(4) Input data. Input data are derived from the same sources available to other modelers: the AIR Data Store, test data specifically developed to input to the model and subjective estimates. "To use data such as the Data Store in the Miller model, it is necessary to assume some analytical form exhibiting the sample means and variances. Miller assumed a normal distribution defined by a mean and standard deviation. Because negative performance times have no meaning, his distributions were terminated at zero time. In practice, the variances are, in general, small compared with the mean times, so sample times less than zero will not be obtained.

"Miller points out that there are no important practical limitations on either the type or number of distinct distributions that could be used in the model; in fact, raw data could be used. However, the use of a single distribution, such as the normal distribution, greatly simplified the formulation of his computer program and improved its flexibility.

TABLE 10

DEFINITION OF MODEL INPUT PARAMETERS
(Taken from Ref. 1)

k_o	-	the initial random number
T	-	the time period allotted for the operator to enter the messages; T will be given in seconds
m_0	-	initial number of messages in message queue
$\frac{dt}{dm}$	-	message arrival times in seconds between messages
M_1	-	first stress threshold - M_1 must be > 0
M_2	-	second stress threshold - $M_2 > M_1$
M_3	-	third stress threshold - $M_3 > M_2$
M_4	-	fourth stress threshold - $M_4 > M_3$
N	-	number of runs desired
F	-	class of the operator A = average; S = slow; F = fast
C_1^T	-	constant multiplier for minimum time
C_2^T	-	constant multiplier for maximum time
C_1^S	-	constant multiplier for standard deviation at minimum time
C_2^S	-	constant multiplier for standard deviation at maximum time
C^P	-	constant multiplier for minimum probability
K_{σ_A}	-	multiplier times standard deviation for average operator
$K_{\sigma_F} = K_{\sigma_S}$	-	multiplier times standard deviation for slow and fast operators
$K_{TF} = K_{TS}$	-	multiplier times standard deviation to give mean time for fast and slow operators

"To arrive at a distribution of performance times for various classes of operators (given that the performance of the whole population is characterized by a known normal distribution), Miller employed a method which is based on the following assumptions: (1) the performance of a class of operators is described by a sub-interval of time within the range of times for the whole population; (2) the frequency of times for a class will be given by a normal density function whose domain is essentially limited to the sub-interval; and (3) the mean and variances of the density functions will be identical with the expected values and variances of the density function of the whole population restricted to a sub-interval" (p. 636, ref. 2).

DIFFERENCES BETWEEN THE MILLER AND SIEGEL MODELS

Differences between the two models are largely technical. One major difference relates to the technique used to sample from the input distributions. Miller considered the use of the rejection method employed by Siegel, but rejected it in favor of a method for which a computer program (Philco 2000) already existed. "The success for failure to perform a subtask is obtained by straightforward sampling of the step function which is unity for all values less than P and zero elsewhere. Equivalently, one selects a random number (≤ 1) and compares the random number with P. If the number is less than P then success is obtained" (p. 50, ref. 1). It does not appear to us that any differences in the sampling technique employed are significant.

Another difference between Miller and Siegel is the application of information theory to develop a measure of TACDEN effectiveness. "The measure derived is the rate of information processing, and is based on the assumption that the TACDEN operation is equivalent to a discrete, noisy data channel" (p. 62, ref. 1). The measure is written

$$R_t = \frac{I}{t} \quad \text{where}$$

$I = p \log_2 mp + (1-p) \log_2 \frac{m}{m-1}(1-p)$ and

p = probability that the operator reproduces a TACDEN characteristic correctly

m = number of TACDEN characters

t = average time to enter a single character

R_t = rate of transmission of information in bits per second.

MANNER IN WHICH THE COMPUTER SIMULATION FUNCTIONS

A very brief description of the computer simulation process was given at the start of the section. In this sub-section we provide more detail.

The computer program first reads in the task description in the form of coded subtasks (see below for list of these) and the various input parameters and data associated with these subtasks. "Execution of the complete set of subtasks represents the performance of one task sequence by the operator. The program repeats the simulation of a message entry over a specified period of time, T. When the operator has used the entire allotted time, the task is complete and his performance record is saved. This process is repeated N times, where N is an input parameter, before a statistical summary of his performance is output.

"The program initializes for one run by zeroing accumulated totals required for output data. It initializes for a task by setting the number of messages done to zero, the time used to zero, the stress level to the initial condition of non-urgent, and printing start of task and the number. The program initializes the entering of one message by clearing all error indications, setting the program to start on the first subtask. It also prints the message number, the stress, the number of messages in the queue, and the time remaining.

"The program then begins processing successive subtasks. There are ten basic types of subtasks, each one involving different processing steps... Two types, 1 and 10, basically represent an "operation" which is an operator or machine action. Three types, 2, 3 and 4, are decision subtasks; 5 and 6 are error checking subtasks, and 7 and 8 are error clearing subtasks. The last type, 9, indicates that the operator has completed and transmitted the message. Each subtask, beginning with i_o, specifies the subtask to be processed next.

"The program takes each subtask successively until the "end of message" subtask is reached. At this time, the message just transmitted is checked for errors, and the number of messages transmitted is increased by 1. If time still remains, the stress level is determined and the message simulation is repeated. The program, therefore, returns to the point of initialization for a message. If the time has been used, the number of tasks performed is increased by one, and compared

to the number requested. If an insufficient number have been completed, the program returns to the task initialization. If sufficient tasks have been completed" . . the output data are compiled and printed (see below) " (pp. 72-73, ref. 1).

A more detailed description of the above process in relation to the various subtasks dealt with is given on pp. 85-91 of ref. 1.

MODEL OUTPUTS

The following is a list of outputs provided at the conclusion of a series of computer runs:

1. Mean and standard deviation of messages entered in time allotted
2. Percent of messages entered incorrectly
3. Mean and standard deviation number of errors per incorrect message
4. Mean and standard deviation time to enter a message
5. Mean and standard deviation time to enter a correct message
6. Mean and standard deviation time to enter an incorrect message
7. Percent of runs ending in a non-urgent level of stress
8. Percent of runs ending in an urgent level of stress
9. Percent of runs ending in a highly urgent level of stress.

LEVEL OF INPUT DATA

The level of input data is quite molecular, since the success of probabilities and completion times entered into the computer are for subtasks, e.g., position stylus, set switch, observe light, etc. This is required because the computer is simulating individual operator actions, which comprise tasks. This is entirely comparable to the level of data

required by Siegel's 1-2 man model. It must be pointed out again that a very molecular level of input data is required by any simulation methodology because the simulation reproduces individual operator actions. In other words, data are required for the simulation model at the same level of detail as indicated by the individual behaviors of the personnel whose performance is being simulated. This imposes a greater burden for data input than might be required by a non-simulation methodology (although some of the latter also require detailed information). The point is that since a non-simulation model is not required to reproduce the individual behaviors of the operator, it can select any level of input which will provide a reasonable prediction.

VALIDATION STUDY

The validation study performed by Miller et al. is of special interest to us because it reflects on the general validity of the digital simulation methodology. The study used 8 operators who had previous familiarity with the TACDEN device and who knew how to type. The experimental session involved 10 messages and 4 stress levels of 12, 10, 8 and 6 minutes to complete. The same 10 messages were used for each stress level.

Analysis of variance statistics were used to determine the effects of operators, messages and stress level and their interactions on message processing time and number of uncorrected errors. Only the results of the first 5 messages were included in the analysis, since large gaps existed in the data of the remaining 5 messages at high stress levels.

All three primary sources of variation had significant influences on processing time, but stress did not appear by itself to affect the number of uncorrected errors significantly.

What is of more interest to us in the degree of correspondence between the TACDEN simulation predictions of performance and the performance actually achieved by operators. The following table provides the essentials of the comparison.

	TACDEN Simulation (Predicted Values)				Validation Study Results			
	Stress Levels				Stress Levels			
	I	II	III	IV	I	II	III	IV
Message Entry Time (secs)	71.5	67	69	72	77	67	60	58
% Incorrect Messages	17	10	7	12	15	8.8	13	4
Errors per Incorrect Message	1.1	1.0	1.0	1.0	1.5	1.3	1.4	1.0

It is apparent that although there are variations between the predictions and the validation study results, in general the predictions are reasonable approximations of actual performance. The differences are referred by Miller et al. to differences in test conditions between the two situation. For example, the computer used messages of the same length and difficulty, whereas the validation study messages varied in length and difficulty. The operator procedure also differed slightly between the two situations. Considering these differences, the correspondence between the two sets of results is quite good. Miller et al. feel that the rate of information transmission as a function of stress level (see Figure 14, taken from Reference 1) appear to contradict some of the assumptions made by Siegel and Wolfe on the stress formulation, namely that the function is piecewise linear and that the function exhibits discontinuities. It is felt, however, that insufficient evidence exists to alter their assumptions. However, standard deviations, average times and percent errors tend to decrease, then increase with stress as predicted (GRAPHDEN simulation results). The peculiarities of the TACDEN situation might well produce some differences in performance as a consequence of stress.

FEASIBILITY OF THE TECHNIQUE

Miller et al. feel that "the structure of the model appears sound" (p. 116, ref. 1). If we consider the model as representative of a general digital simulation technique, then we tend to agree with this conclusion. They point out that "the computer simulation outputs are highly sensitive to assumptions concerning input data both in simulation of relativity

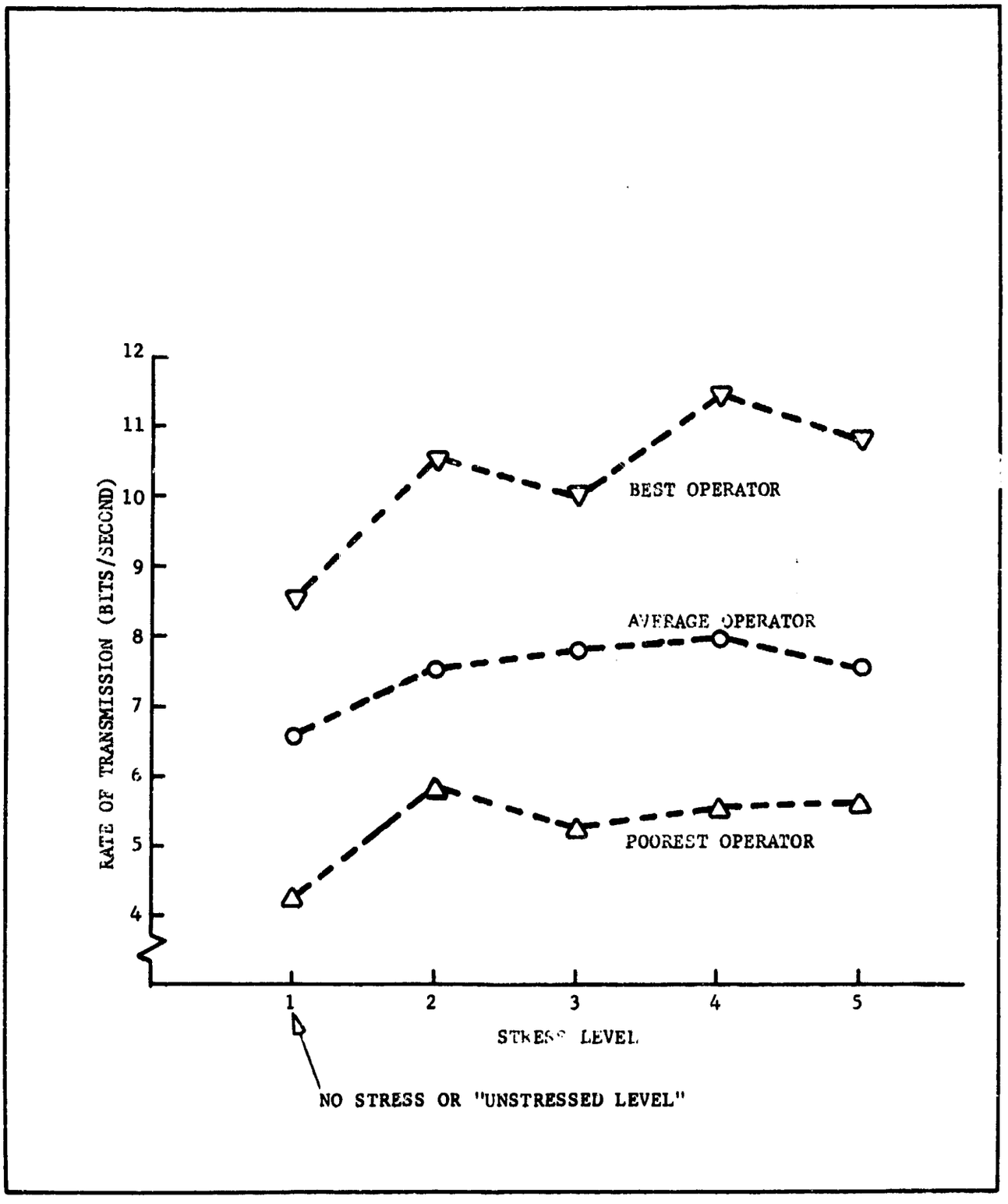


FIGURE 14. APPLICATION OF TRANSMISSION RATE MEASURE
(Taken from Ref. 1)

simple repetive tasks such as the TACDEN operation and relatively complex tasks such as GRAPHDEN operation." (p 116, ref. 1). There is some indication that the sensitivity of the simulation is greater to operator rating (fast, average, slow) than to stress level. 10 computer replications appear to Miller et al. to be adequate to produce their program outputs. This may be because approximately 10 minutes of computer time were required for each simulation of a message entry through TACDEN (238 tasks). When the level of subtask detail was reduced (for the GRAPHDEN study which was similar in terms of its requirements to TACDEN), only 3 minutes of computer time per simulation run was required. They point out that changes in input parameters or program outp can be made within 1-7 days of programmer time plus some computer time to check out the program changes. We mention this point only because this is the first instance found in which some information on time requirements for use of the digital simulation (or any other model) were provided.

The developers of the model present a methodology for integrating the results of the simulation with other factors influencing the overall reliability of the TACDEN. The objective is to obtain a single measure of reliability which will reflect changes in human and machine capabilities.

"The model which is proposed combines two standard examples of waiting line theory The first model gives the steady-state probability that there are n or more messages waiting or being processed when there are (a) machines (and operators) operable. The second model gives the steady-state probability that there are (a) or more machines operable.

"Model

Let n = number of messages which are waiting to be serviced or are being processed
m = number of TACDENS in the system
a = number of machines operable
r = number of repairman
μ = average rate of message processing
λ = average rate of message arrival
α = machine failure rate = $\frac{1}{\text{MTBF}}$
β = machine repair rate = $\frac{1}{\text{MTTR}}$

Assume that message arrivals, message entries and machine failures and repairs are distributed exponentially. Then, for a fixed number of machines (a), the probability P_n that n messages are waiting or are being processed is given by the following relations.

$$P_n = P_o \frac{(\lambda/\mu)^n}{n!} \quad \text{if } n < a$$

$$= P_o \frac{(\lambda/\mu)^n}{a!\, a^{n-1}} \quad \text{if } n \leq a \text{ where}$$

$$\sum_{n=o}^{\infty} P_i = 1$$

"Moreover, if we let s be the number of machines which are not working, it can be shown that in the steady-state condition P_s, the probability that s machines are not working is given by the recursive relations,

$$m \alpha P_o = \beta P_1$$

$$(S+1)\, \beta P_s + 1 = (m-s)\, \alpha P_s \qquad s < r$$

$$r \beta P_s + 1 = (m-s)\, \alpha P_s \qquad s \leq r \text{ , and from}$$

$$\sum_{s=o}^{m} P_s = 1$$

"From these relations we can compile the probabilities P_s that s machines are not operable or $P_a = P_{s'}$ that a = m-s machines are operable." (pp. 140-141, ref. 1).

EVALUATIVE SUMMARY

In view of the fact that the TACIDEN model is only a special case of the Siegel model, the former should probably not be considered as a generally applicable technique. Its value lies in its confirmation of the Siegel model. Hence no formal evaluation is considered necessary.

REFERENCES

1. Miller, G. E., et al. Human Factors Aspects of Reliability, Report U2296, Final Report Contract DA 36-039 SC-90877 for the U. S. Army Electronics Research and Development Laboratories, Ft. Monmouth, New Jersey, January 1964.

2. Meister, D. Methods of Predicting Human Reliability in Man-Machine Systems, Human Factors, 1964, 6(6), 621-646.

III. BOOLEAN PREDICTIVE TECHNIQUE

INTRODUCTION

Another variation on the theme of digital simulation methods pioneered by Siegel is the technique described by Gregg (Reference 1). The major difference between this model and that of Siegel's 1-2 man model is the use of Boolean algebra to model the operator's behavior. "The sequential dependencies between operator responses and machine functions are modeled in Boolean algebra as a set of cause/effect/time relationships" (Ref. 1, p. 44). The model is processed on an IBM 7094 using a set of programs called the Discrete Network Simulator (DNS). In the DNS each event is represented by a Boolean change of state (0 to 1, 1 to 0) as a result of a logical cause and effect relationship among system elements. The program produces a time history of these binary states. The human is modeled "as a set of switching functions or relays" (p. 47, Ref. 1).

GOALS

According to the developer, studies using the methodology should help to determine

(1) The effects of error and equipment malfunctions on task performance.

(2) The effects of stress "and other adverse conditions" on operator performance.

(3) Human error probabilities inherent in system design.

(4) The degree, amount and location within the task sequence of operator loading or under loading.

(5) The comparison of alternative design configurations.

"Modeling of a system in Boolean algebra creates a network analogy of the dependencies and interdependencies among system elements. Each variable is assigned a characteristic activation and deactivation time (i.e., the time required for the effective change of state). The Boolean model is then processed by the simulation program asynchronously in time to present a time history of the binary states of all variables in the system. Varying inputs are used to create different environments,

change activation and deactivation times, insert malfunctions, and so forth. The resulting output time histories are then studied to determine system performance and design adequacy.

"Major characteristics of DNS are as follows:

1. Timing parameters are associated with the variables of the network logic equations.

2. The Boolean equations and the time parameters are converted to machine language and placed in necessary reference tables in one pass through the computer, eliminating the need for processing the model for each simulation desired.

3. Simulation is performed independently of pre-processing or "compilation", with provision for running multiple cases.

4. The time base is the millisecond.

5. A complete behavioral description of the system under study is provided by the printout of an event history, by time priority.

6. Lists of the states of every element in the network can be provided at any specified time during a simulation." (p. 45, Ref. 1).

METHODOLOGICAL SCOPE

Because the methodology models operator behavior as a set of switches or relays, it would seem to be particularly applicable to systems whose operations can be modeled as "go/no go" functions. The highly deterministic nature of the model suggests that it can best deal with highly proceduralized operations but would have difficulty with systems in which decision strategies (cognitive functions) are flexible.

ASSUMPTIONS

The model assumes that two sets of variables are involved in any man-machine system: subtask variables and operator variables. The former consist of the requirements set by the machine design and by the mission, and derived by conventional task analysis, at a level determined

by the needs of the simulation user. Basically subtask variables are the behaviors required of the operator, e.g., meter reading, control responses, etc. Operator variables are such things as degree of learning, motivation, fatigue, skill level, etc. As of the date of the report upon which this discussion is based (1964), operator variables were not included in the model and were assumed to be constant.

The reader will recognize this set of variables as those employed by other models under different terms. For example, operator variables are included by Siegel in his models as an individuality factor.

The Boolean model is completely deterministic. It includes the following assumptions:

"1. Information flow rate, expressed in bits/second, has a concave relationship with error probability. There is an information channel capacity of the human operator, expressed as a point on the information flow rate axis. Underloading channel capacity can be just as detrimental to performance as overloading...

2. Stimulus discriminability or onset is a function of subtask variables. It is primarily dependent upon the hardware characteristics, such as display size, brightness and sensory channel chosen (visual or auditory).

3. The human perceives only one display at a time.

4. Psychological stress is not quantifiable because it cannot be directly measured. Nor can it successfully be expressed as a dependent variable in a mathematical formula whose independent variables are directly measurable. It can, however, be treated as a binary variable, either on or off. Stress is activated when certain conditions or events occur such as the operator falling behind his time schedule in subtask performance, (emphasis that of this author) operator errors which compound the stress condition, or crucial display readings. It is deactivated when such conditions are removed, generally by manual corrective action.

5. Unstressed behavior is characterized by high motivation, facilitation of reactions, and the appearance of psychological set... as an influential agent. When stress is activated, it acts as a

trigger: other behaviors, represented by alternate logical event sequences, interfere with subtask performance. Error probability increases, particularly errors of omission.

6. There is a normal scanning pattern of the displays by the operator. This pattern is cyclic but not normally repetitive. Under stress, the field of perception is narrowed, i.e., the normal scanning pattern is altered. Only those displays which are the stress-producing agents are scanned.

7. Scanning behavior can be conceptualized as a go-no go situation. The operator, acting as a comparator, either senses "go" at each instrument and moves on to the next display or senses "no go" (a difference between an actual reading and the required value), fixes his attention on the critical display and makes a corrective motor response." (p. 46, Ref. 1).

Some of these assumptions will be recognized as having been included in previous models. The inclusion of stress as a primary factor is similar to, and undoubtedly derived from, Siegel's formulation, but is not as conceptually precise. The fact that the assumptions are not presented in mathematical form renders them somewhat suspect. It is difficult to see how, if one cannot quantify stress or express it in a mathematical formula, one can treat it as a binary variable, since common experience does not suggest that the stress response (as opposed to an initiating condition) winks on and off. The concept that stress can be removed by an appropriate corrective action also strikes this author as somewhat dubious.

In contrast to Siegel's formulations, all stress is treated as a negative factor: unstressed behavior is highly motivated, facilitates reactions and creates psychological set. Error probability increases with stress. All of these assumptions are presumably verifiable from experimental research, but not if, as the developer claims, stress is not quantifiable.

Although it is highly probable that error probability is related to information flow rate, it remains to be seen whether the concave relationship postulated actually holds. In any event, the assumption of such a relationship requires that a mathematical equation be provided in order to pick off the particular error probability that corresponds to a given amount of information. (In point of fact, how error probability influences the determination of the operator's Boolean state is not indicated, especially since the model output seems to be solely a time history.) The

developer's model does not suggest this equation. Since error probability also increases with stress, it would seem that some relationship should exist between information flow rate and stress, but no such relationship is indicated.

A major assumption which is only implied by the model description is that error is inconsequential in system performance because the operator will invariably succeed by correcting any erroneous responses he makes. Error is important, but only as a means of causing a chronological mis-match between the mission time requirement and actual event performance. It is, however, possible for a model with this assumption to output a measure of the frequency of successful performance, because degree of success depends solely on whether or not tasks are performed in the required time frame; if they are performed late, the mission fails.

The assumption above is implied by every model which focuses on the time factor and ignores error. Whether one can accept such a sweeping assumption is something the model user must decide for himself. Such an assumption is reasonable only if the system operator has achieved a certain minimal level of proficiency.

PARAMETERS

The model parameters have been discussed in some detail in the section on Assumptions. Particular attention should be given to the stress formulation, although as far as is known, no use has been made of it in testing the model. As the developer states (p. 49, Ref. 1) "The foundations of the theory are stimulus-response, a normal scanning patter of displays, the use of stress as a trigger of alternate behavior chains..."

The developer points out that each subtask element (e.g., monitor attitude display, activate retrothrust switch) is a single stimulus-response unit. There is a determinable start and completion for each such element. The operator's normal scanning pattern can be interrupted by priorities during critical situations such as stress. Apparently there is a priority interrupt feature in the program. Alternative behavior chains are provided "to predict behavior due to stress" (p. 47, Ref. 1), presumably these reflect decision-making sequences which are triggered by a stress condition.

As far as the essential parameters of this model are concerned, the elaboration of the model (at least as presented in reference 1) is insufficiently detailed to indicate how the parameters are used to produce its outputs.

DATA

1. Input Data

The primary input data requirement is for response times for the various subtasks. There is apparently no need for error probability data, on the assumption that error is irrelevant except as a delaying function, humans correcting any erroneous actions inevitably. This ties in with the model's primary interest in the time history of the system (an orientation similar to HOS and ORACLE). The effect of an error is simply to delay performance of the correct response; system failure occurs if the error effect is such as to prevent the last milestone event from occurring at its required point in time.

The fact that the model is very deterministic enforces a requirement for very precise input data (stochastic models can take advantage of the "error slop" involved in input data distributions; deterministic models have much less margin for error).

2. Data Sources

Presumably response time data could be secured from any source, such as the AIR Data Store. However, the developer implies the necessity for gathering these times from a physical simulation apparatus (in his case something called TRANSAT), which resembles the system being modeled. This might increase the accuracy of the model, but it is unlikely that every potential model user would possess a simulator to provide him with the required data. The question might be asked, moreover, why one would need a model at all if one had to have a simulator and exercise it before any significant outputs could be derived.

If one accepted the implied requirement (for simulation data as inputs to the model), it would severely limit the utility of the model. It is our feeling, however, that this requirement was primarily for purposes of testing the model in its formative state and is therefore not a firm requirement for its use.

3. Output Data

The output of the model is a prediction of the times when certain required actions would be performed as related to the mission scenario. As indicated above, error occurs when the time of actual performance of an event does not match the time when it is supposed to occur.

It is suggested, moreover, that a distinct performance measure is not output and that "the simulation is interpretive rather than computational" (p. 46, Ref. 1). Presumably an interpretation must be made of the time histories output by the model. This last feature bears some similarity to the use made of the outputs in Wherry's HOS.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

As with other models a comprehensive task analysis is required as a preliminary to application of the model. Analysis is down to the task level, e.g., "receive/transmit parameters", but there is an implication from the assumptions dealing with information flow rate, scanning patterns, etc. that a more detailed analytic level is required. However, as is usual, the developer provides no details about how the task analysis is to be conducted.

Essential elements of the task analysis are flow charts and circuit diagrams of the subtask sequences. Subtask sequences are coded on the circuit diagram in accordance with the following Boolean operators:

* = and

\+ = inclusive or

/ = not (complementation or negation)

A given subtask sequence would then be written in terms of code names for the subtask elements plus the necessary operators. For example, subtask 5 (monitor velocity display in pre-retrofire) would be represented as MONVEL which is equivalent to MONATT (monitor attitude display) * / VOCRET (receive retrofire command). In other words, monitoring the velocity display involves also monitoring attitude before receiving the retrofire command.

ANTICIPATED MODEL USES

In terms of the potential model uses we have postulated (prediction of system effectiveness, design analysis, manpower selection and training), this model is primarily useful for design analysis. Information received from one of the people who helped apply the model is that the developer's major goal was "to determine where the delays will occur due to human engineering shortcomings" (Ref. 2).

The model could be used for predicting system effectiveness, but only with great difficulty, because despite all the rodomontade provided by its developer, it is not really designed as a predictive tool. Presumably, if one had two different design configurations, and compared performance of the model with each configuration, a choice between the two could be made, but the adequacy of this comparison would depend on how sensitive the model inputs were to design differences, and no information on this point is provided.

In fact, because the task descriptions required for the model do not seem to reflect any design parameters, there is some question in this author's mind about how the design analysis could be performed. The model is not designed for nor could it be very usefully applied to derive selection and training outputs.

VALIDATION/APPLICATION STUDIES

Information on this subject is not provided by the developer. Reference 2 indicates that the model was exercised, simulating some data gathered on the TRANSAT re-entry simulator; and that the model outputs agreed quite well with actual performance data. However, as reference 2 points out, this "should be no surprise" because "many of our parameters were established in TRANSAT for the model (time constants)..."

The author cannot feel therefore that this represents an adequate validation of the model. Apparently also the model has not been applied to any system development project.

It is difficult on the basis of the very inadequate information provided to make an adequate evaluation of this model. There is some doubt whether the methodology can be considered a general model ("theory of operator performance", p. 46, Ref. 1) as its developer apparently intended it to be.

EVALUATIVE SUMMARY

Validity - Very inadequate data available.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Limited to proceduralized tasks.

B. Applicability: Model is not particularly predictive; it is essentially descriptive.

C. Timing: Model can be applied only to systems in later stages of development.

Model Characteristics

A. Objectivity: Relatively high.

B. Structure: Assumptions and parameters not well defined.

REFERENCES

1. Gregg, L. W. A Digital Computer Technique for Operator Performance Studies. Proceedings, Fifth National Symposium on Human Factors in Electronics, May 5-6, 1964, San Diego, California, sponsored by the Professional Technical Group Human Factors in Electronics of the IEEE (pp. 44-51).

2. Freitag, M. (Personal communication to the author, 20 August 1971).

IV. THE HUMAN OPERATOR SIMULATOR

INTRODUCTION

We have encountered other models, e.g., Siegel's, Miller's, which have made use of computers to simulate behavioral processes, but Wherry's Human Operator Simulator (HOS) has certain characteristics which differentiate it from the others.

For example, Wherry makes the point that "HOS is not a man-machine model per se, but rather a general purpose man model" (Ref. 3). By this we infer that Wherry makes a distinction between the simulated human operator (the model itself) and any specific system with which that operator must interact. "For HOS to be used for a specific system, (goal oriented) procedures which the operator is to use with that system must be prepared through the use of the Human Operator Procedures (HOPROC) language and encoded for HOS through the HOPROC Assembler and Leader (HAL) program. Further, the displays' and controls' location must be input to HOS. Finally, there must be a simulation of the system hardware, the environment, and any targets of interest since they will both directly and indirectly determine what occurs during a simulation run" (Ref. 3). This writer is not sure that the distinction, although interesting and technically correct, is critically important to the use of the model.

HOS is not a stochastic model, as are the other computer-driven models we have reviewed. It does not sample distributions of performance data. Rather, it relies on equations describing relationships between parameters and performance outputs. The equations are of course based on experimental data, but the use of functional relationships in equation form rather than sampling from distributions tends to reduce the need for data banks, "since describing the situation properly (i.e., the procedures, tasks, station layouts, etc.) allows HOS to output whatever data is desired" (Ref. 3).

HOS is extremely molecular, much more so than the other models considered, and operates at a very fine level of task element detail. For example, it makes use of inputs such as the state of displays and controls (e.g., 5000 feet altitude), their locations within the cockpit, reach distance and time to reach controls, etc.

Oddly enough, for such a molecular model it is much more behaviroally oriented than the other models considered. We say this because great emphasis is placed on cognitive processes, particularly recall, memory decay, strength of recall, etc. which are basic elements in the decision-making process. Other models include the decision-making process, but deal with it as a total entity rather than in terms of the elements making up the

decision. One might assume from the emphasis on decision-making that the model was not deterministic. However, the particular decisions made depend on deterministic functions, for example, the parameter called "hab strength".

One other point should be noted. The model in its present form is specialized for pilot behavior, although the developer, Cdr. R. J. Wherry, Jr. feels that it can be adapted to other kinds of systems.[1]

All things considered, therefore, one cannot think of HOS as being simply another variation of the general digital simulation model.

GOALS

> "The purpose of HOS is to be able to arrive at accurate data about how specified trained operators will perform specified functions in specified station configurations prior to any 'metal bending' (i.e., early in the development cycle), be it for dynamic simulation or prototype development....." (Ref. 3).

Wherry makes much of a distinction between the simulation per se and any uses to which that simulation can be put. He sees HOS as an effort to substitute computer control for the performance of experimental studies. "The operator model is to be sufficiently excellent so that its output may be analyzed and used in exactly the same manner as that obtained from an experiment with a real human operator. Indeed, it is likely that HOS will provide certain types of data that even the most elaborate human experiments cannot" (p. 1-1, Ref. 2). For its developer, then, HOS is not merely or primarily a methodology for evaluating man-machine system effectiveness.

The developer emphasizes these points:

> "a. The problem of getting adequate simulation of a human operator has little or nothing to do with the problem of analyzing human operator data; b. the problem of being able to relate human operator data to system-effectiveness criteria is another

(1) Wherry notes, "HOS is not a model of a pilot, but a general model for any seated operator whose primary task is to observe displays, compute internal mediated functions, make decisions, and manipulate controls. As such, HOS can be used (i.e., "tailored" to be, through writing appropriate procedures and defining appropriate displays and controls) as a pilot, a sensor operator, a tank driver, a truck driver, a command and control operator, etc." Present plans are to expand HOS to include a capability for operator mobility, transport of objects and object use.

> and separate problem; c. the ability to arrive at suggested modifications to station layouts given that one can relate operator performance to system performance is still another problem; d. let's not get these problems confused; they are separate and distinct ones, but accurate human operator data is critical to this iterative cycle; .....f. the purpose of HOS is to be able to arrive at accurate data....." (Ref. 3)

Although the distinction between the simulation and its uses is well taken, for our purpose, which is to examine predictive models for their use in answering system development questions, the distinction is not important. Implicit in the development of any model are the uses to which that model is to be put, whether or not these uses are explicitly stated. The "pure" scientist need be concerned only about "gathering data", but data gathering without at least an implied use of those data seems somewhat pointless.

Although the reference documents for this model are somewhat reticent about goals for HOS, we can infer these goals from the very distinctions the developer makes. For example, he cites the problem of relating operator data to system-effectiveness criteria, from which we can infer that one of the implied model goals is to estimate the effectiveness of a system. In referring to Human Operator Simulator programs, Wherry says "these programs,would provide valid data on how well trained and motivated operators will be able to perform their tasks in a given system during a given mission" (p. 2-5, Ref. 1). The ability of a model to suggest redesign possibilities is indicated by the earlier reference to station layouts. The application of the model to training requirements is indicated by the statement

> "The outputs of Human Operator Simulator programs could be used to identify periods of maximal load, frequency of usage, etc. and thus determine what procedures and mission segments on which to concentrate during the training period" (pp. 2-6, 2-7, Ref. 1).

Wherry also suggests the possibility that two complete simulations could act as adversaries for each other. "This real pilots and other crew members could be trained against targets which "intelligently" attempt to attack....." (p. 2-7, Ref. 1).

It is apparent therefore that HOS, like the other models described previously, has goals it attempts to achieve, and these goals are very similar to those of the other models reviewed: estimation of system effectiveness, the comparison of alternative design configurations and the suggestion of redesign possibilities and suggestions for training requirements. In a later section we will discuss how well these goals can be achieved.

DEFINITIONS AND ASSUMPTIONS

> "The HOS model considers an operator to be an intermediary function between a mission described as a series of tasks and the objectives of that mission.... The operator has two types of interfaces with the system: controls and displays.... The operator utilizes a tool, the hardware system, under the direction of a set of procedures....." (pp. 1-5, 1-6, Ref. 2).

This concept is essentially the same as that which underlies all other man-machine models.

In consequence HOS consists of four major components:

(1) the simulated human operator;

(2) the procedures governing system operation;

(3) the physical layout of the operator's workspace;

(4) the mission to be run.

Special attention should be paid to item (3) immediately above. The inclusion of the physical layout of the workspace as one of the main components of HOS means that, if that layout is properly taken into account, HOS will be highly sensitive to the design configuration of the system. It will be recalled that one of the comments made by this author relative to the other models is that they seemed to be insufficiently sensitive to design. Apparently that objection cannot be made about HOS.

The other major components have their analogues in the other models, in particular the procedures to be performed and the tasks (mission) to be completed.

Another common sense assumption is that the operator knows what is to be done and how to operate the system. Perhaps for this reason the model does not deal with success probability. Given that an operator recalls an instruction, he will perform it correctly; the only element of uncertainty is how long it will take him to recall the proper procedure. The point is that there appears in this model to be no possibility that the operator will make an error. In view of the fact that pilots do make errors, the failure to consider error in this model -if this is indeed the case - represents a serious deficiency.[2]

---

(2) See Appendix I to this section on HOS and Human Error.

Another important assumption is that "it is a premise of the model that all operator actions are composed of relatively few basic activities" (p. 2-4, Ref. 2). For example, if an action is to "manipulate throttle", we might find that the "manipulate" instruction itself breaks down to "reach" and "twist" actions. The consequences of this assumption is the micro-level of detail noted previously; the simulator must take more molar behaviors and decompose them into their molecular constituents. To do this the HOPROC language contains a set of statements which define the molar instruction (task) as more molecular (subtask) behaviors.

In order to provide a sufficient amount of flexibility in the model, instructions or tasks--- e.g., alter desired position of throttle to 50% --- are defined as goals for the operator rather than as rigid actions for him to take. In the throttle example given previously, the concept of the task as a goal "causes the operator to adjust the throttle setting if the throttle is not at 50 percent, but to do nothing if the throttle is already at 50 percent and he knows it" (p. 2-5, Ref. 2). The concept of the task as a goal to be realized if, as and when circunstances require it provides a certain amount of flexibility to HOS which other models may not have, and is therefore a very desirable feature.

> "This approach to instruction statements enables any operator activity to become, in effect, a subroutine, just as in a computer program, to be accessed when "desired" by the simulated operator. These subroutines are analogous to operating procedures a man would learn in a training program. For example, on a certain type of aircraft a certain fixed procedure is used for the pilot to change heading. A coordinated turn requires actions to the control stick and rudder pedals plus eye movement to monitor the compass, the turn-and-bank indicator and perhaps, the outside world through the window. The same general set of actions is employed each time such a turn is necessary." (p. 2-5, Ref. 2)

More specific assumptions and the structure of the model will be described in the section on MODEL PARAMETERS.

METHODOLOGICAL SCOPE

HOS in its present formulation is strictly pilot-task-oriented. This does not mean, however, that the logic of the model cannot be applied to other types of systems and tasks, although a very substantial rewriting of program specifics would probably be required. "The Human Operator Simulator program should be general enough so that it could be used to simulate any kind of operator". (p. 2-7, Ref. 1).

In its present application, therefore, we must consider the model as restricted to the simulation of pilot behaviors. This is a serious restriction in comparison with models that can be applied to a variety of systems. Wherry may well challenge this point on the grounds that HOS is independent of particular system inputs that are applied to it through HOPROC. However, the HOPROC language is so specialized to flight missions that it would, as indicated previously, require a major revision to fit it for other types of systems, e.g., command/control.

The model has no apparent difficulties with continuous-type monitoring and tracking tasks. This is because the simulated mission is considered "a series of routine (discrete) activities punctuated by exogeneous events, such as the appearance of a target, and certain major decision points that may be specified in time, such as "Return to Base". "The simulated operator may then be.....conceptualized as most of the time performing a general monitoring function interspersed with specific mission-directing and corrective actions" (p. 1-8, Ref. 2). The point is that, having once been given a task, the pilot sets his instruments for the specified parameters and merely corrects any deviations that occur. In other words, continuous activities are formulated in a discrete manner, as points at which the operator is required to take a discrete action because a deviation from the requirement has appeared. This considerably simplifies the model because it is not required to simulate continuing actions.

The fact that, as we shall see, the model output is primarily time means also that no continuing record of errors (deviations in the continuing activity) need be kept.

MODEL PARAMETERS

It is impossible to describe these without describing the structure of the HOS model as a whole. In order to understand the logic of the model it is necessary to start with the inputs to HOS. These require a specific language (Human Operator Procedures or HOPROC) to enable users of HOS to input both operating procedures for the system and mission instructions in English statements. The output of HOPROC is then input to HOS.

> "HOS has been written in FORTRAN IV. HAL, the HOPROC Assembler/Loader program has been written in COBOL for syntactical convenience, and is known as HOPROC-1. HOPROC will shortly be rewritten in FORTRAN, incorporating several types of improvements, and will be known as HOPROC-2." (p. 1-11, Ref. 2).

"The Human Operator Simulator requires many different types of inputs, some of which are most easily supplied by means of an English-like language and some of which require only control cards as a vehicle.... HOPROC statements are translated into HOS-compatible code by means of..... HAL. HAL is the program that translates input into a form usable by HOS. HAL reads in a card deck of HOPROC statements, checks the statements for validity, and outputs encoded HOPROC statements, a set of dictionaries and FORTRAN mediated function statements....." (p. 2-1, Ref. 2)

Five types of directions are provided by HOPROC:

"(1) Operating procedures for the system being run by the HOS operator, for example, the procedure to use in establishing communications contact with his flight commander. Operating procedures are analogous to the ability provided by operator training on the system.
(2) Mission sequence---the scenario to be run on a particular HOS experiment.
(3) Communications received.
(4) Equations he uses for mental computations (mediated functions) whenever necessary.
(5) Equipment terminology---the titles by which displays, controls and o-states are referred.

Items (1), (2) and (3) make use of....English....Items (4) and (5) are input as simple lists." (p. 2-3, Ref. 2).

HOPROC presently supplies nine statements or instructions which are shown in Table 11 (Ref. 2).

Certain terms in Table 11 require further explanation. Display and control are relatively obvious. However, the term "function" refers to mediated function (to be discussed in more detail below) which is a human thought process or mental computation. "The result of a mediated function is a value or setting (of a control or display). For example, a HOPROC statement might be to go to a certain procedure if the value of (resulting from) a mediated function were outside a given tolerance" (p. 2-6, Ref. 2).

The term "o-state" refers to any of several descriptors of human capability. This parameter introduces a very necessary flexibility into HOS. Note the similarity of the o-state to the F parameter of Siegel and Miller. By specifying a particular o-state, "the human factors analyst can specify for each run the operator capability profile he anticipates for the system being developed, as results perhaps from expected type of personnel or level of training to be provided" (p. 1-5, Ref. 2). Unfortunately the documents describing HOS do not indicate clearly what the various o-states may be or how they are derived. Fatigue is specifically listed, and there is a reference to one o-state being a short term memory value. O-states

TABLE 11. HOPROC-1 STATEMENTS

STATEMENT NAME	STATEMENT	USE
Define	DEFINE PROCEDURE procedure-name	A non-action statement to define the name of a procedure. Example: DEFINE PROCEDURE FLAREDROP.
	DEFINE PROCEDURE procedure-name TO ENABLE (display or control). (or) TO ADJUST (display control or function).	Two non-action statements that define procedures to either turn on an on/off switched display or control, or to select a setting on a multi-position switch, control, or level of a mediated function. Examples: DEFINE PROCEDURE RADARON TO ENABLE RADAR, or DEFINE PROCEDURE POWER-SET TO ADJUST THROTTLE
Activate	ACTIVATE PROCEDURE procedure-name.	Allows program control to select the procedure named. Example: ACTIVATE PROCEDURE FLAREDROP or ACTIVATE PROCEDURE RADIO BASE.
End	END PROCEDURE procedure-name.	Used as the normal termination of a procedure. Disallows program control to select the procedure named as one on which to work.
Go To	GO TO STEP label.	Transfers program control to the step of that label. Example: GO TO STEP 640.
	GO TO STEP label NOW.	Transfers program control to the labeled step immediately. (See Note 1) Example: GO TO STEP 640 NOW.

TABLE 11. HOPROC-1 STATEMENTS
(Continued)

STATEMENT NAME	STATEMENT	USE
	GO TO STEP label IF (display control, function, clock or o-state) IS (not) equal to, (not) less than, (not) greater than, (not) within parameter, setting or number). (NOTE: If WITHIN is used followed by a number, a second number must follow as the other limit.)	Conditional transfer of program control to labeled step. Example: GO TO STEP 640 IF FUELQUAN IS LESS THAN LOWER LIMIT .
Commence	COMMENCE MONITORING (display or function)	Causes operator to periodically estimate (recall, observe or compute) the value or setting of the display or mediated function to see if it is within specified range. Periodicity depends on criticality of that display or function, which is computed by HOS. Example: COMMENCE MONITORING CLIMBRATE.
Cease	CEASE MONITORING (display or function)	Exclude the specified display or function from further periodic monitoring. Example CEASE MONITORING CLIMBRATE.
Compute	COMPUTE FUNCTION function-name.	Cause HOS to obtain an estimate of the value of a mediated function, a simulated thought process. Example: COMPUTE FUNCTION ARRIVETIME.
Alter	ALTER parameter OF (display, control, or function) TO number	Change the parameter of the specified display, control or mediated function to the specified value or setting (see Note 2).
	ALTER parameter OF ... AND ALTER parameter OF ...	Links more than one alter instruction in one statement.

TABLE 11. HOPROC-1 STATEMENTS
(Continued)

STATEMENT NAME	STATEMENT	USE
	ALTER (VERY RAPIDLY, SLOWLY) parameter OF ...	Enables user to specify rapidity of the desired change of parameter value. On a scale from 1 to 9, very slowly would be 1, slowly 3, normally 5, rapidly 7, and very rapidly 9.
Set	SET CLOCK clock-number.	Resets to zero and starts a specified simulted mental time clock (clock number may be 1 through 6 for any procedure.

Note 1. In the normal sequence of accomplishing HOPROC statements, a GO TO is equivalent to GO TO NOW. However, if the GO TO transfers control to a preceding step, HOS assumes that this is a waiting loop and will instead search for other current tasks to process. The GO TO NOW statement prevents this searching and forces control immediately to that specified step.

Note 2. "Parameter" represents any of the following seven entries:

CRITICALITY
DESIRED POSITION
LIMIT (LIMITS)
LOWER LIMIT
UPPER LIMIT
PERCENT LIMIT
ABSOLUTE LIMIT

Thus, an example of an ALTER statement would be ALTER PERCENT LIMITS OF CLIMBRATE TO 20. Another example would be ALTER RAPIDLY DESIRED POSITION OF COMPASS TO 150 AND ALTER VERY RAPIDLY DESIRED POSITION OF ALTIMETER TO 8000, which if an aircraft were flying on course 240° at 15,000 feet would cause HOS to initiate procedures necessary to execute a steeply-diving, tight 90° left turn.

apparently enter into mediated functions. For example, the time to perform a mediated function may increase with fatigue, and fatigue will increase over the time of a mission. HOS has the capability of altering o-states as a mission progresses.

> "A mediated function is defined as desired information, not directly displayable to the operator, which he may estimate by his knowledge from several displays, controls or other mediated functions. An example of a mediated function might be "slant range to target" which could be calculated if the pilot knew his altitude and dive angle and was lined up with the target. In actual use, each mediated function used in the program requires that an equation be provided when the mediated function is defined for use in the procedures." (p. 2-15, Ref. 3).

> "Being calculations, mediated functions are expressed as equations in HOS, and are entered as input via one section of HOPROC.... Any given mediated function may require as one of its component variables another mediated function. For example, suppose the desired mediated function was "flight hours remaining" to be calculated on the basis of fuel quantity remaining and fuel flow rate. Fuel remaining is readable directly from a gauge, but flow rate may have to be computed by the operator. HOS will actuate a data gathering sequence that will cause the operator to either recall or observe the fuel quantity reading and perform another mediated function for the flow rate. Results of the flow rate computation and the fuel quantity value are then used to compute "flight hours remaining." (p. 2-7, 2-11, Ref. 2).

The specific equations mediating functions will be described in connection with the ESTIMATOR module of HOS.

A third type of HOS input that uses HOPROC is the titles (names) of displays, controls, settings, and o-states. Each is assigned its own section in the HOPROC input deck. The purpose of these titles is to enable their use as alphabetic variables names in HOPROC statements.

The concept of an internal clock, which keeps track of elapsed time since the clock was "set" was also found necessary for HOS. The internal clock refers to and measures the human's ability to estimate time, for example, if 15 seconds have elapsed since the pilot attempted to start his engines.

Controls, displays and functions

> "may be used in conjunction with various qualifiers. We may refer to the desired position of any source, the criticality of a display or function. The upper limit of, and the lower limit of any display or function may be referred to or we may refer to both upper and lower limits simultaneously by speaking of the limits of a display or a control.... An unqualified source is assumed to refer to the estimated position of that source." (p. 2-17, Ref. 1)

We pass now to HOS, which operates as a processor of instruction statements produced by HAL.

"Its logical processes are closely analogous to the logical processes of a human operator performing the same tasks, and the data it produces is amenable to the same type of analyses as that produced during a human experiment.....

HOS has four principal modules: the DECODER, the MULTIPLEXOR, the ESTIMATOR and the BANKER.... Inputs to HOS are from two sources, the HOPROC compiler via HAL, and a directly-entered data file representing initial states of certain elements and the location of displays and controls....HAL inputs are of two types. The first is the titles dictionary and the encoded procedure string. The following information is provided:

(1) Alphabetic name of the device, function or procedure.
(2) Number and name of settings on the device or function.
(3) Kind of device.
(4) Encoded procedure string.
(5) Pointers to the start of each section.

"The second type of HAL input is a deck of mediated functions in FORTRAN. This deck is combined with special subroutine header cards, compiled and then link-edited with HOS to form the complete simulator.

"Direct inputs (to HOS) include:

(1) Initial states of displays, controls, mediated functions and o-states.

(2) Locations of displays and controls (x, y and z coordinates).

(3) The current values including the actual, the desired, and upper and lower boundaries of devices and functions.

(4) The last estimated location (x, y and z), time of the estimate and hab strength of controls.

(5) The previously estimated value or setting, time of estimate, and hab strength for displays, functions and controls.

"The DECODER is analogous to the human understanding and decision-making function. It recalls or receives an instruction, comprehends it, and decides what further procedures are necessary to accomplish it. The DECODER is the logical branch point that shunts an instruction to the proper program routine for handling.

"The MULTIPLEXOR is the module that decides which procedure should be handled next by the DECODER. It establishes priorities among possible instructions and can generate interrupts when the results of one task logically necessitate an alteration in the instruction sequence.... Whenever a sequence of instructions.... is completed, (it) determines the next procedure to initiate..... Whenever..... a program is "waiting" for the completion of some other event before it can continue.... the MULTIPLEXOR scans for other available tasks to occupy the operator.... The MULTIPLEXOR is the module that initiates actions on the basis of interrupts. Several types of interrupts are recognized. The major type originates when the operator must absorb information from a display which is currently not in an active mode.... The current procedure is interrupted and the procedure to enable that display is activated and work is commenced on that procedure.... A second type (of interrupt) encompasses hardware-originated interrupts that may affect the sequence of a run, such as the appearance of a warning light.... A third type of interrupt occurs when, for example, the simulated operator requires his left arm to activate some device but he is currently using his left arm to adjust another control. The current procedure may be interrupted and a scheduled interrupt arranged for when the left arm becomes available.... In the interim the operator may work on other tasks...

"The ESTIMATOR is the sensing-remembering core of HOS. It also provides the time costs both for performing actions and for absorbing information... The ESTIMATOR utilizes in its computations numerical values provided by... the hardware interface. For instance, following a decision to set the knob to 50 percent, the ESTIMATOR might determine that it will take 0.9 seconds to absorb the information on the present position of the knob and that a knob-turn action will take 1.6 seconds. It will also confirm that the knob's final position will be at 50 percent.

"The ESTIMATOR's function closely simulates a human being's function in the same situation, in that its operation is not fully deterministic. There are three kinds of actions: (1) preparation for information gathering, such as moving eyes to a display, (2) information absorption, such as recognizing the value indicated on a display, and (3) control actuation, such as adjusting a lever... Information gathering is a four-stage process. First, if the operator is already in contact with the display or control... he will absorb its position. If not, recall is attempted, which if successful entails only a small time cost.. Second, if recall is not immediately possible, given more time it may be. This process, known as "almost-recall", involves additional small time costs while the operator continually interates the the recall attempt.

"The success of any recall attempt is probabilistic, depending on (1) the time since the revious observation (t);;(2) how well the operator knows the value referred to as hab strength (H); (3) short term memory capability as given in an o-state (O_{stm}); and (4) a confidence level. The confidence level, expressed as another o-state (O_c), reflects the operator's willingness to believe his own memory, and is the reference against which the probability of recall is tested for success. Hab strength... decreases over time, therefore the probability of recall (P_r) is expressed as

$$P_r = H^{(O_{stm}\sqrt{t})}$$

"If almost-recall fails, the third and fourth stages are entered in which the operator moves his head, eyes, hands, etc. to the display or control and absorbs the information. The time cost in this case includes both the physical movement (computed by means of empirically-derived human performance equations) and the information absorption time. Information absorption time is computed as the sum of separate absorption attempts, the number required depending primarily on hab strength. Each time the operator looks at a display he (1) incurs a certain time charge and (2) a new hab strength is computed according to the equation

$$H = 0.1 + 0.9\,SH_p \quad \text{where}$$

- H = Hab strength
- S = Similarity between the preceding value of the reading or seeing and the present value; $S = \left|1 - (E-E_p)/E\right|$
- H = Previous H value
- E^p = Present estimate of value of reading or setting.
- E_p = Previous estimate of value of reading or setting.

The previous material was extracted from pages 3-5 through 3-11 of Reference 2.

Other computational functions performed by ESTIMATOR deal with distance. The inputs for the distance computation are the coordinates in x, y and z of the two sets of points. The origin of the coordinate system is placed at the cockpit position corresponding to the midline between the pilot's eyes. The equation used is:

$$D = (X_1-X_2)^2 + (Y_1-Y_2)^2 + (Z_1-Z_2)^2$$

This outputs distance in inches.

Another function calculates eye movement and fixation. The inputs are the angle of eye shift and change in convergence angle. Output is the time in seconds for the eye movement. The basic equation is

$$T_{(sec)} = .0175 + .0025A \quad \text{where}$$

A = angle of movement in degrees.

Arm reach time is also calculated in terms of the equation

$$R = 1.8D + 8.0 \quad \text{where}$$

R = reach rate in inches/sec
D = distance between points in inches.

Another function calculates the extrapolated estimate of a value up to the current time. Inputs to this function include last estimate of the value, elapsed time since last estimate, previous estimate of the value, elapsed time since previous estimate, and probability of recall of the previous estimate. Another function computes the modularly decayed value of the estimate.

Time to complete a manipulation (turn) of a knob is based on degrees per unit of change in the knob and the hand force that is used. The formula is

$$T_s = .0482 + .00496F + .00144D + .000147FD \quad \text{where}$$

T_s = turning time in second
F = pounds
D = degrees.

The BANKER module accumulates time cost of the actions performed by the operator and also serves as the HOS interface with the hardware simulator. Obviously time is expended in the decision-making process, when activating a procedure, in using short term memory, making eye, hand, etc. movements, and absorbing information.

A number of points should be made about the model parameters. It is apparent that HOS is based on a number of formalized assumptions about operators. In discussing various approaches to the simulation of the human operator, Wherry indicates that "A more profitable approach would be to develop sophisticated models which require the formalization of assumptions about models" (p. 2-27, Ref. 1). And it is true that, if one could specify a sufficient number of parametric relationships, a comprehensive model of operator behavior could be established without the need for stochastic sampling, the use of data banks, etc. However, these assumptions, as reflected in the various mediating functions, for example, depend on the validity of the relationships specified between various parameters. Wherry feels that considerable data already exist to permit the establishment of these parametric relationships. There is, however, some obscurity about how a number of the parameters are derived, in particular the o-states, hab strength, short term memory capability, memory decay, etc. This makes it difficult to analyze the empirical basis of the more important mediating functions. In any event, the adequacy of these formulations can only be determined by the performance of validation studies (to be discussed below). Despite the molecular character of this model, a number of the parameters, such as hab strength, short term memory, information absorption, etc. are extremely complex (at least in concept). It is impossible with such parameters to estimate their validity merely by inspection; they must be tested experimentally, and until they are verified, they must be viewed with some suspicion.

One factor that redeems the apparent harshness of the previous paragraph is that the model is concerned primarily- perhaps solely, it is not quite clear- with time as an output. This variable is much more tractable than the estimation of success probability. Time can be estimated much more readily than can success. Consequently the model parameters associated with time should also be estimated more easily than would be the case otherwise.

DATA

1. Data Inputs

These have been indicated in the previous section of this description, but for convenience are listed below:

From HAL

(1) Encoded procedure statements (the steps required to perform the flight)
(2) Dictionary of titles for displays, controls and settings
(3) Mediated function statements (equations)

Direct Inputs

(1) Initial states of controls, displays, mediated functions and o-states
(2) Physical location of controls and displays
(3) Information absorption times (e. g. , recall ability)
(4) Output specifications

Of these the most important are the mediated function statements, since these are the only inputs about which any question of validity arises. Two possible sources of error in the function statement may exist; (a) the data on which the equation is based may be erroneous in part; (b) the relationship between parameters expressed in the equation may be formulated erroneously.

2. Data Sources

Data sources for mediated functions and information absorption times are largely the experimental literature. Because of the molecularity of the relationships involved, the data required can only be secured from a highly controlled (i. e. , laboratory) situation, which suggests that prototype engineering or operational test sources - or that favorite of last resort, the subjective estimate - would be inadequate to supply the necessary data. One could, of course, set up one's own experimental situation to derive a function, but in fact it is most unlikely that the data would come from the individual experiment. The problem with the experimental literature, however, is the great disparity in experimental conditions under which parameters are investigated. However, assuming that one once had sufficient data in which one had confidence, the necessity for further recourse to external data sources would be eliminated.

Procedure statements (tasks to be performed by the simulated operator), titles, locations and initial states of controls and displays could be derived without excessive difficulty from an analysis of the physical system to which HOS is to be applied. Given a system whose characteristics are fixed, no question of the validity of these data can be raised.

In view of the extreme minuteness and precision of the system data required, this writer questions whether it is possible to derive these data very early in system development, as Wherry postulates. Will designers be able to provide such detailed information early in development, and, having once been provided, will those data remain unchanged? This is a practical, not a technical question, and can be answered only by applying the model to an actual system development problem.

3. Data Outputs

HOS has the capability to output data in the form of a detailed time-history log of what the simulated operator was doing. This includes such items as:

a. time action started
b time action completed
c. anatomy involved, e.g., eyes, ears, left arm/hand, right foot, voice, mental process
d. display, control, function or procedure involved in the action.

Again it must be emphasized that HOS outputs only time and status information. It does not provide any indication of error or the effect of error on mission performance. At the risk of over emphasizing what the developer may feel is a minor point, it must be said again that this writer feels this is a serious deficiency in HOS. By restricting its output to a time history, the model can be used only in those situations in which the user considers error to be irrelevant. In effect, the measure of system effectiveness becomes the amount of time consumed in flying the mission. Certainly time is a crucial dimension of system effectiveness, but can one be satisfied with time alone?

Lest we be thought unfair, it would be advisable to quote from reference 3.

> "These output tapes can then be used in a variety of data analysis programs (time-line analysis, link analysis, etc.) which will selectively use these data. Additionally there would be data outputs anticipated from the hardware and target simulation programs which would give "state" values of those systems in a time-history log format... These data can be collated and/or corelated with the HOS derived outputs and various system-effectiveness criteria can be imposed to determine the "goodness" of the system... (including the functions assigned to operators, the procedures he is to use, the layouts of the displays and controls, etc.). When sufficient data have been collected to determine where the problem areas

are, ... "optimization" programs can be run to suggest better layouts, procedures can be re-written to shorten time used by the operator for non-critical functions, etc. and these modifications can be tried out using HOS."

For a description of the outputs provided by the human operator data analyser/collate (HODAC) see Appendix II to this section.

PROCEDURES FOR APPLYING MODEL

1. Analytical Methods

Standard task-equipment analysis (TEA) will supply the necessary procedural statements, initial states, etc. Such a TEA is necessarily very detailed, the time increments for updating data being very short (usually in the neighborhood of several seconds). In this connection it would be useful to get some indication of the amount of work in man hours required to make such an analysis, since the usual TEA performed during system development is not that detailed. (This request for development information applies not only to the present model but to all the others reviewed.)

2. Methods for Synthesizing Data

The original intent of this subsection title was to describe how the model combined lower level behavioral units to form more molar behavioral units, e.g., how it combined subtask data to secure task data, task data to secure function data, etc. This is required by static models of the AIR, THERP and TEPPS type. Since simulation models reproduce all required behaviors, combination in the original sense of the term does not apply here. In any event, since the data output is time only, and that is combined additively, the problem, if it existed at all for HOS, would be solved very readily.

ANTICIPATED USES OF THE MODEL

1. Prediction of System Effectiveness

Under this heading we include the following:

a. Determination that the system will perform to specified requirements. In other words, given a system requirement, e.g., the

mission must be performed in 30 minutes, can the system (including the human operator, of course) perform the mission in 30 minutes?

b. Prediction of the absolute level of efficiency the operational system will achieve. For example, assuming that the system is built to present design, how long will any particular mission take? If we were considering success probability as a criterion of system effectiveness, one could word the example in terms such as, what is the probability that the system will accomplish its mission?

Despite the fact that the model developer disavows a specific use for his model- other than to gather data- it is legitimate, for reasons explained previously, to ask whether HOS will answer the above questions. We ask this even though we keep in mind that the model is presently adapted only for flight tasks and that it outputs as a measure only time.

As long as we confine the criterion of system effectiveness to completion or performance time, it is obvious that HOS will answer the above questions. It will tell the user how long a mission will take, and if a time constraint has been placed on that mission, item (a) is a matter of simple comparison of simulated performance time with the system time requirement. Since system lag time is included in HOS, it will output a true system measure in the sense of including both machine and human functions.

Even if no time constraint or requirement has been levied on the system or mission, the time accumulated over one or more simulation runs will suffice to answer item (b).

Note that the above questions were phrased in terms of total system effectiveness. Because of the fine grained detail provided by HOS, items (a) and (b) above can be supplied for individual tasks and subtasks. Ordinarily a time requirement is levied on the total mission only; however, to complete the mission in that maximum time the mission time requirement must be apportioned among the tasks and subtasks making up the mission. In consequence, the individual tasks and subtasks have their own maximum time requirements. From that standpoint the fine detail supplied by HOS will come in handy in determining the effectiveness of the smaller behavioral units.

2. Design Analysis

Under this heading we consider the following:

a. Comparison of alternative design configurations in terms of the speed with which a mission involving each configuration can be performed.

b. Suggestions for redesign of a configuration.

c. Suggestions for initial design.

We can eliminate item (c) immediately. Since the method does not make use of data bank capability estimates associated with individual design features (as, one finds, for example, in the AIR methodology), the designer cannot refer to such a data bank for initial design suggestions. This is not a criticism of HOS; the methodology does not attack this type of problem.

With regard to the comparison of alternative designs, it is apparent that if one can get a measure of system efficiency (performance time) with one system configuration, it takes little more to revise that configuration, run HOS with the second configuration, and then compare the two configurations on the basis of mission time.

HOS is particularly efficient in making this comparison because, as we indicated earlier, the model is highly sensitive to design details. Thus, even a minor change in cockpit configuration should show up in a difference in performance time. Note the implicit assumption (which we consider entirely reasonable) that different design features have an impact on time to respond.

For the same reasons (design sensitivity and fine detail) HOS should clearly indicate where a design change would be desirable. This assumes of course that there is an explicit or implicit (i.e., apportioned) time constraint or requirement levied on individual tasks and subtasks. Analysis of the components (e.g., reach, shift or turning time, etc.) making up the total time for a task or subtask will suggest which components should be modified to reduce that total time.

Wherry points out quite cogently that HOS will output time and status data which can be used by other human factors techniques for analysis purposes (e.g., time-line, link, task and procedural analyses).

3. Manpower Selection

Under this heading we ask whether a model will suggest the types and numbers of personnel needed to operate a system. In particular we are interested in the aptitude/capability of required personnel.

HOS will suggest after fine grained analysis of simulation runs where the operator is having difficulty performing tasks within maximum time requirements. From this one might infer that an operator of greater capability is required, particularly if the source of the difficulty is in terms of information absorption time (i.e., recall capability). However, the model will not suggest any special aptitudes to meet these requirements. Since the model apparently deals only with a one or two man team, it is not sensitive to the need for variable numbers of personnel.

Consequently we can say that HOS does not output much information of value in the manpower selection area.

There is a suggestion in Reference 3 that the model might be of use in function allocation (which can be defined either as the determination of whether a function should be performed automatically or manually; or as the determination of the role of the human relative to equipment). Because of the model's sensitivity to design details, it is quite possible that HOS could be effective in either definition of function allocation.

4. Training

We may ask also whether a model supplies information relative to the content of required operator training or to amount of such training (course duration). As Wherry points out in Reference 1 (pp. 2-6, 2-7), "The outputs of Human Operator Simulator programs could be used to identify periods of maximal load, frequency of usage, etc. and thus determine what procedure and mission segments on which to concentrate during the training period." In particular, high load represented as the ratio of required operations to time allowed to perform these operations will suggest where greater emphasis should be placed during training and on what aspects to train. From that standpoint the application of HOS to training is immediate.

VALIDATION/APPLICATION STUDIES

As of this date information on validation of the HOS model is not available. Initial validation runs will commence in 1973.

From the standpoint of system effectiveness prediction, validation should consist of a comparison of a prediction output by HOS with actual operator performance on the system whose performance was predicted by HOS. The developer of HOS may disagree with this, on the basis that HOS is essentially only a data collection instrument.

Assuming, however, that a validation involving comparison of a HOS simulation with actual system performance was attempted, it could be performed in either of two ways. Operator performance could be measured in a physical simulation of an aircraft system; this would provide a large degree of control over the gathering of data and maximal correspondence between HOS and physical performance conditions.

Alternatively, system performance could be measured by a pilot actually flying the aircraft in the air. Here less control would be possible.

In either case it would be important to be able to measure actual operator performance in as detailed a manner as possible (e.g., down to subtask level and including if possible such elements as reach time, turn time, etc.). The reason is that even though overall system mission performance time might be very much the same in the actual mission as compared with the HOS simulation, the finer grained elements making up that time, such as reach, turn, information absorption time, etc. might differ significantly between HOS and the comparison system. When dealing with such elemental times, differences between HOS predictions and actual mission performance might be covered up by equipment "slop" in the actual system. This is not to suggest that this is what will actually happen; it is merely something to ensure against.

Admittedly it is difficult to instrument in order to measure such molecular behaviors; as Wherry suggests, this is one of the advantages of HOS, that it will permit prediction of elemental times that are difficult to measure ordinarily. However, in order to have confidence in HOS' parametric relationships, it is desirable to validate not only the overall model but also its elements.

In this connection a very detailed exposition of the experimental data on which hypothesized parametric relationships are based would supply partial validation of these relationships, if their empirical validation (by comparison with operational performance) is not possible.

Wherry notes (personal communication): "It must be understood that most operator data that have been collected and reported in the literature do not break time measurements down into the "microscopic" level needed by the Human Operator Simulator. The derivations which I have gone through would require a rather lengthy description and would not prove whether they will ultimately work in HOS or not. For this reason I see no great purpose being served by telling how I arrived at the equations at this point - we will ultimately know whether they work by next FY. If they work it really won't matter how they were derived; if they don't work I'll change them or collect the necessary data to derive them hopefully better than they presently are."

EVALUATIVE SUMMARY

Validity - No data available, but formal validation studies are planned.

Reliability - No data available, but the validation studies referred to above should be suggestive in this regard.

System Development Applicability

A. Comprehensiveness: Presently restricted to piloting operations but developmental work on the model will expand its capability to include almost any type of task.

B. Applicability: Model will output measure of system effectiveness (in terms of time), but is more specialized as a design diagnostic tool.

C. Timing: Model can be used at any stage of system development, provided that the detailed system information is made available.

Model Characteristics

A. Objectivity: Highly objective, few judgments required.

B. Structure: Reasonably well defined, although some pockets of obscurity relative to mediating processes remain.

REFERENCES

1. Wherry, R. J. The Development of Sophisticated Models of Man-Machine System Performance, in Levy, G. W. (ed.), Symposium on Applied Models of Man-Machine Systems Performance, Report NR69H-591, North American Aviation/Columbus, 3 November 1969.

2. Wherry, R. J. The Human Operator Simulator, Naval Air Development Center, Warminster, Pa., 25 May 1971.

3. Letter, Cdr. Robert J. Wherry, Jr., Naval Air Development Center, Warminster, Pa., to author, 25 June 1971.

APPENDIX I

HOS and "Operator Error"

Wherry has provided the following discussion:

It is obvious that from the discussion of HOS output, primary emphasis has been placed on the time it will take an operator to accomplish his tasks rather than on whether he has accomplished his tasks well or not. Indeed, from other sections of HOS it may even appear that the operator is incapable of making an error. To a certain extent this is true. For example, there are no provisions in the model for the simulated operator to either forget a step or to perform a step out of sequence. This is not an oversight, but a planned feature of HOS. The reasoning back of this design feature is that operators do not omit or accomplish steps out of sequence (either intentionally or unintentionally) unless they are inadequately trained or they are pressed for time because they have been given too many tasks to do for a specified time period. Certainly no evaluator would agree to using untrained personnel in a system and call it a "fair" test of the system, even though they didn't know what they were supposed to do. Secondly, there are two legitimate ways to evaluate whether the tasks an individual has been assigned are "too many" or "too complex" or both. One way is inform him that he must follow all the procedural steps and maintain his performance as best he can, regardless of how long he takes to accomplish his tasks. The second way is to tell him that he must complete all his tasks within a certain period of time, regardless of how well he does them, and he is free to omit steps if he feels that he must.

If a system is being evaluated for which sufficient time is available to accomplish all required tasks, following prescribed procedures, and to perform those tasks within required performance limits, then it should make little difference which set of instructions are given to the individual; his performance should be the same. Note that it is not maintained that it will be the same, but merely that it should be the same. It is possible that individuals working under the latter rule will omit steps that need not have been left out. If a system is being evaluated for which insufficient time is available to accomplish all the tasks "accurately" enough if all procedural steps are adhered to, then either method should uncover this problem. The results will be different depending on which set of instructions are given to the operator, but both should indicate that he could not do everything he was supposed to in the prescribed time period. Omitting steps in the "real world" may be making the best of a badly designed

system (it must be insisted that prescribed procedures are part of the system), but that does not change the fact that it is a badly designed system.

It may be argued that there are certain times and certain circumstances in which the operator should follow a simplified set of procedures. This, of course, is allowed in HOS, by defining both sets of procedures and the decision which must be made (i. e. , where the information comes from that allows the operator to decide which set to use).

While the above discussion indicates why HOS is designed in such a way that the operator cannot intentionally or unintentionally forget a procedural step, this does not mean that the simulated operator cannot make other kinds of errors. Below are partial list of the many types of "operator errors" that can be made in HOS:

a. The operator relies on recall of some display which, unknown to him has changed in value since he last looked at it. This in turn leads to an erroneous decision or calculation on his part.

b. The operator desires to maintain some display between some upper and lower limit values, however, he is so busy doing other steps that he fails to notice, for some time period, that the display has exceeded the allowable limits.

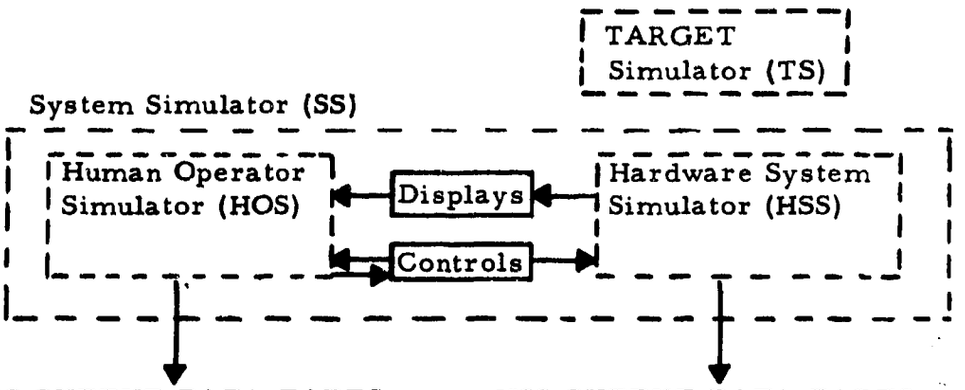
c. The operator is attempting to maneuver his vehicle to a certain position, but because of the complexity of the functions he must calculate (e. g. , no predictive display is available), and the disparate locations of the displays he must use to provide him information for the functions to be calculated, he takes too long in accomplishing the function and is incapable of following an acceptable path (e. g. , in a carrier landing he may get "behind the problem" which leads him to over correct and land too short, too long, or too hard to get a "waveoff".

The ability to identify these kinds of errors is contingent on the ability to correlate the output of HOS with the output of the Hardware Simulator, and to have criteria for what the system should be doing. It certainly cannot be argued that in the above cases no error occurred. However, it is questionable whether all of these constitute human errors. If one approaches the problem from the standpoint of whether the problems could be alleviated through better human engineering (inclusion of warning signals, allocation of certain functions to the hardware rather than to the

man, reduction of the complexity of tasks assigned, etc.) it must be recognized that the same man might be able to do the job if the system were adequately human engineered. If the operator is doing the best he can, given the constraints of too many tasks, too complex tasks, poor station layout, etc., should the operator be blamed because he can not do everything we would like him to be able to do? The poor performance of the system is not the operator's fault; it is the system designers who are culpable.

APPENDIX II

LIST OF HOS DATA OUTPUTS



HOS OUTPUT DATA TAPES

- o Time Log of What Operator is Doing
- o Sampled Time History of Displays Being Monitored Which Are Out-of-Limits

HSS OUTPUT DATA TAPES

- Sampled Time History of What Hardware States Are

Human Operator Data Analyser/Collator (HODAC)

- o Analyses HOS Output Data Tapes
- o May Be Cued To Ignore Certain Time Periods
- o May Be Cued To Perform Only Selected Analyses

  - Display Usage Analyses
    - Link Values
    - Time Spent:
      - Activating
      - Accessing
      - Absorbing & "Dwelling"
      - Monitoring
      - Deactivating
      - ...

- Control Usages Analyses
  - Link Values
  - Time Spent:
    - Activating
    - Accessing
    - Manipulating & "Dwelling"
    - Deactivating
    - Estimating
- Mediated Function Usage Analyses
  - Link Values
  - Time Spent:
    - Estimating
    - Calculating
- Anatomy Usage Analyses
  - Eyes
  - Left Arm/Hand
  - Right Arm/Hand
  - Feet

- Memory Usage Analyses
  - Time Spent Recalling
  - "Successful" Recall
  - "Unsuccessful" Recall

- Decision Usage Analyses
  - Time Spent Deciding
  - Number of Decisions Made
- Loading Analysis
  - Looking Toward Devices
  - Looking At Devices
  - Manipulating Devices
  - Reaching For Devices
  - Memory
  - Decisions
    - ...

- Procedure Tracking Analyses
  - Number of Times Used (Activated)
  - Average "Start" to "Stop" Time
  - Average Time Spent Actually Working On
  - Average "Dwell" Time While Working On

Average Times Since Procedure Start To Step i
Percent of Times Step i Reached (Alternate Path Analysis)

Monitoring Behavior Analyses
- Displays Monitored
- Functions Monitored
- Number of "Out-of-limits" Occurances
- Average Time "Out-of-limits" Before Detection
- Average Time to Bring Back within limits
- Number of Times Looked At While Monitoring
- Time Spent Accessing and Looking At While Monitoring

Multi-Mode Analyses (Accomplished Only for Multiple Discrete Controls)
- Total T ıe in Each Mode (Position)
- Average Time in Each Mode
- Link Analysis of Multiple Positions (Modes)

Interrupt Analyses
- Number of Interrupts
- Type of Interrupts
- Time to Process Interrupts

V. ORACLE - DYNAMIC TIME LINE SIMULATION

INTRODUCTION

ORACLE- whose acronym stands for Operations Research and Critical Link Evaluation- is, as the name implies, operations research-oriented. By this we mean that it betrays this influence by its ancestry and the types of concepts it incorporates. The model derives from attempts to develop commodity flow models to determine the maximum capability of the system (see Reference 1 for the antecedents of the model). The essence of the model is that it performs what OR specialists call a "traffic analysis" of the system which describes the flow of events over system channels over time and the queues that may build up. More details about this will be given later.

The point is that this model derives from a different conceptual framework than the others we have considered, essentially a non-behavioral framework, although it will be seen that many of the model's inputs and outputs are very similar to what more behaviorally oriented models supply.

One of the paragraphs in Reference 1 (p. 3) summarizes the technique very nicely:

> "ORACLE... is a general purpose system simulation technique. The technique allows modeling any multicommodity system flow on a scientific digital computer. The system to be analyzed is defined in terms of a general link-node network which is stressed with inputs at one or more points and the resulting queues, delays and throughput rates are presented for the analyst."

GOALS

ORACLE is a diagnostic rather than an evaluational tool. It does not provide a quantitative assessment of the effectiveness of a system, but by supplying a picture ("snapshot" is a term frequently used by its developers) of the operation of a system, it permits diagnosis of system elements that have resulted in an inbalance between inputs and outputs, with consequent performance queues. "A performance queue is defined as a system component (man-man, man-machine, machine-machine) interface capability that is less than optimum in response to given input conditions" (p. 2, Ref. 2).

Through its diagnostic capability the model "provides a ready means of assessing the effects of possible changes to the system under study" (p. 3, Ref. 1). We recognize this as a means of redesigning the system by introducing various changes in system elements and determining the effect of the change on system performance. This appears to be the primary goal of the model, although when we discuss the uses of the model it will be seen that this general methodology has a number of possible uses.

It is also possible to use the model for initial design.

> "For example, ORACLE was implemented in a command and control study performed for the United States Air Force during the definition phase of the Airborne Warning and Control System (AWACS). In that study, a complete baseline was derived, including personnel, display and control, data processing, and communications subsystems. The initial step was to establish a "strawman" configuration for each of the subsystems included in the analysis. No attempt was made to relate personnel to speciality areas during initial simulation cycles. The basis for the "strawman" was derived from paper analyses, engineering experience and judgment, and in some cases, intuition. As further iterations to the "strawman" were made using the outputs from ORACLE, the qualitative/quantitative measures of man-machine requirements became firmer, resulting in a baseline system configuration." (p. 2, Ref. 2).

The developers also point out that the model can be used as a source of empirical data. As a simulation of the system, it can be tested and retested any number of times under varying conditions, and the history of the system performance can be examined critically. Hence it is possible to test an hypothesized system before it is even built. This capability is of course inherent in any simulation model and in fact represents the outstanding capability of such models.

ASSUMPTIONS

Because the model is not behaviorally oriented, it has relatively few assumptions (at least of the behavioral type) to be considered. One basic assumption is of course that one can conceptualize systems in general in terms of a flow of data, information or events from one point

to another over time. From this conceptualization stems the basic characteristics of the model. For example,

> "... events are composed of nodes and links. A node may be considered as an element in time that receives, processes, and responds to specific single or multiple inputs, and provides a single or multi-channel output. Nodes represent combinations of man-machine interfaces performing discrete tasks to complete an event. Interconnections between nodes are provided by unit transfer paths called links. Links provide for the transfer of data, information, materials, goods, and so forth between nodes, and also represent man-machine interface combinations..." (p. 3, Ref. 2).

It is of course the developer's perogative to make whatever assumptions he wishes. One's only concern is for the consequences of the assumption. In the case of ORACLE, the assumption that all systems can be conceptualized in terms of a flow of events raises the question whether all types of systems can in fact be so modeled. Obviously communications systems can, but can systems emphasizing decision-making or sensing or motor operations? It would appear that the concept is general enough to encompass them, but the question arises as to the relative efficiency with which various systems can be handled with such a concept. Evidence on this point will be considered in the discussion of validation studies.

Another assumption inherent in the model is the concept of system effectiveness represented as a balance between inputs and outputs. Given a system requirement, that system is effective when input rates produce required outputs without any performance queue developing (or rather any "intolerable" queues, since it would be unreasonable to expect a complete absence of queueing). The components of the system are visualized as required to process these inputs (which define the "load" in the system).

In the truest sense of the word ORACLE is a workload model. The concept of workload has been encountered before, particularly in Siegel's digital simulation technique, although in the latter case it was necessary to introduce behavioral concepts such as "stress" to account for overload performance.

The concept of workload as being the critical measure of system effectiveness carries with it the implied assumption that time is the critical dimension of system performance. We have encountered this assumption in other models also, e.g., Wherry's HOS. There is an

implication here that the system and the operator will always perform well enough to accomplish the system requirement; the only question is, how long will it take the operator/system to accomplish its goal. In other words, "quality" of performance is thought of purely in terms of the time dimension. Presumably delays and queueing could be great enough to threaten accomplishment of the system requirement, but this would be accomplished only through "intolerable" queues.

The question to be asked is whether a system evaluation in terms of time alone would be completely satisfying to a design engineer who may feel that even when input rates are not excessive, the operator may still make errors and perhaps not accomplish his task.

Finally, although the model employs no overt behavioral assumptions, it derives its input data from a variety of sources, some of which, like the AIR Data Store, have behavioral assumptions of their own. Does one, as it were, assume the assumptions of one's input data sources? This is a moot question, since it is not feasible to regress indefinitely in questioning data sources.

METHODOLOGICAL SCOPE

The developers of ORACLE stress its diversified application "to any system involved with transfer of material and/or information from point to point" (p. 11, Ref. 1). There are three major system applications: to military operations, production and routing (movement, information, transfer). As noted previously, the basic concept of traffic analysis "is not intended to imply only an analysis of data traffic". However, the major proposed applications of the model seem to be to systems whose operations can be conceptualized in terms of movement, information transfer, communication and the like. The developers also recognize that continuous tasks, "that is, tasks of a monitoring nature, present a problem..." (p. 9, Ref. 2), but they presumably have a mechanism to handle the difficulty. "Continuous task time delays may be controlled by employment of an "and gate" node, by incorporation of a "clever" node... or by assigning a time delay to a "normal" node..." (p. 9, Ref. 2), terms to be defined later].

MODEL PARAMETERS

We have already mentioned that model events are composed of nodes and links (note the similarity- in part- to Siegel's formulation of DEI in terms of information theory).

> "Five basic node types have been developed to provide the manipulation of traffic flow through an operational sequence. These nodes are: "normal", "clever", "and gate", "or gate", and "dummy". The "normal" node represents a specific system entity with a defined man-machine interface operating at a specific percentage of loading level and which requires a discrete response time. Similarly, the "clever" node also represents a system entity; however, it possesses the capability of providing a Monte Carlo input source for generation of an alternate path. It is also capable of accepting time milestones from the scenario to terminate a sequence, thus initiating the next phase upon expiration of a predetermined clock time. The "and gate" node is similar to both the "normal" and "clever" nodes, except that its initiation depends upon completion of a predetermined number of combination of input links. The "or gate" node is identical to the "and gate" node, except that it initiates an output over the appropriate link to the first available noder, that is, the "or gate" is an optimum-time path locator. The "dummy" node is provided only to allow continuity in the flow sequence without imposing time constraints upon units passing throughtthe system. The "dummy" node represents a nonentity, and is assigned zero time and an infinite capability (loading level)." (pp. 4,6, Ref. 2).

A queue can also be considered a model parameter. A queue is formed at a component if the rate of inputs arriving at the component is greater than the processing rate.

DATA

1. Input Data

Input data to be applied to the model are largely time oriented. A list of input data would include:

a. input rates for units;
b. initiation times;
c. response times for events;

d. event (node) priorities;
e. probabilities of event occurrence based on machine availability and reliability criteria.

In addition the development of the model will specify the functional elements of the model (i.e., its man and machine components) and the links between them.

The uses of these input data are fairly clear. Input rates (e.g., the number of messages per unit time) impose loads on the system. Initiation times are required to know when to initiate an event, and obviously response time is needed to determine delays before the next event can be triggered.

Event priorities must be assigned because some units are more critical than others.

> "In such cases, the noncritical task is deferred and will go into queue until the priority unit has been processed and transmitted. It is also possible to limit the maximum queue size, or time in queue, for low-priority units. This is desirable because such units may never be processed if the input rate for critical priority units is very high. That is, if the critical unit input rate is almost continuous, or if a queue builds up, low-priority units would be deferred each time they tried to enter the node. Therefore the logic of priority assignment depends on system requirements. For example, some unit types may require immediate processing, while others may be deferred for a certain period of time without system degradation. Also, certain units may be deferred until a specific quantity accumulates, which will then be processed..." (pp. 8-9, Ref. 2).

Probabilities of event occurrence are required to account for delays resulting from the need to repeat erroneous actions or to select an alternative pathway when one pathway is clogged.

> "Alternative path diagrams can also be constructed to provide for incorporation of contingencies. The alternate path capability provides a means of injecting probabilities of failure into the simulation by using Monte Carlo techniques to provide a predetermined probability of occurrence. The probability of failure may represent equipment reliability or human error probabilities.

An alternate path may be incorporated at any node within the sequence diagram, provided that the probability of alternate path occurrence and alternate route sequences are designated." (p. 3, Ref. 2).

Obviously the model accounts for erroneous actions (failure to accomplish the task), but is interested in these primarily in terms of the time delays they produce.

2. Input Data Sources

The model makes use of the same data sources that almost all other models use. "Time estimates for planned systems are extracted from previous dynamic tests, and through experimentation with similar system tasks." (p. 4, Ref. 2). A second primary source is the AIR Data Store. "Time estimates for existing systems are derived from time studies available from sources such as industrial engineering or from personal judgments provided by experienced personnel" (p. 4, Ref. 2).

Obviously very detailed information is required to simulate individual operator/machine actions. Some of these data derive from the characteristics of the system being modeled. (i.e., input rates, initiation times and event priorities). This suggests that fairly detailed analysis of the system being modeled is required. Operator error (event) probabilities and operator response times could be derived from experimental literature, once one knew the characteristics of the individual machine components with which the operator interacted. Equipment reliability values could be derived from the fairly extensive equipment reliability literature on equipment components.

The need to define the system to be modeled in fairly complete detail might present a problem during development. "For example, a display and control device required for a specific task must be completely defined and its characteristics determined... The information required, information available, evaluation and decision-making process... must all be taken into account" (p. 4, Ref. 2).

Because the input data are so heavily dependent on system characteristics, the need for an extensive external data base would seem to be correspondingly reduced.

3. Output Data

"The basic measure of performance is the achievement of critical events within the time span dictated by the scenario, i.e., did a particular component achieve required processing rates, etc., or did queues form due to an inability to complete the operational procedures within required time spans?" (p. 7, Ref. 1).

The developers consider that ORACLE can supply various outputs. As we understand them, they are listed below:

a. A prediction of the total processing time required for a given sequence of events.

b. The identification of queues (length, type) on critical paths.

c. Related to (2) the determination of those components (e.g., machine, man) associated with these queues.

These outputs can be summarized in the following sentence: "The primary output generated by ORACLE is a qualitative/quantitative measure of man-machine workload as related to both system requirements and time constraints imposed by tasks" (p. 1, Ref. 2).

As indicated initially, the model does not output an estimate of system effectiveness, but rather (as in the case of HOS) a time line history of system performance which can then be analyzed for the particular feature of interest to the modeler. Thus, one might look for factors responsible for a queue, or the characteristics of a component at which a queue develops. This analysis is enhanced by the capability of calling

"for any number of snapshots at selected times. Each snapshot consists of a complete instantaneous picture of the system at that time. That is, the location of every unit, identified by specific position (link or node) within the system is listed. The analyst then knows which units are in queue at that exact time, the size of all queues, and the loading of all nodes and links within the system" (p. 10, Ref. 2).

FUNCTION AIR-TO-GROUND CCOR
EVENT ASSESS FINAL TRACK PARAMETERS
MODE NO. 102016
DATE 11/19/68

TASK/SUB-TASK NO.	TASK/SUB-TASK DESCRIPTION	TIME (SEC)	DISPLAY/CONTROL DEVICE	DISPLAY INDICATION &/OR CRITICAL VAL.	OPERATOR IDENT	OPERATOR ACTION/DECISION-MAKING	RESPONSE ADEQUACY	CHARACTERISTIC ERRORS, MALFUNCTIONS &/OR REMARKS
1.	ASSESSMENT OF FINAL TARGET TRACK PARAMETERS ON HUD AND MSD	5.0	HUD MSD ADI-BACK-UP HSI BACK-UP AIR DATA o ALTIMETER o RATE OF CLIMB o ANGLE OF ATTACK o TAS o I AS o "G" METER o MACH	HUD o TAS o COMMAND AIR-SPEED o TRUE ALTITUDE o COMMAND ALTITUDE o HEADING o COMMAND HEADING o HORIZON LINE MSD (OFFSET PPI) o OWN A/C VECTOR o TARGET POSITION o TARGET PRIORITY FF	PILOT	ASSESSMENT IS MADE ON THE BASIS OF MEETING THE SELECTED MISSION PROFILE VERSUS FLIGHT DYNAMICS ENVELOPE. THE DIGITAL COMPUTER WILL PROVIDE A RECOMMENDED ATTACK VECTOR BASED ON TRACK PARAMETERS AND FLIGHT DYNAMICS, ETC.	TO BE DETERMINED	PILOT CORRECTS FLIGHT PROFILE IF REQUIRED TO MEET WEAPON RELEASE REQUIREMENTS AS PREDICTED BY DIGITAL COMPUTER AND SENSOR TRACKING DATA. CORRELATION IS MADE BETWEEN HUD AND MSD AS REQUIRED TO MAINTAIN TACTICAL AND FLIGHT DYNAMICS BALANCE.
2.	INITIATION OF TARGET ENGAGEMENT FOR ATTACK	2.0	- ADDRESS o NUMBERIC (0-9) - MISSION CONTROL o TARGET ENGAGE - HUD ANNUNCIATOR o TGT ENGAGE - MSD o TGT ENGAGE	- HUD o ENGAGE SYMBOL o TIME TO GO -MSD o TARGET LOCATION o ENGAGE SYMBOL o A/C VECTOR	PILOT	IF THE PILOT AGREES WITH COMPUTER VECTORING, HE ACTUATES THE TARGET ENGAGE SWITCH TO INITIATE STEERING COMPUTATION.	TO BE DETERMINED	TARGET ENGAGE SWITCH LIGHT ON MISSION CONTROL PANEL IS ACTIVATED. HUD ENGAGE ANNUNCIATOR LIGHTS, AND ENGAGE SYMBOL IS PLACED ON TARGET IN MSD.
3.	PERFORMANCE OF FINAL DETAILED CHECK OF AIRCRAFT PERFORMANCE STATUS	2.0	SUBSYSTEM STATUS ENGINE STATUS	SUB-SYSTEMS GO NO GO INTOLERANCE OPERATING RANGE	PILOT	PILOT MONITORS SUB-SYSTEM STATUS PANEL FOR EQUIPMENT STATUS, ALSO SCANS ENGINE INSTRUMENTS FOR PERFORMANCE CHECK.	TO BE DETERMINED.	THE SUB-SYSTEM STATUS PANEL PROVIDES DISCRETE STATUS OF EQUIPMENTS BY NORMAL, DEGRADED, MALFUNCTION AND EMERGENCY INDICATIONS. AN AURAL TONE IS SUPPLIED WITH EMERGENCY INDICATION. NORMAL-A "GO" CONDITION EXISTS INDICATING THAT ALL EQUIPMENT IS OPERATING WITHIN

Page 1 of 2
9802A-VB-3

Figure 15. Operational Procedures Analysis Format
(Taken from Ref. 1)

PROCEDURES FOR MODEL APPLICATION

I. Analytic Methods

Three primary tasks are required before the model can be run:

a. Establishment of a scenario which structures the operational environment of the system in terms of time. It is necessary to specify the quantities of particular units flowing through the system, establish unit input rates, initiation times, times of occurrence of selected contingencies, etc.

b. Development of operational sequence diagrams (OSD), whose purpose is to identify the events which must occur to perform each mission function and to define unit traffic flow within the system. The OSD also includes alternative path diagrams to account for contingencies, etc.

The OSD referred to here is not the same as the one familiar to human factors specialists, as defined by Kurke [1]. It is at a detailed function or gross task level ("identify and receive new track, radar continuous scan"), whereas the Kurke OSD describes the detailed task element level. The detail required for system modeling is provided by the operational procedures analysis.

c. Performance of an operational procedures analysis: "The operational procedures analysis is required in order to define the specific task characteristics necessary for event performance within each operational sequence diagram. A secondary purpose is to provide a data base for the evaluation of events shown to be critical by the simulation output. The complexity of the analysis is dependent on the level of detail required to obtain sufficient data for simulation input purposes". (p. 3, Ref. 2). Figure15 presents a representative operational analysis format. Note that performance times for each task is indicated, together with control/display device characteristics and the critical values of these, the operator action or decision-making process, response adequacy and characteristic errors.

Obviously a very considerable amount of analysis of the system to be modeled is required. Although this analysis provides most of the information produced by a behavioral function/task analysis, it seems to require no behavioral judgments such as might be required to assess stress level, for example. This is perhaps an exaggeration; to describe decision-making

1. Kurke, M. I., Operational Sequence Diagrams in System Design. Human Factors, 1961, 3, 66-73.

processes, errors, etc. must implicitly involve at least a minimal amount of behavioral framework.

2. Methods of Synthesizing Data

Because this is a simulation technique which provides finely detailed time histories, no synthesis (combination of outputs at a lower level to derive a value for a higher level of behavioral unit) is required. As noted in other discussions, this is a definite advantage for simulation models, because it avoids the necessity for combinatorial statistics. At the same time, however, it is a disadvantage in the sense that it does not provide a single figure of effectiveness merit for the system as a whole. This must be inferred from an analysis of the individual features of the time history and must therefore be somewhat judgmental.

3. Input Data Application

The input data required for each event are related to the node and link units, that is, to the individual operator and his control/display interfaces. This follows because the only system elements in the model are the node and link units.

ANTICIPATED MODEL USES

We have considered other models in terms of the following possible uses:

(1) Prediction of System Effectiveness

(a) Determination of the absolute efficiency of the system (i. e., how long does it take the system to perform its mission, how reliably will it perform that mission).

(b) Determination of relative system efficiency (i. e., in relation to a specified system requirement, how reliably will the system accomplish that requirement).

(2) Design Analysis

(a) Comparison of alternative system configurations in terms of system performance criteria.

(b) Determination of design inadequacies in the system and indication of redesign alternatives.

(c) Assistance in the initial design of the system.

(3) Manpower Selection

Specification of the required numbers and types of personnel needed to run the system, with emphasis on needed aptitudes and skills.

(4) Training

Assistance in the determination of training required by personnel to operate the system.

ORACLE, like most of the other models we have reviewed, appears to be primarily effective in the prediction of system effectiveness and in design analysis. Obviously, exercising a system with a given load of jobs to be performed and accumulating the time required by the system to process these jobs will indicate how long it takes the system to perform its mission. Given that one knows what the system requirement is (in terms of maximum time to perform), it is a simple matter to compare actual performance with the time constraint and determine whether or not the system can perform within that constraint.

The particular utility of the model lies in its capability to compare alternative system configurations; indeed, it almost seems from the descriptions provided that that is what it was intended to do.

> "The analyst may then revise the system configuration by changing the number of machines or personnel, or by altering capability at critical points. Stated simply, a man-machine interface may be adjusted in either of two ways to eliminate or reduce a performance queue to a tolerable level. Output rate may be increased by improving man and/or machine performance through qualitative/quantitative adjustment of personnel or machine, or by adjustment of operational sequence of procedure in order to reduce response time. Input rate may be decreased only by changing mission goal requirements or by changing the system operational environment. After the revision is made, the model is re-run and again analyzed. The new queue structure is formed and analyzed, and the impact of stable and tolerable performance queues can be assessed..." (p. 1, Ref. 2).

It should be noted that because the model simulates the system so closely, it should be highly sensitive even to minor variations in configuration, although only practical application experience would demonstrate this. In general, models focussing on the time dimension as an output measure appear to be more sensitive to equipment configurational differences than are models which rely largely on error probabilities.

Because of the detailed time history that is provided and the detail with which the system is simulated, the model is particularly effective for redesign. The model points out factors related to mission success, e.g., queues on critical paths.

> "If a "fix" or adjustment to the system is made, and its impact on time sensitivity is evaluated by manual timeline techniques, a major effort would be required by the analyst. Using the simulation model, it is possible to change several data parameters and to determine the effect of the proposed change in the time required to submit the job to the computer, and at a momentary investment of from 5 to 30 minutes of computer time." (p. 10, Ref. 2).

> "Another type of output obtained is a measure of the impact of planned system revisions. There are two aspects of planned system revisions. One is related to changes in the design of the units flowing through the system. The second involves changes or revisions to the operational sequence or procedures. Both of these involve a simple matter of changing a few data parameters and/or changing the operational sequence flow diagram." (p. 10, Ref. 2).

It has already been pointed out that the model can be used for initial design. This involves setting up a preliminary crude system configuration, testing it, analyzing the output of the simulation, revising the system configuration in the manner described above, running another simulation, comparing it with the results of the first test, etc. - an iterative procedure that involves successive comparisons of alternative configurations each of which becomes progressively more detailed.

Because the model does not include specific behavioral parameters such as skill level, it is not very effective for manpower selection. However, it is possible to test the effects of adding or deleting manpower in the system, and of changing the allocation of functions between man and machine (e.g., redistributing task responsibilities).

Similarly the model outputs do not suggest what the training content required for personnel would be. Again this may be because the model does not include behavioral parameters indicative of training needs, although we have noted this same difficulty in other more behaviorally oriented models. Presumably, however, if intolerable queues built up at a human interface, it might be decided after detailed analysis of the problem that additional training was required. However, the specification of training content or required duration would have to come from analysis of system requirements rather than of model outputs. It is interesting to note, however, that the developers indicate "the model will accept predicted learning curve data and as the number of simulated missions increases, system performance will correspondingly improve" (p. 11, Ref. 2).

The developers indicate that "specific usages of ORACLE include, but are not limited to:

"a. Determining the number and type of personnel required for a given task mix and system configuration.

b. Measuring the effect of changes in personnel/task/configuration on the system effectiveness.

c. Testing the capability of the configuration to handle the simulation load.

d. Identifying critical elements in the operational sequence.

e. Measuring the effect of degradation of individual system functions." (p. 5, Ref. 1).

Some information about ORACLE as a computerized technique might be of interest.

"ORACLE is a system of 46 functional subroutines which are assembled and loaded as a "program" according to the definition of the system being modeled. These routines perform the following functions:

* Dynamic core allocation
* Event calendar generation
* Random number generation

* Output generation
* Statistical analysis
* Diagnostics

Of the 46 subroutines, 26 provide the functional attributes of Simscript. ORACLE is written in standard ANSI Fortran IV. . . Two additional features . . bear mentioning. First, the routines use list processing to make maximum utilization of core storage. This permits a relatively large and complex system to be modeled without overburdening the computer and/or the programmer. The second feature is the use of an event calendar rather than a fixed time step resulting in more efficient machine execution time." (pp 9-10, Ref. 1).

". . . the model requires approximately 50,000 words of core storage. In addition, five high-speed "scratch" files providing a capacity of up to 100,000 words, and one magnetic tape file are required. The model has been successfully run on the UNIVAC 1107 and 1108 machines. Simulations have also been made on the IBM 7000 series although on a much smaller scale. Actual running time on the 1108 were within a range of from 5 to 30 minutes for computations with printout requiring from 40 to 240 minutes. As the complexity of the model increased within the creation of additional subroutines, the running time and required core allocation prohibited use of the model to full capacity. This was overcome by the use of dynamic core allocation routines to replace the large number of partially used dimension arrays with a single array which is continuously in use." (p. 11, Ref. 2).

VALIDATION/APPLICATION STUDIES

The model was initially developed for the AWACS system for USAF, in which it was used in the design of the command/control system. Subsequently it was used to model a radar production line. It has also been used in the design of the Test Complex Surveillance and Control System (TCSCS) and in the evaluation of Mallard. The model has also been applied to other production line situations and to a traffic management study.

This listing reveals that the model has been applied to a variety of systems and problems, presumably with some success. However, the

degree of success in the application of the model cannot be determined from the available reports.

Ordinarily the major evaluational criterion for a model is validity - the degree to which a prediction is replicated by real-world testing performance. The case of a model which is used primarily for diagnosis is somewhat different. Obviously the model output must correspond to some extent with real world experience, although because the model does not derive a single figure of merit it may be difficult to determine that degree of correspondence. However, there are applications where other factors may be of equal importance. We refer to assistance in the initial design or redesign of a system. In the case of initial design, the criterion of validity does not really apply, because one is not predicting anything except that the system developed in this way will work with reasonable effectiveness. In the case of redesign, the fixes recommended must remedy the difficulty. In such cases the capability of the model may be difficult to isolate from the effects of applying other (e.g., human engineering) techniques.

In these situations we are dealing with a utility rather than a validity criterion. Because of the very applied nature of these situations, a model may be highly useful even where it supplies only partial answers or is in error to a certain extent.

The validity criterion applies only to the first of the four uses of a model: prediction of system effectiveness. Naturally a model cannot be absolutely invalid and still be useful for design, manpower and training purposes. However, if one had two models, the first of which had slightly greater validity than a second, while the second had significantly greater utility than the first, the second would, in our opinion, be preferable. It is incorrect therefore to place sole reliance on a model's predictive validity.

What we have said should not be interpreted as reflecting on ORACLE's validity; we have no evidence for it, one way or another. We are, however, impressed by its potential as a design instrument.

EVALUATIVE SUMMARY

Validity - No formal data available, although there are reports of a number of successful applications.

Reliability - No formal data available, although if the applications are successful, one can infer a reasonable degree of reliability.

System Development Applicability

A. Comprehensiveness: Model may be limited to systems whose operations involve movement, information transfer, communication, etc.

B. Applicability: Model outputs measures which can be evaluated in terms of system effectiveness, but its primary use appears to be for design diagnosis.

C. Timing: Can be applied at all stages of system development.

Model Characteristics

A. Objectivity: Highly objective; few judgments required.

B. Structure: Few behavioral assumptions necessary.

REFERENCES

1. Biddlecomb, R. et al. The ORACLE System Simulation Model. Unnumbered, undated report, Westinghouse Aerospace Division, Baltimore, Maryland.

2. Stinnett, T. A. ORACLE- A Dynamic Timeline Simulation Technique. Unnumbered, undated Westinghouse report reproducing a paper presented at the IEEE-GMMS ERS International Symposium on Man-Machine Systems, 8-12 September 1969, Cambridge, England.

VI. PERSONNEL SUBSYSTEM EFFECTIVENESS MODEL

INTRODUCTION

The model described in this section was developed by Dr. D. T. Hanifan working at the Western Division of Dunlap and Associates. The model falls into the simulation category because of its use of a computer to pace through the tasks which will be performed operationally by the system. Although the model is not fully articulated (the developer indicated in a phone conversation that the report on which this description is based did not include all its details, which have unfortunately not been written elsewhere), the reader will recognize a number of features familiar to him from previous model descriptions.

GOALS

The purpose of the Personnel Subsystem Effectiveness (PSE) model is generally to assess the adequacy of a planned system design (i.e., its operator and/or maintainer aspects, with the term "design" implying much more than merely equipment details) during system development. Assuming that quantitative PSE standards or requirements are specified, the probability that the human tasks included in system design will satisfy these requirements can be tested.

A glimpse at the more specific objectives of the method can be secured by examining what its outputs are supposed to be.

"The products of such analyses can be used to:

- ... improve hardware design
- ... evaluate and improve operational and maintenance procedures
- ... evaluate computer software to improve human performance compatability
- ... assess the adequacy of planned training in terms of impact on system effectiveness
- ... provide a test and evaluation tool
- ... provide a valuable data base for future operational and system design improvement" (Ref. 1, p. 20).

If one can draw valid interences from the title of the model, the latter is designed to deal with the entire Personnel Subsystem of major systems. We say that because the associations with the term "personnel subsystem" call up images of large complex systems like the Atlas ICBM with which

the author was associated. Consequently, the model is a very ambitious one, but not more so than a number of other models we have so fa. considered.

ASSUMPTIONS

A basic assumption is that "effectiveness of the personnel subsystem is defined as the probability that mission-critical demands on personnel will not exceed personnel capability during a given mission (or across many missions)" (Ref. 1, p. 1). This assumption is implied in a number of other models reviewed (e.g., Siegel's digital simulation model), as it is of course a fundamental Human Factors tenet expressed, for example, in time-line analysis; but the PSE model makes this assumption a keystone of the methodology. For example, "The mission model or scenario defines the requirement (demand) for performance levels and task accomplishment time" (Ref. 1, p. 6). Those requirements are likely to be very time-oriented, because the phrase "mission profile" is used with reference to the model's input data; the profile suggests a time-line analysis. However, as we shall see, the profile may include requirements in terms other than time, e.g., error probability.

In any event, the definition in the previous paragraph suggests a process of matching requirements against anticipated performance, which would necessitate that the system description on which the PSE model is based contain very explicit requirements for each task being simulated. This may pose relatively little difficulty with regard to completion time requirements, but may with regard to performance level requirements, since the latter are not ordinarily specified in most of the system development projects with which the author is familiar. It should be noted that "Effectiveness has both personnel capability and availability terms. Personnel capability is a complex function of preassignment training, on-the-job training, experience, fatigue, motivation, morale and numerous incidental factors, and is generally difficult to cope with satisfactorily in a quantitative way during typical system development programs. Personnel availability can be determined in a more straightforward manner by accounting for demands and assignments throughout the system... "The availability term is simply the probability that appropriate personnel will be available upon demand given a complete accounting of all conflicting demands and personnel locations throughout the system, including such non-mission activities as eating, sleeping, sickbay, and the like." (Ref. 1, p. 1)

Personnel capability can be expressed as probability of task success. However, because capability is recognized as being affected by diverse factors (note the resemblance here to Swain's "performance shaping factors"), the input data distributions which are matched against task requirements must specifically include these factors. In other words, it is not satisfactory to specify that the probability of accomplishing task X is simply .9875, without indicating that the probability will, in the presense of high or low motivation, for example, be increased or decreased. This means that to include these diverse capability factors it is necessary to supply not one but several input data distributions relevant to any given task. This is not to say that this cannot be done, but it imposes a severe strain on input data requirements. In the present inadequate state of our knowledge of behavioral processes, it is possible to utilize the model, but it is unlikely that all the many factors influencing human performance can be included- at least at present.

Despite the recognition of the many factors that may influence personnel capability, the model includes no further behavioral assumptions of the type we have seen in other models, e.g., the relationship between stress and performance, although the developer recognizes the potential effect of stress ("Thus, time stress may decrease performance level..." Ref. 1, p. 9). The model assumes various personnel types, but it is unclear whether the term "type" refers to personnel specialized for a particular job or to some other dimension. We would interpret it to mean personnel specialization, rather than something like Siegel's F distribution, because personnel type is closely linked to the definition of personnel availability.

METHODOLOGICAL SCOPE

The scope of the model is quite broad, including both operator and maintenance personnel. There is apparently no restriction on the type of task or behavior which the model can presumably handle.

PARAMETERS

An essential parameter of a requirement-oriented model such as this is the "demand" imposed by the task, job or position. Demand is not further defined, although the developer indicates (p. 2, Ref. 1) that "demands... are distributed in accordance with... the probability density

of demands on personnel capability..." One could of course define demands as being equivalent to the performance requirement imposed by the task on personnel, but in that case the parameter would seem to be superfluous.

Whether or not it means the same as demand, performance requirement is a critical model parameter. Presumably a requirement is associated with each task or subtask in the system scenario. There may in fact be several such requirements, particularly time and performance level which "may be expressed in terms such as accuracy, force exerted, tracks initiated, etc." (Ref. 1, p. 9).

We have already discussed the concepts of personnel capability and availability.

DATA

1. Input Data

A great variety of input data are required, as may be seen from Table 12 (Ref. 1, p. 8). Input data are of the following types:

1. The mission profile (the scenario) which is derived from the system description by means of task analysis.

2. The requirements associated with tasks, also derived from the same source.

3. Proposed system manning, also derived from the system description.

4. Probability distributions of task completion times and successful task performance, which are derived from a variety of external sources, such as data bank, the results of simulation testing, subjective judgments, etc.

5. Priorities associated with contingent probabilities and alternative action strategies.

6. For the associated Life Cycle Cost Model (which we do not consider in this description because it is not an effectiveness model), man-hour and associated support costs.

TABLE 12. Input-Output diagram for PSE Model.

Sources	Inputs	PSE and LCC Models	
• Work Study & Human Factors	Proposed Manning (numbers of each type by work-space, task, job or position	1. Combines analytic and simulation approaches to compute personnel subsystem effectiveness, P, and change in effectiveness, P, due to changes in PS design:	EFFECTIVENESS
• Scenarios, Wargame Models	Mission profiles with normal and contingent task demands (tree diagram)	$P = \prod_j \left[1 - U_j\left(1 - \sum_i W_{ij} \int_x F_{ij}(x) g_j(x) dx\right)\right]$ →	P, ΔP
• Threat Analyses • Reliability Analyses • Vulnerability " • Personnel Injury • Logistics Analyses	Contingency probabilities	where U_j = probability of mission failure due to failure of the j th task W_{ij} = relative frequency with which personnel type i is assigned to task j $F_{ij}(x)$ = cumulative probability distribution of performance capability of the i th type of personnel with respect to the j th task $g_j(x)$ = probability density of demands on personnel capability by the j th task	
• Work Study • Maintenance Engineering Analyses • R/M/L Models	Probability distributions of normal and contingent task times (operating, administrative and maintenance/support)		
• Human Factors	Probability of successful performance of each task by each personnel type	2. The Life Cycle Cost Model computes change in total life cycle cost, C_t, due to changes in PS design strategy:	COST
• Military Doctrine • PMS • Human Factors	Scheduled routines and routine task assignments by personnel type	$\Delta C_t = f(\Delta C_h, \Delta C_l, \Delta C_s, \Delta_a)$	ΔC_t
• Military Doctrine • Mission Analysis • R/M/L Analysis	Priorities	where C_h = hardware costs C_p = personnel costs C_s = support costs C_a = administrative costs	
• LCC Data	Manhour and associated support costs by type (rating); other costs as required		
• Military Doctrine • Results of Special Studies	Administrative policies		

The data requirements of the model have been listed by the developer as follows:

1. Information required to construct the Personnel Subsystem Effectiveness (PSE) Model:
   ... Relation of tasks to missions (performance specifications, scenarios)
   ... Subtask breakdown of each task (using detailed task analyses)
   ... Significant contingencies and alternatives at both task and subtask levels (operational/mission data)

2. Data required to exercise the PSE Model:
   ... Distribution of time to perform each subtask (using time lines)
   ... Time constraint (if any) on each subtask or collection of subtasks
   ... Distribution of performance levels achievable on each subtask or probability of error (whichever is applicable)
   ... Performance level required on each subtask (required by operational requirements or system/subsystem specifications)
   ... Contingency probabilities or frequency of action alternatives (conditional probabilities of each branch leading from each branching point)" (Ref. 1, p. 12).

A number of preliminary formats for accumulating PSE data have been suggested by the developer. "... The data bank would include a task/event sequence network (ESN) together with success criteria, operational requirements and branching probabilities... Equations would also be developed for each task or subtask network..." (Ref. 1, p. 15). It is desirable to determine the time and performance level distributions associated with each subtask. This avoids the necessity for converting the input data into individual probabilities of the .9999 type, but apparently enables the input data to function as probabilities. Each moment is referenced to (derived from, presumably) a graphical depiction of the distribution, such as

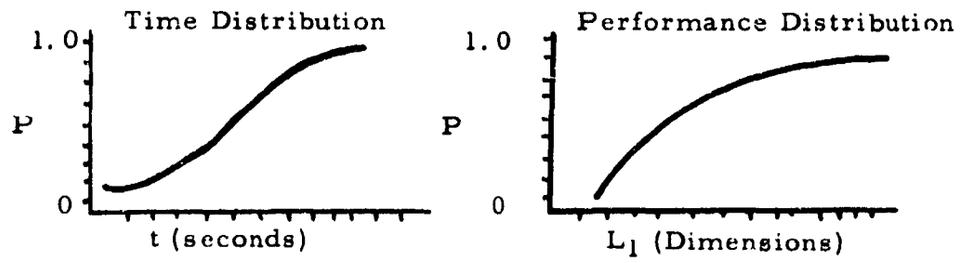


FIGURE 16. EXAMPLES OF THE REFERENCE DISTRIBUTIONS

Obviously the input data must be complete enough that a graphical function can be derived. Again this imposes a severe strain on input data requirements.

"In some cases, performance level is inappropriate or unobtainable and a probability of error (P_e) can be estimated. For each data set, the assumptions regarding personnel, hardware (including environment and facilities) and software are given by coded references to appropriate system documentation." (Ref. 1, p. 15).

2. Data Sources

Obviously to exercise the model to its full capabilities very complete system documentation and masses of input data are needed. This requirement is lightened only slightly by the fact that raw data can be utilized as inputs, without transforming each datum into a probability. It is therefore likely that without a very extensive data collection effort prior to the application of the model, the latter cannot be fully exercised.

Like other models, external reference data will be secured from any available source. "Initially, data can be obtained from existing sources and structured judgments. Sessions involving multiple judges qualified by operational or design experience as appropriate can be used to obtain judgment data. Where a series of judgments are to be made, the use of paired comparison procedures should be considered together with special techniques for detecting and correcting for systematic bias. In such cases, intra- and inter-judge reliability should be reported... In critical problem areas, experimental procedures for estimating human performance should be considered, including trial runs in full-scale mockups and live simulations. Live simulations are generally used only for areas of critical human performance in which available data are inadequate, structured judgments are insufficiently credible, and hardware, software or personnel design problems have been identified." (Ref. 1, p. 14)... "Human capability data (from previous experience, from the literature, from judgments or derived experimentally) are used to estimate task performance levels and times at the gross level or at the detailed, microstructure level of decisions and actions as required. Branch points defining contingencies or alternatives are estimated, either from human factors, system reliability or mission data" (Ref. 1, p. 22).

The need for extensive data inputs may serve as a possible deterrent to the use of the model, where system development requirements are very

time-demanding, except on major system development projects where explicit requirements for PSE modeling are recognized from the start of the project. As is the case with other models, it would be useful to have some estimates as to the cost in manpower and time to gather the necessary data and exercise the model. Conceivably one could build a generalized data bank and store it in the computer's memory, but the necessity for developing a mission profile and associated requirements may be costly in time.

The PSE model does not, however, impose any requirements for data different from those required by comparable simulation models.

3. Data Output

The model is similar to other methods in that its primary data output is a probability of task success which "is the joint probability (P_{Lt}) that a task will be performed at the level L or greater and in time t or less, where the level L and the time t are defined by the demands of the mission" (Ref. 1, p. 9). The model therefore emphasizes both the time and the performance level aspects of human performance, in contrast to other models which emphasize one or the other.

"In some instances, either performance level or performance time will be the single criterion of task success, or performance level will need to be measured along multiple dimensions which may be either interdependent or independent. Techniques exist for treating either case. In most instances, the criterion for task success can be reduced to a relatively simple bidimensional one made up of time and performance level. Required task time may be in terms of clock time (e.g., must be completed within 10 seconds) or it may be referenced to an external event (e.g., must be completed before occurrence of event "a") (Ref. 1, p. 9).

Probability of task or mission success is not, however, the only output that can be derived from exercising the model. Because the computer can print out each successive model operation, it is possible to obtain the time to perform any collection of subtasks up to the total task. As we have seen with other computerized models, this is an extremely valuable aspect, because it enables the model user to diagnose those points in the mission profile where personnel cannot satisfy demands.

Since an overall system requirement is available (it must be, otherwise it would be impossible to develop individual task/subtask requirements), it is also possible to determine whether the system requirement can be achieved by the personnel subsystem. Thus, "the PSE model, suitably programmed on a computer, is used to compute "achieved" performance times and levels at a level of definition consistent with the derived "requirements". "Achieved" and "required" are compared and probability of task success is computed... If the derived system/mission task requirements are met, the allocation is complete and forms the basis for training standards and for test measures to be used during Personnel Subsystem Test and Evaluation (PSTE). If the requirements are not met, a search is made for a compatible reallocation or special studies are made to refine the human capability data" (Ref. 1, p. 22).

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

Although a variety of analyses are listed in Table 12, it is apparent that the basic analytic tool, as with the other models, is task analysis. As indicated in connection with input data, this analysis requires development of a task/event sequence network (ESN) which is a form of block or flow diagram. This may or may not be the same as the tree diagram referred to in Table 12. The similarity to Swain's probability tree and Blanchard's GSSM and MSSM should be noted.

2. Method of Synthesis

Although we have indicated that PSE is a simulation model, it requires combination of probabilities in some manner the same way that the non-simulation methods do. This may result from the fact that probability of success is based on two criteria, i.e., performance level and time, which means that the probability of successful performance on each criterion is determined separately and the results multiplied.

In order to understand how the model operates conceptually (we shall deal later with its functional operations), it is necessary to examine the logic of its approach.

"Consider a situation in which the personnel capability demands of the jth task, job or position are distributed in accordance with the function

$g_j(x)$, i.e., the probability density of demands on personnel capability, where x represents the performance parameter or vector. Further suppose that the distribution of personnel capability of the ith personnel type for the jth task is $F_{ij}(x)$, i.e., the cumulative distribution of personnel capability. Then the probability of successful performance of the jth task by the ith personnel type is

$$P_{ij} = \int_x F_{ij}(x)\, g_j(x)\, dx \qquad (1)$$

Equation (1) states that the jth task will be successfully performed if the personnel performance capability of the ith type is equal to or greater than the mission demand on personnel capability. Unavailability is included since a performance capability of zero is assigned to unavailable personnel.

"Since the probability, Pj, of successful performance of the jth task is across all available personnel types i = 1, 2, ... n, then

$$P_j = \sum_{i=1}^{n} P_{ij}\, W_{ij} \qquad (2)$$

where W_{ij} is the relative frequency with which personnel type i is assigned to task j.

"When the variable x in equation (1) is one-dimensional, the factors in the equation are ordinary distribution and density functions represented by well-known types of formulae. When there is more than one performance requirement or criterion associated with a given task and the variable x is a vector quantity, the distribution and density functions involved are less well-known and in the general case, less tractable.

"Suppose, for example, that successful performance of a task requires an operator to perform at certain minimum levels on each of two criteria, one of which is a time requirement. Then

$$P_{ij} = \int_x \int_t F_{ij}(x, t)\, g_j(x, t)\, dx dt \qquad (3)$$

In the case in which the variables x and t are independent, we have by definition

$$F_{ij}(x, t) = F_{ij}(x)\, F_{ij}(t)$$

and

$$g_j(x,t) = g_j(x)\, g_j(t).$$

Therefore

$$P_{ij} = \int_x F_{ij}(x)\, g_j(x)dx \int_t F_{ij}(t)\, g_j(t)\, dt \qquad (4)$$

which means that the probability of successful performance on each criterion is determined separately and the results multiplied.

"Note that if $F_{ij}(x,t)$ is interpreted as the distribution of maximal performance capacity, it can usually be considered independent of $g_j(x,t)$, the density of mission demands on performance. Such an interpretation of $F_{ij}(x,t)$ is reasonable since we are interested in the probability that mission demands will be met or exceeded, and personnel performance under critical conditions tends to increase as necessary, depending on mission requirements, up to "maximal capacity". However, when mission demands create an environment which degrades maximal performance capacity, $F_{ij}(x,t)$ and $g_j(x,t)$ can no longer be considered independent and the formulation of P_{ij} must take into account that dependency. Fortunately, in most cases of practical interest, it suffices to treat $F_{ij}(x,t)$ and $g_j(x,t)$ as two joint distributions which are independent of each other and equation (3) is applicable.

"Now, let U_j be the conditional probability that the mission will fail if the jth task is failed (not completed satisfactorily). Since $1-P_j$ is the probability that the jth task is failed, then the probability of mission failure due to the jth task is

$$U_j(1-P_j)$$

and the probability of mission success considering the jth task is

$$P_j = 1-U_j(1-P_j) \qquad (5)$$

"Personnel subsystem effectiveness is simply the overall probability of mission success considering all independent tasks j = 1, 2, ... m

$$P = \prod_{j=1}^{m} P_j = \prod_{j=1}^{m} \left[1-U_j(1-P_j)\right] = \prod_{j=1}^{m} \left[1-U_j\left(1-\sum_{i=1}^{n} W_{ij} \int_x F_{ij}(x)\, g_j(x)\, dx\right)\right] \qquad (6)$$

where $x = (x_1, \ldots x_n)$ for N performance dimensions" (Ref. 1, pp. 3-5).

This last equation requires the determination of availability and appropriateness of each of the personnel types i = 1, 2 ... n. This requires the model to have the capability of storing tables of tasks vs. personnel types and their performance level and task time distributions. The mission model, as indicated previously, defines the performance level and time requirement for each task. P_{ij} is determined by sampling from the performance level/time distributions for the ith personnel type and comparing the sampled values with the values demanded by the mission. This is done by the computer "at convenient increments of time during the mission" to determine whether the predetermined demands have occurred, continued or ended.

"If a demand occurs, the personnel-available list is searched for available personnel and the one with the highest P_{ij} is assigned (and deleted from the list). If a demand is continued, the list is unchanged. If a demand is ended, the personnel are added to the list. If a demand occurs but no personnel with $P_{ij}>0$ are available, then the computer determines whether a mission failure occurs by solving the appropriate function of U_j or by sampling a distribution with mean U_j. If a mission failure does not occur, the process is repeated at successive intervals until failure occurs or until appropriate personnel become available. Probability of mission failure, (1-P), is computed on a running basis, and sampling is carried out to determine whether the mission has failed or continues.

"The procedures and algorithms for generating and analyzing task sequences (including alternatives and contingencies) have already been formulated by Dunlap and Associates, Inc. Programs for combining distributions (convolution and mixing or furcation) currently exist for the Olivetti Programma 101 desk top computer and for high speed digital computer (e.g., CDC 6600 or GE 360)." (Ref. 1, pp. 6-7).

3. Level of System Description

The precise level to which input data are applied is somewhat obscure. The model is "developed down to the perception, evaluation, decision action level of detail in critical problem areas" (Ref. 1, p. 22) which suggests the subtask level. On the other hand, the requirement, e.g., time constraint, may be at the level of "a collection of subtasks" which suggests that probabilities for subtasks may have to be combined in order to make the required comparison between "achieved" and "required" performance.

ANTICIPATED MODEL USES

The uses for which the developer feels the model is effective have been in part indicated previously. In addition, "the approach described... responds to the frequently-stated requirement that

(1) critical tasks be assigned a quantitative human performance "reliability" index
(2) human performance measurements be made during PSTE for critical tasks
(3) a systematic method be provided for identifying and recording human-initiated malfunctions which can be
    ... correlated with equipment performance data to determine the inter-action of human and equipment performance
    ... converted to reliability indices which can be related to system functions (functional flows) for use in predicting system performance
(4) human performance quantification and evaluation be done on each critical task and used to determine:
    ... the contribution of each critical task to system effectiveness and reliability, and
    ... the minimal level of human performance required to meet system operating requirements." (Ref. 1, pp. 20-21).

Let us examine how well the model satisfies the potential uses we have specified for other models: prediction of system effectiveness; comparison of alternative design configurations; design analysis; manpower selection and training.

It seems reasonable to suppose that if the model lives up to expectations, it should be able to estimate future operational performance because it applies predictive input data to tasks just as other operability models do. Since the existence of a system requirement is critical to model operations, the model obviously will indicate whether or not that requirement will be fulfilled.

Similarly the model should be able to compare alternative configurations based on anticipated PSE measures. How sensitive it will be to design differences, however, depends on the equipment parameters to which the behavioral input data are related. This is as obscure in this model as it is in most models. Insofar as the designs are based on different procedures, the model should be able to evaluate alternative procedures.

In the same way, the model should be useful in design analysis- because of the time-historical record it provides- if the behavioral input data are linked specifically to equipment design parameters, such that input probabilities of success for task N vary as a function of different design factors. There is no implication in Reference 1 that this is so except that hardware data (of an unspecified type) are mentioned as one of the model inputs.

The model specifically requires manning inputs; consequently it should be able to supply information on the effectiveness of different manning configurations. It is much less likely that the results of a model exercise will suggest the particular aptitudes for which system personnel should be selected, but then the model does not say it can.

It will be recalled that the model is supposed to "assess the adequacy of planned training in terms of impact on system effectiveness". It is difficult for the author to see in this statement anything but the usual tendency of model developers to inflate the range of capability of their models. It is difficult to see how one could include a planned training program as an input to the model unless one hypothesized a given level of effectiveness of that training and adjusted input probabilities and task times correspondingly.

Similarly one sees absolutely no way in which human-initiated malfunctions can be identified by exercising this model, unless an effect factor, like Swain's F_i, were included; however, the model logic does not differentiate between errors that have an equipment consequence and those that do not.

This model does have one capability that perhaps the others do not. It could be used as a test and evaluation tool (during the PSTE phase) by serving as a framework for gathering PSTE data, and then using these data as inputs to the model to secure a total system figure of merit. And it is true that the model could be used as a data base to the extent that input data are stored in the computer's memory.

In the light of the model's half developed state (and it can be considered only a skeleton at present), the claims that are made for it should be viewed with caution.

VALIDATION/APPLICATION STUDIES

According to the developer the method has neither been validated nor applied. No supporting data are supplied in the basic reference. However, as was pointed out previously, procedures and algorithms have already been formulated.

Even with this, however, it would seem to us to be difficult to apply the model to an actual system development project because so many of the model details remain obscure. This is unfortunate because for systems large enough to afford personnel subsystems projects, the model (assuming it works) would seem to be ideal. In consequence, it is impossible to view this model as having anything more than potential.

EVALUATION SUMMARY

Validity - No data available; model has never been applied.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Model is not limited to specific types of equipment, tasks, behaviors. One potentially important limiting factor may be that quantitative system requirements must be specified in detail.

B. Applicability: Model outputs a prediction of system effectiveness and can be used for design analysis and manning.

C. Timing: Model can be applied to systems at all stages of system development, depending on how concretely system details have been conceptualized.

Model Characteristics

A. Objectivity: It is hypothesized (from other writings of the developer) that the model leans heavily on expert estimates, but specific procedures for securing these are not indicated.

B. Structure: Model structure is not complete. A number of basic parameters are not explicitly defined.

REFERENCES

1. Hanifan, D. T. Human Performance Quantification in System Development: Recent Advances. Final Report, Contract N62462-70-M-0733, performed for Naval Applied Science Laboratory, New York, by Dunlap and Associates, Inc., Santa Monica, Calif. February 1970 (revised July 1970).

C. MAINTAINABILITY PREDICTION MODELS

1. ELEMENTARY RELIABILITY UNIT PARAMETER TECHNIQUE (ERUPT)

INTRODUCTORY DESCRIPTION

ERUPT is a method for inferring two measures of maintenance technician performance as part of a model describing weapon system readiness (availability). The method does not require reporting of human-initiated malfunctions but makes use of failure and maintenance data from operations. The two human performance parameters inferred are:

A. The probability that a failure is detected and repaired during maintenance.

B. The probability that maintenance does not induce failure.

The model consists of equations and computational routines that enable one to solve for A and B on the basis of known values of failure and maintenance inputs.

GOALS

The developers of this technique (see refs. 1 and 2) speak of it as an "indirect" approach, because it was deliberately derived to avoid the necessity of collecting data on human-initiated failures. After a review of all the failure/maintenance reporting systems in the Navy as of 1967 (e.g., 3M), they decided that these systems provided no usable data relative to human reliability (HR). Therefore they sought to develop a technique in which HR parameters could be inferred from more readily available indices.

(It might be noted incidentally that studies performed by the author and others found essentially the same negative situation with regard to Air Force reporting systems.)

This introductory explanation seems necessary because the technique was developed essentially to provide information on certain HR parameters and only indirectly to answer certain specific prediction questions. For example, "The most significant feature of the technique is that the quantification of these human performance parameters can be accomplished by using equipment failure and maintenance data without relying on human-initiated failure reporting" (p. 56, ref. 1).

ERUPT is part of a model which evaluates system readiness. From the standpoint of HR, therefore, the technique has the primary goal of estimating two human performance parameters:

(1) α = the probability that failure is detected and repaired during maintenance;

(2) β = the probability that maintenance does not induce failure.

Instead of using the Greek symbols for the alpha and beta above, we shall, for convenience's sake, refer to these parameters henceforth as A and B. Actually we are interested in 1.0-A and 1.0-B, because these values indicate deficiencies in technician performance that suggest the need for remedial action.

A and B can be considered as estimates of the efficiency of maintenance technician performance. If the technician detects a malfunction and remedies it (A) and if his actions do not result in additional maintenance requirements (B), this essentially summarizes his efficiency. If these parameters could be estimated with reasonable precision, it might be unnecessary to make use of methods that attempt to measure maintenance behaviors directly, like the personnel reliability index (already discussed) or the techniques used by the author and his colleagues (ref. 4).

The question that needs to be answered about this technique is whether the method can be used to perform the various functions a predictive technique should perform: (a) to estimate during design/development the effectiveness that the human component of the system will assume during operations; (b) to permit comparisons of alternative design configurations in terms of the effectiveness of that human component; (c) to suggest initial design and redesign possibilities; (d) to provide information relative to manpower selection and training. We shall consider these points later.

ERUPT is part of or can be used to develop a measure of what the developers call "readiness reliability". This is defined as the "probability that the weapon is operable at the time of its operating mission or, more generally, probability that the weapon is in "go" condition when it is needed" (p. 56, ref. 1). The reader will recognize immediately that what is being described here is more commonly known as system availability.

Presumably the model will also predict something called "mission-tactic reliability" - the "probability that the weapon will successfully carry out a given mission with a prescribed tactic, assuming the weapon is ready (operable) at the beginning of the mission" (p. 56, ref. 1). This is more commonly understood as system reliability. However, only readiness reliability was considered in the research effort, and it is doubtful whether the technique can in fact handle mission-tactic reliability, since, as we shall see later, none of the input data relate to operator performance. Hence, if the technique were used for estimating system reliability, it could do so only for systems so completely automated that no operator interface (or hardly any) is needed.

Conceivably what the developers had in mind in referring to mission-tactic reliability was that the general strategy of solving for A and B by inputting known values for other parameters could be applied where A and

B were operator performance parameters rather than maintenance technician parameters. A and B would then have to be redefined in terms of operator performance, as would the other model parameters. What this redefinition would be is unclear and in any event need not concern us further because it is completely speculative. One can view this as another example of the tendency of developers to expand the scope of their techniques beyond what they were intended to do.

DEFINITIONS AND ASSUMPTIONS

In the course of the preceding discussion we defined the two human performance parameters with which the model is concerned and the two measures of system effectiveness it seeks to derive. Definitions will be supplied of the input data and output metric when these are discussed later. It is, however, desirable to define what is meant by elementary reliability unit (ERU) which apparently plays a major part in the model because it defines the level at which input data are gathered. It would be best to permit the developers to speak for themselves: "One of the most important concepts associated with the application of ERUPT is the grouping of system components into ERUs. The selection of ERUs is based on the maintenance level established for the system. Maintenance level in this context is defined as the lowest type of equipment indenture at which maintenance is performed." (p. 57, ref. 1).

To translate into more commonly recognised terminology, ERU is the same as the lowest level replaceable unit. This is the level of indenture at which equipments or components will be maintained and repaired rather than being replaced. In other words, if maintenance involves only removal and replacement of hermetically sealed cannisters, then the cannister level is the ERU and the inherent reliability (π) used in the model refers to cannisters. If, on the other hand, technicians maintain at the circuit level, then π is applied at the circuit (ERU) level.

On to assumptions. The first assumption is that "the scope of the research (on the technique) is limited to human reliability as related to the operational and maintenance phases of the weapon systems life cycle" (p. v, ref. 1). From this we infer that the methodology is not usable during the design/development stages of system life. Later, when we examine required input data, we shall examine this inference in more detail.

A second assumption of great importance is expressed by "The general thought is that, if "inherent" or "true" hardware component failure rates can be determined under controlled conditions, then the expected reliability of the ERUs can be calculated. Then, based on failure rates of these ERUs under actual conditions obtained from maintenance reports, it will be possible to infer the α's and β's from the differences" (p. 59, ref. 1).

To express the concept somewhat differently, it is assumed that the hardware components of a system have an inherent, true reliability (i.e., unsullied by operational malpractice or human hands). If this value is known, and if one then can determine operational failure rates, and can calculate operational reliability, then the difference between the inherent and operational reliabilities must have been caused by the only other factor influencing operational reliability, i.e., maintenance actions.

Several points must be noted about this assumption. (1) The assumption implies that inherent reliability is derived from system test data gathered under controlled conditions. Estimates of inherent reliability have generally been grossly overestimated. For example, it is known that reliability predictions for the electronic equipment on the F-111 (in terms of MTBF) were at least several times lower than actual (operational) MTBF for that equipment. The natural tendency of contractors to represent their equipment in the best possible light leads to highly optimistic estimates of reliability during development. It might be noted incidentally that many reliability engineers are quite skeptical about the existence of an "inherent" reliability. In any event, should inherent reliability be overestimated, it is inevitable that A and B will be overly pessimistic.

(2) Even assuming that inherent reliability were estimated on the basis of system tests conducted during development, it is well known that such system tests rarely resemble operational usage conditions, so that the inherent reliability values derives will again be distorted, with undesirable consequences for A and B.

(3) Since A and B alone are considered the cause of the difference between inherent and operational reliability, no attention is paid to other factors that influence maintenance-induced failure, e.g., inadequate logistics, poor job aids, etc. Or rather A and B subsume these other variables. In consequence the human component of the system (A and B) is given the discredit for non-human factors that also tend to produce maintenance-induced failures.

It may, however, be argued that since inadequate logistics, poor job aids, etc. influence the system only through the maintenance technician, A and B do in fact represent these other non-behavioral factors.

(4) It must be noted that even in these "enlightened" days, many systems are developed without the benefit of reliability predictions. This is not an objection to the ERUPT methodology, since all systems should have such reliability predictions; it is merely a cautionary note.

Perhaps because the methodology was developed on the basis of operations research concepts and specifically sought to avoid the need for behavioral data, the technique makes no explicit assumptions of a behavioral nature. However, the model makes <u>implicit</u> behavioral assumptions which

should be clarified. For example, one implicit assumption is that the probability of maintenance-induced failures increases with the number of opportunities to perform preventive and corrective maintenance operations. This can be inferred from the fact that among the input data is the number of maintenances before repair/replacement of the ERU or since last repair. Since A and B are derived from an equation which includes such parameters as maintenance opportunities, time between exercises and tests, number of ERU tests, number of corrective maintenances, etc. we might expect the developers to indicate how these factors influence A and B. They do not, however. The equation describes system availability and by inserting input values for all parameters but A and B, and then solving for A and B, the human parameters are inferred. However, we do not know from the model what causes A and B; we merely extract a value for these. This is not necessarily an objection to the model, if the values derived for A and B are correct; but it represents an inadequacy.

One might expect A and B to be influenced by such factors as the complexity of the equipment being maintained and the technician's skill level. ERUPT recognizes the influence of skill level by relating it to the level of indenture at which the equipment is maintained, but does not include skill level in the equation for deriving A and B. Nor does it include equipment complexity in that equation. It might, however, be argued that complexity is assumed in the level of indenture for the ERU, but such an indirect relationship is weak at best.

The consequences of not including factors such as equipment complexity and skill level in the equation for A and B means that it is difficult to interpret estimates of A and B in terms of design and personnel implications. We shall have more to say about this in discussing the uses of the technique.

METHODOLOGICAL SCOPE

"It should be emphasized that the equations developed for readiness reliability are applicable to any weapon system which meets the underlying assumptions of the model. However, it is also true that the readiness reliability measure of effectiveness is not the only measure that can accept ERUPT human reliability parameters; they can be incorporated equally well into other models of system reliability" (p. 59, ref. 1).

Attention has already been drawn to the fact that the model predicts availability only. From that standpoint the first statement in the preceding paragraph appears correct. However, the assumption that values for A and B, once derived, using ERUPT, can be incorporated in other models, appears unjustified. Since A and B are derived from other parameters like inherent reliability, number of maintenance actions, etc. any system making use of A and B must include among its parameters the same parameters that ERUPT includes; otherwise the significance of A and B will be distorted.

The model therefore is specialized to derive an availability estimate and an estimate of maintenance technician performance. It is not limited to any particular type of system or any particular kind of maintenance. It will not, however, predict system reliability during missions or operator reliability.

MODEL PARAMETERS

"The model consists essentially of equations and computational routines..." (p. 56, ref. 1). In this section we will define and discuss the elements of the model equations. We shall not discuss the computational routines, assuming these to be adequate for the purpose for which they were developed.

The equations include the following terms:

(1) α_i = probability that failure of the i^{th} ERU, if it exists, is detected and repaired during the first maintenance following the failure.

(2) β_i = probability that maintenance does not induce failure in the i^{th} ERU given that the ERU is in nonfailed condition at the time maintenance is initiated.

(3) π_{ij} = probability that i^{th} ERU, which is in nonfailed condition at the time the j^{th} exercise or test is initiated, survives the exercise or test.

(4) t_i = storage time between exercises or tests of i^{th} ERU.

(5) $1-G_i(x)$ = probability that i^{th} ERU survives a storage time x given that it was in "new" condition at beginning of storage, i.e., at zero storage time. This is determined from a theoretical distribution estimated under laboratory or test conditions.

(6) NK = number of different integral values of k in the sample of corrective maintenances.

(7) k_i = number of maintenances before repair/replacement since beginning of storage or since last repair of the ERU if previous corrective maintenance has been done (including maintenance when last repair/replacement was done). A sample of k_i's is required from actual experience to derive maximum likelihood estimates of α and β.

$$(i = 1, 2, \ldots\ldots NK).$$

(8) N_i = frequency of the i^{th} value of k_i in the sample of corrective maintenances (i = 1, 2, ... NK). Size of sample is equal to $\sum_{i=2}^{NK} N_i$.

(9) τ_i = storage time of i^{th} ERU between time when last exercised and maintained to time of operating mission.

(10) ν_i = number of exercises or tests of ERU before operating mission.

(11) $P_j(\nu_i)$ = probability that the i^{th} ERU survives the i^{th} exercise or test and was last repaired during the $(\nu_i - j)$ maintenance period.

The meaning of most of the elements of the model is apparent. However, the precise significance of a few elements is doubtful. We assume that π_{ij} represents the inherent reliability of the ERU. In effect, if the ERU is exercised, it will perform without malfunction. $1-G_i(x)$ is also presumably a form of inherent reliability, except that it represents reliability during "shelf" or "storage" life.

With regard to NK and k_i, an example is provided. "Assume that a sample of ERU corrective maintenances shows that they have occurred on the first, third, third, second and first preventive maintenances. Thus, under the definitions,

NK = 3 $K_1 = 1$ $N_1 = 2$

$K_2 = 2$ $N_2 = 1$

$K_3 = 3$ $N_3 = 2$ " (p. A-3, ref. 1).

The parameters involved in the model are therefore:

(1) inherent reliability
(2) storage time
(3) shelf life reliability
(4) number of corrective maintenances
(5) number of preventive maintenances
(6) number of equipment tests or exercises

and of course

(7) malfunction detection/correction probability (A)
(8) maintenance-induced failure probability (B)

which are derived from 1-6.

DATA

1. Measures Used

These are the input data which were discussed in the section on model parameters.

2. Data Sources

"The method for estimation of values for α and β parameters for an ERU is based on actual shipboard failure experience and maintenance data for the ERU" (p. 59, ref. 1). This appears to be in contradiction to the assumption we made that π_{ij} represents inherent or true hardware reliability which is ordinarily estimated during system development. It is difficult to see how one could derive a measure of inherent hardware reliability from operational data, because operational data represents the influence of operational factors acting on that inherent reliability. In this connection on page 58 of ref. 1 it is asserted that the "probability that the ERU which is in a non-failed condition at the time a test is initiated, survives the test" is derived from laboratory test data, rather than from shipboard data. There is an obscurity there that needs explanation.

Another apparent contradiction is that G (xt) is supposedly derived "from a theoretical distribution estimated under laboratory or test conditions" (p. A-1, ref. 1). If the distribution is derived from laboratory conditions, it is obviously not gathered from shipboard data. This ambiguity also needs explication.

Even if one assumes that all the data needed for the model could in fact be derived from shipboard, the manner in which those data would be gathered requires elaboration. Ideally the various reporting systems aboard ship should be able to provide the data inputs from which the model parameters could be calculated, but these would have to be analyzed and transformed into the individual input measures required. For example, π_{ij}, even if it could be derived from shipboard data, would not be represented on the reporting form in this manner.

It appears then that there is considerable obscurity concerning the manner in which input data could be secured. It is regrettable that the test case described in reference 1 did not utilize data from an operating weapon system. If it did, some of the problems relative to data sources could have been clarified.

3. Output Measures

The model outputs A and B, of course. It provides a series of F values (probability that a set of A and B values will produce the corrective maintenance experience reflected in the result of the input data) for each of the values of A and B, in increments of .01. The various

output measures are:

(1) A

(2) B

(3) Probability that this sample of corrective maintenances could have occurred given this set of A and B values.

(4) The maximum probability that this sample of corrective maintenance could have occurred.

(5) Value of A at which (4) occurs.

(6) Value of B at which (4) occurs.

The preceding output measures dealt only with A and B. Another output is a prediction of the readiness reliability of the i^{th} ERU. Given this value, the readiness reliability values of the individual ERUs are multiplied to provide a measure of system readiness reliability.

Note that the multiplicative relationship inherent in the system readiness reliability assumes that all ERUs are needed for reliability; in other words, the equation does not take into account the possible effects of any redundancy. Whether this assumption is justifiable is left up to the reader's judgment.

It should also be noted that if system readiness reliability (availability) is calculated on the basis of shipboard (operational) data, there are simpler ways of measuring this directly. Availability can be simply defined as $\frac{\text{total uptime}}{\text{total time}}$. Given that one is predicting availability from operational data, availability determination for either the ERU or the system as a whole is relatively simple, requiring only the determination of the proportion of total time that a given unit has been down (as a result of malfunction and maintenance).

PROCEDURES FOR MODEL APPLICATION

1. Method of Analysis

The initial step in the development of the equation is the selection of ERUs. This is required in order to identify the components for which data will be collected. The developers indicate that this selection process poses no problem for anyone familiar with the system, and this is probably correct, because this type of analysis is routinely performed by reliability and maintainability personnel. For this reason, perhaps, no detailed description of the process is provided in reference 1. In order to make such a selection it is necessary to know the level of maintenance for the

system being evaluated. Presumably this is the same level at which maintenance data are reported routinely aboard ship. However, it is necessary to ensure that the ERU level selected is in accordance with reporting procedures; otherwise data will not be available.

If laboratory or test data secured during development are to be utilized, it will be necessary to ensure that such data matches the ERU level previously selected.

It is noteworthy that no behavioral analysis is required by the methodology, in contrast with the other methods reviewed.

2. Method of Synthesis

π is an average reliability which is applied to each ERU. This means that although each ERU probably has a different probability value, the average of these probabilities is applied to all ERUs being modeled.

$P_k(r)$ is the probability that corrective maintenance takes place on the k^{th} maintenance.

$P_x(f)$ is the probability that a failure occurs before the x^{th} maintenance but not before the $(x-i)^{th}$ maintenance. In consequence

$$P_1(f) = 1 - \left[1 - G(t)\right] \pi$$

This can be expanded as follows:

$$P_x(f) = \beta^{x-2} \left[1 - G(x-1)t\right] \pi^{x-1} - \beta^{x-1} \left[1 - G(xt)\right] \pi^x \text{ for } x \geq 2.$$

Assuming failure before the first maintenance, the probability that the corrective maintenance will take place on the fourth maintenance is

$$P_1(f) \cdot (1-\alpha)^3 \alpha = 1 - \left[-G(t)\right] \pi \left(1-\alpha^3\right) \alpha$$

Without going through the intervening steps (for which see ref. 1), the probability that corrective maintenance has occurred on the k^{th} maintenance regardless of when the failure has occurred is

$$P_k(r) = \sum_{x=1}^{k} \left\{ \beta^{x-2} \left[1 - G(x-1)t\right] \pi^{x-1} - \beta^{x-1} \left[1 - G(xt)\right] \pi^x \right\} (1-\alpha)^{k-x} \alpha +$$
$$\left\{ 1 - \left[1 - G(t)\right] \pi \right\} (1-\alpha)^{k-1} \alpha$$

The calculation for A and B involves the following expansion of the preceding equations:

$$\text{MAX} \prod_{i=1}^{NK} \left[\sum_{x=2}^{k_i} \left\{ \beta^{x-2} \left[1 - G(x-1)t\right] \pi^{x-1} - \beta^{x-1} \left[1 - G(xt)\right] \pi^x \right\} (1-\alpha)^{k_i - x} \alpha \right.$$
$$\left. + \left\{ 1 - \left[1 - G(t)\right] \pi \right\} (1-\alpha)^{k_i - 1} a \right]^{N_i}$$
$$\text{for } \left[\alpha, \beta \mid 0 < \alpha < 1,\ 0 < \beta \leq 1\right]$$

Since the complexity of the equation makes manual computation impractical, a computer program to perform the calculations is available. A computer search is initiated to define the set of values for A and B which maximize the probability that a specified distribution of failure probabilities (G(xt)) will occur. The computer program computes a probability (F value) that a set of A and B values will produce the corrective maintenance experience reflected in G(xt).

Since we are only indirectly concerned with the derivation of the equations for estimating system readiness, we shall not present them here. A complete exposition is presented on pp. A-4 and A-5 of ref. 1.

3. Model Output

These have already been discussed.

ANTICIPATED MODEL USES

1. Prediction of System Effectiveness

It has already been pointed out that the model derives A and B which can be used as estimates of the efficiency of the maintenance technician. A predicts the efficiency with which malfunctions are diagnosed and remedied; B predicts the liklihood that maintenance-induced failures will not occur during preventive maintenance. Ideally therefore both A and B should be high.

Interestingly enough, the model does not deal with a situation in which a second malfunction is induced by the technician during corrective maintenance of an original malfunction. This is actually a more common situation than inducing a malfunction during preventive maintenance. The reason is simply that many systems do not permit preventive maintenance or else preventive maintenance is very severly restricted. Of the two parameters estimated, therefore, A would seem to be more valuable.

On the other hand, the value for A should usually be quite high. Unless the system is extraordinarily complex or very poorly designed, malfunctions will be detected and repaired, although mean time to restore (MTTR) may be quite high. Hence A should always approach 1.00. It is difficult to understand therefore why the illustrative example in ref. 1 has A values down to .16. In the case of this low value, the developers talk of malfunctions not being detected promptly during preventive maintenance. It is possible that A should be interpreted as probability of detection and repair of malfunctions encountered solely during preventive maintenance. However, if this is the case, the objection raised to B applies also to A. In any event, there is an obscurity here that needs clarification.

The model also outputs an estimate of system readiness reliability (availability).

2. Design Analysis

We are concerned here with the way in which the derived A and B values can be used either for comparing alternative design configurations or for initial design/redesign suggestions. In other words, what can one do with A and B?

Because the data inputs to the model are operational, it would seem that the model would have no use in the design/development phase and consequently could have no use for design analysis. If one were fanciful, however, one could conceive of applying the model during design with estimated parameters. In other words, based on experience with previous systems of the same sort as the one being evaluated, derive estimates of π_{ij}, t_i, $1-G_i(x)$, etc. and work the equation through. This would provide an estimate of A and B which would obviously be less precise than one derived from operational data, but still perhaps usable. One could then predict maintenance technician efficiency for a system still in the developmental stage.

It is unlikely, however, that one would compare design configurations in this way, however, because it is not at all clear whether the input parameters to the model are sensitive to differences in equipment design. The operational data would undoubtedly be sensitive to these differences, but the estimated input data would not be, because the model does not indicate any relationship of input parameters to design. By this we mean that in actual operations design A will have a different π, a different storage life, number of corrective maintenances, etc. than would design B. But in the absence of operational data and without design-input parameter relationships specified, we could not assign different estimated input values to the different configurations.

Conceivably the model could have value as a redesign tool. Given a low A and B in actual operations, what should one do to correct the situation? Again, however, the absence of design relationships with model parameters would leave us without any way of interpreting A and B in terms of these relationships. In this respect, this model is no different from the others. One could of course look at a ERU with a low A and B and try to infer the characteristics that led to the low A and B, but how successful one would be is anyone's guess.

3. Manpower Selection

Not applicable.

4. Training

A low A and/or B could be interpreted as meaning that maintenance technicians need more training; but equally well such a low A and/or B could mean that the complexity of the maintenance required is too much

for the technician. In any event, the model would do nothing to suggest how much or what type of training would be needed during development.

VALIDATION/APPLICATION STUDIES

The proof of the pudding, it is said, is in the eating. This means that although the model may be excellent for the limited uses to which it may be put - the estimation of A and B - one cannot be sure until a validation study has been performed. Like the other models reviewed (except the digital simulation model), this one has not been validated, or at least we have no information on validation.

Validation here would require an independent estimate of A and B derived from actual shipboard failure and maintenance data. Even this would not be a completely satisfactory test of the model because it would not test A and B directly, but only A and B as inferred from other operational parameters. Only measurements of maintenance-induced failures (as performed by the author and his colleagues in ref. 3) would indicate whether model estimates of A and B were correct. The difficulty here is that the model was developed in the first place only because the authors of ref. 1 felt that they could not measure these HR parameters directly. If one accepted their proposition (which the author does not), no true validation of the model would be possible.

One other possibility exists: to use the model to estimate system readiness reliability and then check on alternative ways of measuring availability. Since A and B are required for estimating system readiness reliability, if the two different ways of estimating readiness reliability coincided (within reasonable limits), one could infer the correctness of the model and hence of A and B.

As pointed out above, no validation study is reported in reference 1. An application test of the model was performed with assumed data and the derived A and B values made what appeared to be sense to the developers. The results indicated that the HR parameters were sensitive to the number of preventive maintenance actions preceding each corrective maintenance action. However, this cannot be considered validation, since the absolute values achieved might have been highly erroneous.

A sensitivity analysis of A and B was also performed on an actual weapon system. This indicated that readiness reliability is quite sensitive to human maintenance parameters.

What then can we say about the model? Its primary advantage is that it does estimate A and B, and no other model so far reviewed does so. Its greatest deficiency is that it cannot (apparently) be applied during the design/development phase.

EVALUATION SUMMARY

Validity - No data available; model never applied.

Reliability - No data available; model never applied.

System Development Applicability

A. Comprehensiveness: Method highly limited.

B. Applicability: Model does not predict future performance but estimates system availability on the basis of operational data.

C. Timing: Useful only with operational systems.

Model Characteristics

A. Objectivity: No subjective estimates required.

B. Structure: Ill defined; model parameters obscure.

REFERENCES

1. Beek, C., Haynam, K. and Markisohn, G. Human Reliability Research, Report PRR-67-2, New Developments Research Branch, BuPers, Sept. 1967.

2. Markisohn, G. Human Initiated Malfunctions. Proceedings of the Symposium on Human Performance Quantification in Systems Effectiveness, Washington, D.C., Jan. 17-18, 1967.

3. Meister, D. et al. The Effect of Operator Performance Variables on Airborne Equipment Reliability. Report RADC-TR-70-140, Rome Air Development Center, Griffiss AFB, New York, July 1970.

II. MIL-HDBK 472

MAINTAINABILITY PREDICTION METHODS

INTRODUCTION

The four methods described in this section are taken from MIL HDBK 472 (Ref. 3). They do not include all available prediction methods and this section does not deal with these other methods because to do so would require a far more extensive examination of the maintainability prediction area than we can provide. The reader who is particularly interested in this field would do well to consult Rigney and Bond (1966, Ref. 5) and Smith et al. (1970, Ref. 6).

Examination of the methods included in Reference 3 should be adequate for our purposes because (1) they represent the methods accepted by the great majority of maintainability engineers, since inclusion of these methods in a military handbook gives them at least semi-official status; (2) the research on which these methods are based has been the most outstanding of the recent past; (3) our examination of maintainability prediction methods is intended only to be representative of these methods, not to be exhaustive.

This last point is most important. Our goal in this report is to review methods that estimate the efficiency of the technician's performance, as do ERUPT and the personnel reliability index. The maintainability prediction methods discussed in this section attempt to predict system down time which is something quite different. Because maintenance task completion times are at the core of all these methods, their output measure is obviously strongly influenced by the technician's effectiveness (along with other factors such as design and logistics) that results in shorter or longer times; but the methods themselves do not seek to evaluate efficiency as such. They have somewhat lesser interest for us because of this. However, in reviewing these methods we shall place primary emphasis on those aspects that relate to or reflect the influence of technician performance.

We shall discuss each model in turn, but it would be desirable to start with a capsule summary of their major characteristics.

The models may be classified (following the lead of References 5 and 6) into two general types: (a) the time-synthesis methods in which times associated with certain elemental maintenance tasks are synthesized or combined to derive an overall expected mean time to restore; (b) the correlation methods in which an equipment is evaluated using a checklist and the checklist scores are inserted into a multiple regression equation which yields an estimate of the expected maintenance burden. The four methods also differ in terms of the types of systems to which they are applicable and their point of application in the design cycle. For example, one method is for predicting flightline maintenance of avionics equipment, a second to predict shipboard and shore electronics gear, etc.

It must be noted that the focus of the prediction is exclusively in terms of time, not error. This is because equipment maintainability is defined in terms of downtime. It may be assumed that given infinite time all systems (except those destroyed or discarded) can be restored to operational use. Hence, the effect of error is simply to increase task time and thus downtime.

METHOD I (ARINC) [1]

GOALS

The purpose of this procedure is "to predict system downtime of airborne electronic and electro-mechanical systems involving modular replacement at the flightline" (p. 1-1, Ref. 3).

Note that this goal is simpler and hence more straightforward than the goals of the models previously reviewed (i.e., operability models). Again it must be emphasized that the procedure's avowed purpose is not to predict man, machine or system effectiveness, but merely system downtime. Nevertheless, every maintainability engineer recognizes that downtime is the resultant of the combined effects and effectiveness of equipment, personnel, logistics, administration, etc.

ASSUMPTIONS

Reference 3 indicates that "the philosophy of the entire prediction procedure is based on the principles of synthesis and transferability" (p. 1-2, Ref. 3).

The principle of synthesis refers to the assumption that total system downtime can be derived by the addition of downtime distributions for more elemental activities. "Application of the model consists of synthesizing various combinations of the elementary activities to form higher order maintenance tasks which are specific to a given system" (p. 273, Ref. 6). This assumption will be quite familiar to the reader who has read the description of the AIR Data Store and similar reliability-oriented methods. It is entirely justifiable because response times are obviously additive.

The concept of transferability refers to the assumption that "data applicable to one type of system can be applied to similar systems under like conditions of use and environment....." (p. 1-2, Ref. 3). Again the reader will recognize this assumption as being at the heart of the data bank concept. In terms of the elemental activities with which completion times are sufficiently molecular, they will be completely generalizable. "The Elemental Activity is a simple maintenance action of short duration and relatively small variance which does not vary appreciably from one system to another. An example of a basic elemental activity would be the opening and shutting of a door...." (p. 1-1, Ref. 3).

---

1 This designation refers to the developer of the methodology. For specific references to authors of the research reports on which this and the other methods are based, see Reference 6.

Several points should be noted about the concept of transferability and elemental activities. To be valid the concept must assume independence of the elemental activity from the effects of system design, although not necessarily from the effects of skill level. This poses certain difficulties. At a certain elementaristic level, the task activity may be virtually independent of system design, but where this level is can only be determined by testing across systems.

The concept of elementary activities also forces a highly analytic process on the procedure. System operations must be "decomposed", as it were, into progressively more elementary components; these must then be reconstructed to form the total system. This is standard reliability methodology and we have seen it as part of the reliability-oriented analytic models like the AIR Data Store, THERP, etc.

More specific assumptions, related directly to elemental activities, underlie the methodology:

"(a) The mean time required for the performance of an Elemental Activity is independent of system design and support facilities.

(b) The frequency of occurrence of an Elemental Activity correlates with some factor of system design or support facilities.

(c) The Elemental Activities in any maintenance category are independent of each other.

(d) The total time required in any maintenance category is completely accounted for by one or more of the Elemental Activities in the category." (p. 1-14, Ref. 3)

Assumption (d) above means in effect that the elements in any maintenance function (like localization, repair, etc.) completely account for the time required to perform that function. This is a reasonable assumption, given any reasonable taxonomy of maintenance behaviors.

Assumption (a) was referred to previously. We are dubious that any elementary activity can be completely independent of system design, because even the opening or closing of a door will depend in part on such aspects as the number of fasteners for the door or the characteristics of the door handle.

Assumption (b) is probably acceptable. For example, the number of connectors in a system will in part determine the number of times the technician must connect and disconnect this type of component.

Assumption (c) is almost certainly not correct. If the maintenance category is, for example, fault location, it cannot be divorced from what happens in the preceding category of malfunction verification. Rigney and Bond (p. 61, Ref. 5) point out that "the assumption of independence among

separate tasks is almost certainly inaccurate in some instances." However, they also point out that "the ARINC studies have shown a surprising insensitivity to category independence conditions; the synthesized distributions (of times) were about the same when category correlations were taken into account and when random combinations of categories were effected. It is difficult to see why this should be so unless the synthesis technique includes so many different items that a masking or blurring of separate effects occurs in the total-time distributions" (p. 61, Ref. 5).[2] Smith et al (Ref. 7) point out that errors in time estimation of individual categories do not seem to have an excessive effect on the product of such combined time estimates.

The reader should remember, however, that the statements above apply only to time; one could not generalize to the multiplicative combination of error probabilities.

The elemental activities referred to as the "lowest common behavioral denominator" of this methodology do not correspond to the subtask or task element levels found in other models. They (43 in all) more closely correspond to the task or even the gross task level. For example, sample behavioral elements are:

(1) Using test equipment to verify malfunctions....

(2) Interpreting symptoms by mental analysis.....

(3) Performing standard test.

The grossness of these elements makes it doubtful that they can be independent of system design.

Behavioral parameters are largely missing in this method, which is entirely understandable considering that its conceptual orientation is derived from engineering, not from Human Factors. Rigney and Bond point out that "the maintainability time synthesis schemes are relatively free of psychological variables such as commitment, urgency, susceptibility to stress and multi-person interaction" (p. 61, Ref. 5). The lack of behavioral parameters actually implies a further assumption: that behavioral factors have little impact upon maintenance task times, an assumption which a behavioral specialist will find it difficult to accept.

METHODOLOGICAL SCOPE

The procedure is restricted to airborne avionics and flight line replacement. This restriction is determined by the empirical data sources on which the task times for this procedure are based. Since the method depends largely on the application of its time data bank, the source of that bank restricts its applicability. This may appear somewhat at variance with the concept of transferability, but the transferability refers only to generalization across airborne avionics equipment removed and replaced at the flight line.

2. See also the studies by Mills and Lamb in Appendix C.

On the other hand, if we are really dealing with elemental activities, and if the times collected pertain to these activities, there is logically no reason why the data could not be used for other types of systems. Since this is apparently not true, one can only suspect that these so-called elemental activities are not really as elemental as they perhaps should be.

The implication of the procedural restriction to avionics is that the method will be employed only if one wishes to predict downtime for avionics equipment; if other types of equipment are involved, one of the other methods will be selected. In view of the differences among the other (i.e., operability) models reviewed previously, it might well be that this procedure could be followed generally: to look not for a single general purpose model, but to select the one most effective for a given application.

PARAMETERS

If we think of model parameters as being the elements that are involved in the application of the procedure, the following would be included:

(1) the behavioral elements already discussed (i.e., the maintenance tasks);

(2) system design characteristics, i.e., number, type and location of components;

(3) component failure rates;

(4) distribution of downtimes for the elements and for the maintenance categories.

The maintenance categories referred to are:

(1) preparation time;

(2) malfunction verification time;

(3) fault location time;

(4) part procurement time;

(5) repair time;

(6) final malfunction test time.

From a behavioral standpoint the above correspond to the functions into which active repair time can be subdivided. The categories are similar to others generally accepted by workers in the area.

DATA

1. Input Data Required

The following information must be made available to perform the prediction:

(a) location and failure rate of each component;

(b) number of flight-line replaceable components of each type;

(c) number and type of readouts;

(d) number of types of spares carried;

(e) number of pressure-retaining connectors and magnetrons;

(f) number of test points;

(g) nature of special test equipment;

(h) estimates of durations of average mission;

(i) manning schedules for operations and maintenance personnel;

(j) estimates of time occupied by unscheduled activities.

Some of the above items will be found in any system; others, like number of magnetrons, seem highly system-specific.

The major input data are the task completion times associated with the elemental activities, together with the appropriate distributions to be used. Three types of time distributions are available:

(1) the fitted normal distribution;

(2) the fitted log-normal distribution;

(3) a corrected time log-normal distribution.

The original prediction method assumed a normal distribution of times for the elemental activities. However, further refinement of the procedure resulted in the three distributions above.

Elemental activities having standard deviations less than an arithmetic mean of an hour are assumed to be normally distributed. This is because they are assumed to be of routine nature and hence, execution time should not be significantly influenced by changes of personnel, characteristics or surrounding events.

If, however, the standard deviations are greater than the arithmetic mean, or the mean is more than an hour, the elemental activities are considered "both as more complex and as containing many possible subactivities" (p. 1-12, Ref. 3). In this case the tendency is for the applicable distribution to be skewed to the right and is assumed to be of a log-normal nature.

The use of distributions (rather than simply the mean of the distribution) represents a greater degree of sophistication than one has encountered so far and is a procedure which should be followed with the other models reviewed. It should be noted, however, that it may well be easier to develop time distributions than similar distributions of errors.

2. Data Sources

The task times needed to perform the prediction are given to the user; his only responsibility is to select the appropriate distribution.

Parts lists and location data are available from system documentation.

Failure rates for components are derived from MIL HDBK 217A (Ref. 2) or a comparable source like FARADA.[2A] No subjective judgments are required except for the decision as to which time distribution should be selected.

3. Output Measures

> "The ultimate measure of maintainability is the distribution of System Downtime. Intermediate measures include the distribution of times for the various Elemental Activities, Maintenance Categories, Malfunction Active Repair Time, Malfunction Repair Time, System Repair Time and System Downtime." (p. 1-2, Ref. 3)

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

Obviously the breakdown to elemental activities requires a form of function/task analysis, although as with the other methods reviewed in this report a specific procedure for performing this analysis is not described. This analysis merely requires identification of the elemental activity, not any consideration of the parameters influencing these activities.

2. Method of Synthesis

The complexity of this procedure results from its "building block" assumption. Having associated an appropriate task time with the elemental activities, it is necessary to combine them in the following manner. The

2A. Computer Applications, Inc., Army, Navy, Air Force and NASA Failure Rate Data (FARADA) Program, Vol. I, Contract N00123-67-C-2528, New York, New York, 1968.

steps listed below are not to be considered a complete description of the procedure, which is extremely complex, but are merely illustrative.

(a) Calculate occurrence probability of each elemental activity;

(b) Determine probability of completing an elemental activity by transforming time values into Z values and looking up the equivalent probability in a table of the normal conditions;

(c) For each maintenance category and for each of the 20 time values multiply the probability of occurrence of the elemental activity by the probability of occurrence of each of the 20 times. For each of the 20 time values sum the probabilities for all the elemental activities comprising the maintenance category. Plot the distribution of times for each of the categories.[3]

The procedure then goes on to make use of the Weibull equation to determine administrative time. "Predict the cumulative distribution of administrative time by repeating this process for a minimum of 200 times and plotting the results" (p. 1-20, Ref. 3). From the distribution of Malfunction Active Repair time randomly select by a table of random numbers a minimum of 200 random values of this time. To each time add a value of administrative time. Multiply each of the 200 time values secured previously by 0.95N where N is the total average number of malfunctions following a flight which provokes a complaint. Plot the distribution of system logistic time and draw the best fitting line through the plotted points. Determine its probability of occurrence. Multiply each system final test time value by the "readout factor"; plot the distribution of system final test times and draw a best fitting line through the plotted points. All of the preceding steps (which are considerably abbreviated in this description) lead to the determination of system repair time which can be made by drawing a number of random samples from a distribution of times already available.

ANTICIPATED MODEL USES

As indicated previously, the procedure is restricted to the prediction of system down time. In terms of the use categories we have employed with other models, the method does not predict effectiveness. Presumably, if there is an explicit or implicit standard relating to an acceptable system downtime, then one could use the downtime estimate derived from this procedure as an effectiveness measure, but this use of the procedure would be tenuous at best.

---

3. The author had intended to provide a step by step description of the statistical method involved, but at this point he threw up his hands at the difficulty of making that complex procedure intelligible. Hence, the following paragraph (above) is only a summary.

As far as its use for comparison of alternative design configurations, the procedure is far too complex for this. Moreover, by its assumption that the elemental activities are relatively independent of system design, the procedure eliminates any design sensitivity.

With regard to design analysis, Smith et. al. (Ref. 6) complain that the method cannot be used during design. Rigney and Bond (p. 61, Ref. 5) indicate that "But often the design significance has not been well enough worked out to give detailed guidance. A project to clarify and extend the design meaning of various parameters in time synthesis might be worthwhile. For the results of such a project to be practically applied, the designer should be told how to identify those features that contribute most to the time simulation" (underlining that of the author). We find then the same problem of design-relevancy in maintainability prediction that we encountered in the models previously reviewed.

Applications of the method to selection and training are not relevant.

VALIDATION/APPLICATION STUDIES

In the original development of the method data were gathered from the AN/ASB-4 Bomb/Nav system in the B-52 bomber and then applied to 7 other systems. Graphs are presented in Reference 1 that indicate good agreement between predicted and observed times (pp. 1-5 through 1-10, Ref. 3). Rigney and Bond (p. 60, Ref. 5) indicate that "validation so far appears rather positive". Cunningham (Ref. 1) indicates that "this prediction technique, with at least 42 separate equations, is not for the beginning (maintainability) engineer nor for the beginning statistician. It is our feeling that the prediction method is highly complex..... At the EIA (maintainability) Workshop, May 1968, one trial was reported using this procedure and that was dropped for lack of data.....".

The impression one receives is that although this method demonstrates reasonably good validity, it is unlikely to be used because of its complexity. Apparently utility criteria are as important as -or more so than- validity in the actual use of the method.

EVALUATION SUMMARY

Validity - Method has been formally validated. However, it has been infrequently applied because of its complexity.

Reliability - No formal data available, but reasonable reliability can be inferred from validity testing.

System Development Applicability

A. Comprehensiveness: Like all time-synthesis methods, limited by its input data to flight line avionics.

B. Applicability: Does not output prediction of system effectiveness, only downtime. Has no applicability for design analysis, selection or training.

C. Timing: Any time after design concept has been established, provided requisite data are available.

Model Characteristics

A. Objectivity: Relatively few subjective judgments required.

B. Structure: Some objection to basic assumptions; otherwise model structure is well defined.

METHOD II (Federal Electric Corporation)

GOALS

These are essentially the same as Method I. However, the procedure is specialized for shipboard and shore electronic systems. The method has two parts, part A predicting corrective maintenance time, part B predicting active maintenance time (which includes both corrective and preventive maintenance). Part A results in a maintainability prediction expressed in hours because it utilizes tabulated maintenance task repair times, recorded in hours. Part B does not use tabulated task times; instead it uses estimates of man-hours which are based on past experience or an analysis of the design with respect to maintenance.

ASSUMPTIONS

This method is also a time synthesis method. It assumes that "the magnitude of the repair time....is the sum of the individual maintenance task times which are required for its completion" (p. 2-2, Ref. 3). However, the times involved are average times instead of distributions. It also uses standard electronic part failure rates.

The individual maintenance tasks referred to in the previous paragraph are actually functions rather than tasks, i.e., localization, isolation, disassembly, interchange, reassembly, alignment and checkout. We are dealing in this procedure with behaviors conceptualized at a much grosser level than Method I; in fact, the tasks of Method II resemble the maintenance categories of Method I.

Another basic assumption which differentiates Method II from Method I is that of maintenance level. It is assumed that "at the part level of repair it takes longer to perform a repair task than at the subassembly or equipment level because less time is required for the discrete steps involved at the latter levels" (p. 2-3, Ref. 3). The maintenance times vary then "as a function of the depth to which a technician must penetrate an equipment to localize and isolate failures" (p. 274, Ref. 6). An essential part of the technique is the selection of task completion times in terms of the level at which maintenance is to be conducted. There are 9 functional levels: system, subsystem, equipment, group, unit, assembly, subassembly, stage, and part.

The concept of functional level makes Method II much more sophisticated than Method I. It should also make the downtime estimates more realistic than those supplied by Method I; we shall see later if this is true.

Another assumption, which is crucial for part B of this method, is that judgments can be made of the tasks required for maintenance by analyzing basic features of design and that the time required for these can be estimated from these judgments. This assumption directly links the methodology with design features, a characteristic that was not found in Method I.

Like Method I there are no behavioral assumptions in this procedure.

The fact that standard times are utilized and not time distributions as in Method I implies an assumption that such distributions are unnecessary. One may criticize this assumption as being statistically naive.

METHODOLOGICAL SCOPE

This methodology is restricted to naval equipment because of the source of its time data. However, the basic concept structure, e.g., completion time as a function of maintenance level, would seem to be capable of generalization to other system-types, if the appropriate task time data were gathered. The fact that the method is apparently restricted by its data source seems to imply an assumption that the influence of design on task time is such that no generalization across systems is possible. Yet the 7 maintenance tasks (or functions, as we would term them) are utilized in describing maintenance of all systems.

PARAMETERS

One finds in this methodology essentially the same parameters as those of Method I, except for the crucial parameter of functional maintenance level. Again, system design features and failure rates for components are important. Probability of occurrence is, however, not specifically utilized in this method, although component failure rate does influence that probability.

DATA

1. Input Data

These are essentially the same as those required by Method I. The specific items of information needed are listed below:

a. Packaging, to the extent that a detailed breakdown into the various equipment groups, items, etc., can be determined.

b. Diagnostics, i.e., the diagnostic procedure that would be followed in the event of failure of each part in the equipment.

c. Repair methods.

d. Parts listing.

e. Electrical and environmental stresses.

f. Mounting, the method by which each individually replaced part is mounted.

g. Functional levels, the functional levels at which alignment and checkout are performed.

Since the prediction method requires that the sum of the maintenance task times be multiplied by the failure rate to obtain an estimate of number of hours required for each maintenance action, failure rates must be secured from MIL HDBK 217 and NAVSHIPS 93820.

The other major input data needed are the maintenance task times, which are of two types. For part A of the procedure there are tabulations which are supplied for each maintenance level and each major maintenance function. For part B of the procedure one has to estimate the time for these functions.

A specified method for performing this estimation is not provided. For example, an instruction might be:

> "(1) Assuming that each component fails in its most likely mode, note the fault localization features and determine the necessary steps to localize the fault to the module or function. Estimate the average localization time.....
>
> (4) Observe the method used to attach a failed component to its mounting surface and perform an analysis to estimate the interchange time." (p. 2-28, Ref. 3).

This is subjectivity at its most extreme, since no ground rules at all are provided, and renders the part B methodology highly suspect. It is interesting, however, that Cunningham (Ref. 1) prefers part B to part A. However, he indicates that applicable data for part B "can be obtained from a credible data bank or from engineering judgment".

2. Data Sources

The task times used in part A were derived from over 300 observations of maintenance activity in the fleet. The tables for part A provided by Reference 1 serve as the data source. For part B presumably the data source for task times could be anything.

3. Output Data

For part A the basic maintainability measure is something called equipment repair time (ERT) expressed in hours. This is the median of the individual repair times and can be determined as follows:

a. When repair times follow a normal distribution, ERT is equal to MTTR.

$$ERT = \frac{\Sigma(\lambda R_p)}{\Sigma\lambda} = MTTR$$

where: λ = average part failure rate in failures per 10^6 hours.

R_p = repair time required to perform a corrective maintenance action in hours.

b. When repair times follow an exponential distribution:

$$ERT = 0.69\ MTTR$$

c. When repair times follow a log-normal distribution of repair times

$$ERT = \frac{MTTR}{\text{antilog}\ (1.15\sigma^2)}$$

where σ is the standard deviation of the logarithms to the base 10 of repair times.

d. When repair times follow a log-normal distribution, the geometric mean time to repair ($MTTR_G$) occurs at the median, therefore it is a measure of ERT.

$$MTTR_G = \text{antilog}\ \frac{\Sigma(\lambda \log R_p)}{\Sigma\lambda}$$

For part B the applicable formulation to obtain the mean corrective maintenance time ($\overline{M}_c$) in man-hours is

$$\overline{M}_c = \frac{\Sigma(\lambda M_c)}{\Sigma\lambda}$$

where λ = average part failure rates in failures per 10^6 hours.

$\overline{M}_c$ = the man-hours required to perform a corrective maintenance task.

Mean preventive maintenance time ($\overline{M}_p$) is equal to $\frac{\Sigma(fM_p)}{\Sigma f}$

where M_p = the man-hours required to perform a preventive maintenance action.

f = the frequency of occurrence of preventive maintenance actions per 10^6 hours.

Active maintenance time is the sum of the preventive and corrective maintenance man-hours required to maintain a product for a specified period, divided by the total number of preventive and corrective maintenance tasks required during that time. Mathematically it is expressed as

$$M = \frac{(\Sigma\lambda)\overline{M}_c t_i + (\Sigma f)\overline{M}_p t_j}{\Sigma\lambda t_i + \Sigma f t_j}$$

where: $\overline{M}$ = mean active maintenance time

$\overline{M}_c$ = mean corrective maintenance time (resulting from time, t_i)

$\overline{M}_p$ = mean preventive maintenance time (during time t_j)

$\Sigma\lambda$ = the sum of the part failure rates

Σf = the sum of the frequency of occurrences of preventive maintenance tasks

t_i = operating time during period t_j

t_j = calendar time in operating inventory.

The total maintenance effort (maintainability index MI) required to maintain a product in operational status per unit of operating time is

$$MI = \frac{(\Sigma\lambda)\overline{M}_c t_i + (\Sigma f)\overline{M}_p t_j}{t}$$

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

The analytic method is the customary function/task analysis (although without any behavioral implications). The first step is to determine the functional level breakdown of the equipment or system. This is done by

dividing the equipment or system into its various physical subdivisions beginning with the highest subdivision and continuing down to the part, subassembly, assembly, etc., level. A functional level diagram (see Figure 2-1 of Ref. 3) is usually prepared.

After the functional level breakdown has been established and the diagram prepared, the functional levels at which localization, isolation, access, alignment and checkout features are applicable are determined. These are indicated directly on the diagram.

2. Method of Synthesis

The method of synthesis is essentially the same as that in Method I, although substantially less complex, and involving the notion of functional level.

For part A the prediction involves listing the component item, method of repair, the circuit designation, type of part and the failure rate of each part (referring back to MIL HDBK 217, etc.). The failure rate values are then summed. Maintenance task times for each component part are determined by reference to a table of these (which table automatically includes the function level parameter). These maintenance task times for each part are then added to secure R_p. The failure rate for the part is multiplied by the summed maintenance task time to secure λR_p and transformed into logs. The sum of the λR_p values is determined. These calculations will then supply equipment failure rate ($\Sigma\lambda$), $\Sigma\lambda R_p$ and $\Sigma\lambda \log R_p$. Mean time to repair (MTTR) is calculated by

$$MTTR = \frac{\Sigma\lambda R_p}{\Sigma\lambda}$$

Geometric MTTR is calculated as

$$MTTR_G = \text{antilog}\left[\frac{\Sigma(\lambda \log R_p)}{\Sigma\lambda}\right]$$

Part B of the procedure for corrective maintenance prediction is essentially the same except that the maintenance task times are estimated rather than determined from already existent tabulations. For preventive maintenance prediction the frequency of occurrence of the preventive maintenance task and the man-hours required for that task are substituted for the corrective maintenance time estimates.

ANTICIPATED MODEL USES

Same as for Method I. The method is limited to the final design phase.

VALIDATION/APPLICATION STUDIES

A validation study carried out on the AN/URC-32 Transceiver and the AN/SRT-16 Transmitter showed reasonably good correlation between predicted and observed maintenance times. However, the actual correlation values were not indicated, nor was there any distinction made between parts A and B of the procedure.

From a utilization standpoint this method is much simpler and easier to apply than Method I, which would suggest that as between the two methods, the former would be preferred.

EVALUATION SUMMARY

Validity - Formal validation tests indicate reasonably high validity.

Reliability - No data available.

System Development Applicability

A. Comprehensiveness: Limited to ship systems.

B. Applicability: Like most maintainability prediction models, does not predict system effectiveness, only downtime.

C. Timing: Only during final design stage.

Model Characteristics

A. Objectivity: Part A highly objective; Part B highly subjective.

B. Structure: Part B poorly defined.

METHOD III (Radio Corporation of America)

GOALS

This method is an example of the correlational approach mentioned previously. It differs from the previous methods reviewed by involving the analysis of maintenance tasks required by a selected set of malfunctions; this analysis is then used to complete 3 checklists whose scores are inserted into a multiple regression equation which then outputs a time prediction. It is significant that of all the methods reviewed in this report, this is the only one which makes use of checklists as a means of inferring from system characteristics to some performance prediction.

Beyond this feature, the method has the same objectives as the two previous ones: to predict system downtime.

ASSUMPTIONS

A fundamental assumption underlying this procedure is that system downtime "is assumed to be a function of specific design parameters which relate to: the physical configuration of the system; the facilities provided for maintenance....and the degree of maintenance skills required of personnel...." (p. 3-1, Ref. 3). Although Method II assumed that one could estimate downtime by analyzing design features, these were not specified in any detail. From that standpoint (and particularly with reference to skills), this maintainability model is far more behaviorally oriented than the ones previously reviewed.

The procedure also assumes that because of basic uniformity of design, a random selection of items by class will provide a representative sample of maintenance tasks. In other words, an essential element of the method is the use of random sampling procedures. The use of sampling in predicting downtime implies that on the average it should take the same time to correct one resistor or capacitor failure as for another resistor or capacitor failure. Apparently the possibility of an effect of other design features on maintenance time is not considered. Whether or not this assumption is justifiable cannot be determined; in any event, it is heuristically defensible.

It might also be noted that the random sampling aspect of this procedure has become standard for the selection of maintenance tasks to be used in maintainability demonstrations (Cunningham, Ref. 1).

One of the characteristics of all the maintainability methods reviewed is the emphasis on standardized equipment units to which predictive data can be applied. In other words, a resistor is a resistor is a resistor, regardless of the system in which the resistor is found. This is somewhat at variance

with procedures in the operating (non-maintenance) models reviewed, where the behavioral unit to which predictive data are applied is not very specifically defined. Thus, these latter models do not make use of a standard set of behaviors or subtasks. It is of course much more difficult to define these than to define equipment components. The difficulty is reflected in the failure over many years to develop a universally accepted behavior taxonomy. Yet one might ask whether a viable operability prediction method can be developed without such a taxonomy.

One final note. The checklist methodology is based on the original research which adopted a correlational approach. It should be noted that only one other maintainability model considered (Siegel's personnel reliability index) is based on a correlational approach and none of the operability models fall into this category. One may speculate why this orientation has not been utilized more frequently.

METHODOLOGICAL SCOPE

The method is restricted to the prediction of ground electronics equipment maintainability, apparently on the basis of the systems on which the original research was performed. However, since the validity of the method was demonstrated for ground electronic equipment and the checklists were sufficiently generalizable, there is no reason why the method could not be expanded, even if the original research had to be redone in part.

PARAMETERS

The elements of the methodology include:

(1) system design characteristics;

(2) test equipment;

(3) personnel requirements and skills;

(4) component failure rates;

(5) malfunction diagnosis procedures.

Items (1), (2) and (3) above are specifically included in the three checklists which are an essential foundation for the method. Component failure rates are used in the random sampling procedure for selection of maintenance tasks. "The size of each n sample (Task Sample) is determined by considering the relative frequency of failure for a particular class of replaceable items" (p. 3-5, Ref. 3). The steps in the malfunction diagnostic procedure are the raw material used for the checklist analyses.

As indicated previously, the method manifests a greater awareness than the other maintainability prediction methods reviewed of the behavioral elements involved in system restore time.

DATA

1. Input Data

A very considerable amount of system-specific information is required. Reference 3 (p. 3-3) indicates that the evaluator should be at least familiar with following: schematic diagrams, physical layouts, functional operation of the equipment, tools and test equipment, maintenance aids and the operational and maintenance environment.

Input data of the data bank variety (i.e., task times) are not required because the ultimate time predictions made depend on checklist judgments which are inherent in the anlytic process.

There are three checklists to be applied. Checklist A, which scores physical design factors impacting on maintenance of the failed item, resembles the typical human engineering checklist for maintainability. That is, it deals with such characteristics as access, fasteners, packaging, displays, test points, etc. Checklist B, which scores design facilities, actually includes two separate factors: additional design features (e.g., test equipment, connectors and jigs) and certain non-design (e.g., personnel) features like assistance received from others and visual contact between personnel. Checklist C deals with maintenance skills required and hence requires very complex judgments. For example, it demands measures of the amount of energy expended, sensory, alertness, etc. and quantitative evaluation of these items is measured in terms of the amount of effort required in applying these factors. Scores on these checklists vary from zero to 4.

The original research demonstrated a multiple correlation of .74 between the checklists and maintenance time. However, the primary factor in this correlation was checklist A, which correlated .56 with the criterion. Checklists B and C correlated .28 and .16 respectively with the criterion. The low correlations of the latter two checklists is quite understandable because of the complexity of the judgments involved. Rigney and Bond (p. 74, Ref. 5) indicate that the support and personnel variables were dropped because of their low correlation with maintenance time; however, reference 3 still retains them. Rigney and Bond (p. 75, Ref. 5) suggest that "Perhaps the negligible P correlations (personnel) are due partly to restriction in range. Military technicians are selected on intelligence (average IQ about 116); about 80 percent of them are at or above the 6th Stanine of the Electronics Aptitude Index; their amounts of technical training are apt to be similar; and because of high turnover rates the average technical experience

on any crew is apt to be a little over 2 years. Thus, over a series of eight field sites, there may not have been sufficient variability to produce important relationships, even though correlations in a less standardized population (or either Personnel or Support circumstances) would prove to be significant.... Perhaps a second reason for the negligible P and S correlations lies in the time criterion. It could be that individual differences in Personnel are reflected in qualitative ways not evident to a simple time measure....."

Nonetheless, the use of a checklist methodology based on a multiple regression correlational approach is still quite interesting. The use of checklists to derive some sort of maintainability score has a number of precedents, notably the efforts of Munger and Willis (Ref. 4) and of Topmiller (Ref. 8). Moreover, the use of a checklist methodology permits a more direct relationship between the system prediction and design analysis, a relationship which is lacking in the other maintainability (and operability) models.

2. Data Sources

These are inherent in the system description.

3. Output Measures

There are four of these.

(a) mean corrective maintenance time, $\overline{M}_{ct}$;

(b) mean preventive maintenance time, $\overline{M}_{pt}$;

(c) mean downtime, $\overline{M}_{t}$;

(d) maximum corrective maintenance time, M_{max}.

The mathematical equations for these are:

(1) Mean corrective maintenance time

$$\overline{M}_{ct} = \frac{\sum_{i=1}^{N} M_{ct}}{N}$$

where: $\overline{M}_{ct}$ = mean corrective maintenance time;

N = sample size of corrective maintenance tasks;

M_{ct} = corrective maintenance time of individual maintenance tasks.

(2) Mean preventive maintenance time

$$\overline{M}_{pt} = \frac{\sum_{i=1}^{N} M_{pt}}{N}$$

where: $\overline{M}_{pt}$ = mean preventive maintenance time

M_{pt} = preventive maintenance time of individual maintenance tasks.

(3) M_{max} is expressed as

$$M_{max} = \text{antilog} \left[\overline{\log M_{ct}} + 1.645\sigma \log M_{ct}\right]$$

where: $\overline{\log M_{ct}} = \frac{\sum_{i=1}^{N_c} \log M_{cti}}{N_c}$ = mean of $\log M_{ct}$ and

$$\sigma \log M_{ct} = \sqrt{\frac{\sum_{i=1}^{N_c} (\log M_{cti})^2 - (\sum_{i=1}^{N_c} \log M_{cti})^2 / N_c}{N_c - 1}}$$

PROCEDURES FOR MODEL APPLICATION

To apply the method the evaluator selects a group of malfunctions on the basis of failure tables. Each maintenance task required to resolve the malfunction is traced out in detail, including out of tolerance indications that result, special problems of access or removal, etc. The separate tasks are then scored on the three checklists and the sum of the derived scores yield values that are inserted into the regression equation. Solution of this equation produces the estimated maintenance times.

1. Analytic Method

The steps involved are:

(1) Determine the sample size and list the component items whose simulated failure (on paper) will lead to the maintenance analysis.

The selection process is far too complex to describe in this section; interested readers are referred to pp. 3-5 through 3-11 of Ref. 3.

(2) Perform a maintenance analysis by specifying the type of failure expected of each component selected in (1) above. This involves determination of the mode of failure and the malfunction symptoms. The analysis is conducted by listing each step required to resolve the simulated failure. The maintenance analysis has certain similarities to task analysis, at least in the specification of the required tasks; and to the probability tree analysis used in THERP, although not at the latter's level of detail. However, behavioral analysis is left to the application of the checklists.

(3) Apply the checklists to the maintenance steps. Each checklist item is applied to the totality of the maintenance steps, not to each step. Thus, regardless of the number of maintenance steps required by any failure, there will be 15 scores in checklist A, 7 in B, and 10 in C.

2. Synthetic Method

The individual checklist scores are then summed and inserted into the regression equation which takes the form:

$$M_{ct} = \text{antilog}\ (3.54651 - 0.02512A - 0.03055B - 0.01093C)$$

To facilitate the calculation of M_{ct} a nomograph is available which permits the prediction of downtime directly in real time instead of log values. $\overline{M}_{pt}$, M_t and M_{max} can be derived by substitution in the equations described under Output Measures.

Cunningham (Ref. 1) indicates that the nomograph "is unfortunately limited to a minimum task time of about 6 minutes, which, in today's maintenance world of quick-restore-time, is incomplete".

ANTICIPATED MODEL USES

Like the other maintainability models, this one does not predict system effectiveness. However, the use of the checklists permits very direct design analysis and redesign suggestions.

VALIDATION/APPLICATION STUDIES

The model was developed on the basis of research on three pieces of ground electronic equipment, a long range search radar, a data processor and a data link transmitting equipment. Validation was performed on two

equipments, the AN/FPS-6 radar and the AN/GRT-3/GRR-7. A comparison of predicted versus actual downtime showed good results for the latter, but relatively poor results for the former. There was a rather high standard error (0.32 times log M_{ct}) which implies a wide spread of obtained values around the predicted value.

> "Such a large error term, of course, results from the characteristics of the original standardization data; the N's are necessarily rather small, the simple correlations between checklists and the time criterion are not extremely high, and the checklists are themselves moderately related to each other. If, now, one were to perform multiple regressions on the new data, one would expect the correlations between checklists and criterion to be higher and the standard errors smaller, if only because the prediction procedure would be applied to the same set of failures that was observed in the field.....
>
> RCA has carried through new multiple regressions for the AN/FPS-6 and the AN/GRT-3/GRR-7, using the field-observed troubles as the basis for the maintenance analysis and checklist scoring. As expected, the newly-fitted equations provide accurate 'predictions'; indeed, the multiple correlations were extremely high (0.95 for radar, 0.96 for communications)...." (p. 70, Ref. 5).

Cunningham (Ref. 1) indicates that "the author's experience with this prediction technique when used for the final design phase prediction, is that the measured values were roughly twenty percent higher than predicted, in the three or four times used. When used in the preliminary design phase, the predictions were found to be totally unreliable."

Smith et. al. (p. 276, Ref. 6) indicate that the predictions derived from this method either over or under-estimated actual maintenance time.

It would appear then that there are serious deficiencies in this method. Nonetheless, if one assumes a relationship between maintenance task times and maintainability design and support features, additional research using this approach would seem quite worthwhile. In fact, the concept of a checklist approach to operability models might well be considered, because of the desirability of relating system effectiveness to design features.

EVALUATION SUMMARY

Validity - Formal validation tests prevent conflicting evidence about validity, some tests showing low validity, others, high validity.

Reliability - No data available, but one would suspect lowered reliability because of the subjectivity involved in checklist evaluations.

System Development Applicability

A. Comprehensiveness: Limited to ground electronics systems.

B. Applicability: Like most maintainability prediction models, does not predict system effectiveness, only downtime. May be good tool for design analysis because checklist A relates design characteristics directly to downtime.

C. Timing: Can be applied during design/development.

Model Characteristics

A. Objectivity: Many subjective elements.

B. Structure: Parameters included in checklist instruments are poorly defined.

METHOD IV (Republic Aviation)

This procedure is another of the time-synthesis methods. As Smith et. al. (p. 274, Ref. 6) indicate, it "basically consists of a framework for compiling and combining data on any system. Its generality principally derives from the fact that the model user generates all input data for relatively basic, 'macroscopic' equations".

GOALS

These are the same as the other methods reviewed: To predict the mean and/or total corrective and preventive maintenance downtime of systems and equipments.

ASSUMPTIONS

A major assumption is that maintenance time depends on the specific operational function which is in progress. "In other words, the procedure requires the development of a mission/maintenance profile which specifies the various operational functions of the system and the scheduled preventive maintenance actions required for each operational function" (p. 4-1, Ref. 3). The method is therefore highly mission oriented.

Another major assumption, about which the author is highly dubious, is that estimates of task times can best be made by the maintenance analyst working closely with the design engineer, or by the design engineer himself. However, no procedure is supplied for making such estimates. The reader will have noted that the author places much importance on the necessity for describing in detailed form all procedures to be employed. Otherwise, one of the essential evaluative criteria for man-machine predictive models, i.e., reliability, cannot be applied.

The other assumptions inherent in the time-synthesis approach (which have been described previously) also apply to the method. The times to be input to the model are, like Method II, mean values rather than distributions. This is supposedly permissable because of the high degree of correlation between predicted and observed values (a correlation which, unfortunately, is not specified).

As in the case of some of the other maintainability models, no behavioral assumptions are included in the model.

METHODOLOGICAL SCOPE

Because of the generality of the model and the absence of an input time data bank derived from observations on actual equipment, the method

is considered applicable to all systems. "Because of the nature of the time estimating techniques (historical and idiosyncratic), this procedure is applicable to all systems/equipment" (p. 4-2, Ref. 3). For the same reasons the method can also be applied at any design stage, including the very early period. The output measures derived do not include estimates of administrative or delay times, because these are not normally definable during the design of the equipment.

PARAMETERS

These are the same as found in the other methods, e.g., system design characteristics, end item components, failure rates, etc. The list of information required (see below) suggests an effort to "cover the waterfront" in terms of factors influencing maintenance task time. Personnel are mentioned as part of the operational resources for which information is needed, but nothing further is done with this parameter.

DATA

1. Input Data

All data needed to apply the model are inherent in the system and in the experience of the evaluator. The following information is required:

a. system block diagram, functional flow diagrams and subsystem block and flow diagrams;

b. end item list and end item failure rates;

c. maintenance concept, maintainability goals, operational resources, facilities, personnel, support equipment;

d. definition of the task being performed, location, and environmental constraints.

2. Data Sources

These are completely unspecified. There is total reliance on the judgment of the analyst.

3. Output Data

Three measures are provided by the model: (a) mean corrective downtime (MCDT); (b) mean preventive downtime (MPDT); (c) total downtime (TMDT). These in no way differ from output measures of the other models.

PROCEDURES FOR MODEL APPLICATION

1. Analytic Method

There is heavy emphasis on task analysis (without, however, any consideration of behavioral parameters). For example,

> "The estimated elapsed time required to perform maintenance on a system will vary as a function of the conceptual and physical constraints within which the estimation was made. These constraints consist of the availability of physical resources (i.e., personnel, spares and consumables, support equipment and facilities) and applicable maintenance and operational concepts (i.e., testing concept, level of repair, mission descriptions, etc.)....." (p. 4-4, Ref. 3)

These physical and conceptual constraints are very inadequately defined in the above procedure. Or, "A series of mission/maintenance profiles will be established based on the system operational requirements....." (p. 4-5, Ref. 3).

More detailed steps involved in the analysis are:

(1) The end items of the system are identified down to the smallest piece of equipment on which a specific maintenance action is to be performed.

(2) The failure rate is identified for each item.

(3) The preventive and corrective maintenance actions to be performed with these items are identified and defined.

(4) A task analysis is conducted for each preventive and corrective maintenance action in (3).

(5) A distribution of task times for each end item action is generated.

(6) The total task time for an operational function is compared to the allocated time to determine if the maintainability design of the equipment is adequate. If not, an analysis is made of critical design points.

(7) All task times and associated downtimes are integrated over calendar time to derive total preventive downtime, total corrective downtime and total mean downtime.

It should be noted that the method is extremely general.[1] Because of this it includes one feature not present in the other methods. We refer in particular to the comparison of total predicted task time with

1. However, note that it contains all the major steps included in the previous methods.

allocated time to determine if the maintainability design of the equipment is adequate. This is the only maintainability method which attempts to establish an evaluation of system effectiveness in maintainability terms.

2. Synthesis Method

This procedure is essentially the same as other time-synthesis methods. Given that times are estimated for preventive maintenance tasks, these are added together, i.e.,

$$PDT_m = \sum_{i=1}^{m} T_{i_m}$$

where: PDT_m = the total preventive maintenance performance time for action P_m.

T_{i_m} = the time to perform the maintenance task on end item I_i as required by action P_m.

To determine corrective maintenance times, the troubleshooting, repair and verification times are derived in the following manner:

$$T_{i_m} = (\Sigma T_{s_{i_m}}) + T_{c_{i_m}} + T_{v_{i_m}}$$

where: T_{i_m} = the total time required to correct malfunctioning end item I_i during action P_m of an operational function.

$T_{s_{i_m}}$ = the troubleshooting test times required to isolate end item I_i during action P_m.

$T_{c_{i_m}}$ = the time required to remove, replace, adjust or otherwise repair malfunctioning end item I_i during action P_m.

$T_{v_{i_m}}$ = the time required to verify that the system is good, given that I_i is replaced, repaired, adjusted, etc. during action P_m.

The same general procedure is employed throughout. MCDT for the system is given by the weighted (normalized failure rates) of the MCDT as below:

$$MCDT_s = \frac{\Sigma(\lambda_{i_r} + \lambda_{i_{gr}})\ MCDT_r + \Sigma(\lambda_{i_m} + \lambda_{i_{gm}})\ MCDT_m}{\Sigma(\lambda_{i_r} + \lambda_{i_{gr}}) + \Sigma(\lambda_{i_m} + \lambda_{i_{jm}})}$$

where: $MCDT_s$ = the mean-corrective-downtime for the system for the given mission/maintenance profile.

The total mean corrective downtime of the system for the mission/maintenance profile is given by

$$MCDT_t = f\,(MCDT_s)$$

where: $MCDT_t$ = the total mean-corrective-downtime of the system for the mission/maintenance profile

f = the number of detectable failures occurring during the calendar time.

Some rather abstract matrix forms to be used in the prediction are shown in Reference 3, but these are very abstract and when analyzed they turn out to involve essentially the same operations as those of the preceding maintenance models.

ANTICIPATED MODEL USES

Besides the prediction of system downtime (which is characteristic of all these maintainability models) this method can (presumably, but not actually) be used to predict system effectiveness from a maintainability standpoint, in terms of the comparison between required and allocated downtime. Presumably also the method could be used for design analysis, but how this would be accomplished is unspecified. Selection and training analysis is non-applicable.

VALIDATION/APPLICATION STUDIES

The handbook from which the preceding description was extracted makes a vague reference to correlations between predicted and observed values, but no data are provided. The method is extremely general, but because of its generality it is difficult to see how it could be used practically. As a model which is a generic form of the time-synthesis methods, it may have been used, but in the form in which it is presented in Reference 3, it cannot be used. Cunningham (Ref. 1) indicates that he can say nothing about the method, because he has never used it.

EVALUATION SUMMARY

Validity - No data available, no evidence method has ever been used.

Reliability - The lack of structure in the model would suggest high subjectivity and correspondingly low reliability.

System Development Applicability

A. Comprehensiveness: Presumably applicable to all systems.

B. Applicability: Presumably predicts maintainability design adequacy as well as downtime, but not system effectiveness.

C. Timing: Can be applied at any stage.

Model Characteristics

A. Objectivity: Highly subjective.

B. Structure: Model is so poorly defined that its use would be difficult.

SUMMARY COMMENTS ON THE MAINTAINABILITY PREDICTION MODELS

Certain characteristics of these models stand out:

(1) There is very little consideration of personnel factors in these models, even considering Method III and despite the analytic procedure which seems similar to behavioral task analysis.

(2) Despite the complexity of the mathematical equations employed, they appear very simplistic, depending as they do largely on the simple addition of task times. The concept of functional levels in Method II does, however, represent an interesting feature.

(3) A number of the models depend heavily on the individual judgment of the maintenance analyst, without specifying procedures to be followed in securing these judgments. This tends to support Swain's contention that engineers are willing to accept data secured very unsystematically.

(4) More effort has gone into attempts to validate maintainability models than has gone into operability models. Even so, the governmental support of model validation has been sparse.

(5) The maintainability models appear to be restricted by the characteristics of their input data sources. In other words, with the exception of Method IV, there appears not to have been an attempt to develop a general purpose maintainability prediction method, contrary to the situation with the operability models.

(6) The maintainability models make no effort to use computer simulation methods, probably because the complexity of the troubleshooting process would make this inordinately difficult. There has been some research on simulation in maintenance (at least in terms of mockup test devices (see Reference 5) but this has been directed solely to research on maintainability parameters rather than model development. Although the maintainability models we have dealt with are of the analytic type, they include certain features which we do not find in analytic operability models, namely the use of checklists and multiple regression equations (correlational approach).

(7) The maintainability predictive models seem to be somewhat more advanced than the operability models reviewed in terms of larger data bases derived from operational environments and certainly in terms of actual application to system development uses. One gets the impression of more governmental support in the development of these methods. There appears also to have been somewhat more practical direction to the research leading to the development of these models (in other words, the immediate research goal was development of a practical predictive method) than one finds in the operability models. All of this suggests that the development of the operability models would proceed further with more consistent government support and a more practical direction.

REFERENCES

1. Cunningham, C. Applied Maintainability Engineering (to be published).

2. Military Standardization Handbook, Reliability Stress and Failure Rate Data for Electronic Equipment, MIL-HDBK 217.

3. Military Standardization Handbook, Maintainability Prediction, MIL-HDBK 472, Dept. of Defense, May 1966.

4. Munger, M. R. and Willis, M. P., Development of an Index of Electronic Maintainability, Report AIR-275-59-FR-207, American Institute for Research, Pittsburgh, Pa., 1959.

5. Rigney, J. W. and Bond, N. A., Maintainability Prediction: Methods and Results, Technical Report 40, University of Southern California, June 1964.

6. Smith, R. L., et. al., The Status of Maintainability Models: A Critical Review, Human Factors, 12 (3), 1970, 271-283.

7. Smith, R. L., et. al., Subjective Judgment as a Means of Generating Corrective Maintenance Time Data, Integrated Sciences Corporation, Santa Monica, California, July 1970 (to be published as AMRL report).

8. Topmiller, D. A., A Factor Analytic Approach to Human Engineering Analysis and Prediction of System Maintainability, Report AMRL-TR-64-115, Aerospace Medical Reserach Laboratory, Wright-Patterson AFB, Ohio, Dec. 1964.

III. PERSONNEL RELIABILITY INDEX

INTRODUCTION

The personnel reliability index developed at Applied Psychological Services by Dr. Arthur Siegel and his coworkers is a method for providing feedback on the technical proficiency of Navy electronic maintenance personnel.

The index is based on the compounding of probability of successful performance values of each of 9 electronic job dimensions derived on the basis of multi-dimensional scaling analysis. Supervisors were asked to report the number of uncommonly effective (UE) and uncommonly ineffective (UI) performance observed over a two month period in their technicians for each job factor. A ratio,

$$\frac{\sum UE}{\sum UE + \sum UI}$$

derived for each job factor, yields a value which varies between 0.00 and 1.00. When ratios for each job factor are combined the result can be used to "achieve a total effectiveness value for a technician" (p. 11, Ref. 3).

The effectiveness value can be interpreted as being equivalent to a prediction of the reliability with which the technician will perform his job. "Thus, ... effective performance can be called "reliable" performance and... ineffective performance... can be termed "unreliable" performance" (p. 11, Ref. 3).

The personnel reliability index therefore describes the performance of the maintenance technician in much the same way that a probability value derived from THERP, for example, describes the performance of the operator in a system. (This applies of course only to the way in which the output measure of both techniques can be interpreted; the manner in which THERP and the personnel reliability index derive these output measures is entirely different.)

Despite the simplicity of the technique, the concepts underlying it (i.e., the meaning of the job factors being measured) are extremely complex, although superficially they are readily understandable.

The importance of the personnel reliability index is that it is the only technique we are aware of that attempts to predict the efficiency of maintenance technician performance. The other maintainability prediction techniques reviewed in this report either predict equipment or system downtime only or the probability that a failure will be detected and repaired during maintenance (ERUPT).

In the discussion that follows it is necessary to distinguish between (1) the technique for gathering the data on which the personnel reliability index is based and (2) the use presently being made of those data by Siegel and his co-workers in a digital simulation model to derive predictions of maintenance downtime (Ref. 2).

GOALS

The purpose of the personnel reliability method as an index of performance is to provide "feedback on the technical proficiency of Navy electronic maintenance personnel" (p. i, Ref. 3). As a performance evaluative measure it is "modeled after an equipment reliability index" (p. i, Ref. 3).

In descriptions of models reviewed previously we have suggested that a predictive model may have one or more uses or goals which can be summarized as follows:

(1) prediction of the performance effectiveness of the personnel component of a system;

(2) comparison of alternative system configurations in terms of the personnel performance effectiveness anticipated with each configuration;

(3) use of the technique in design analysis, i.e., to suggest new design possibilities or redesign of a system configuration found inadequate in predicted performance;

(4) use of the technique to suggest manpower selection requirements, e.g., number of personnel required and aptitudes/skills needed;

(5) use of the technique to establish training requirements, i.e., suggestions as to the amount of training, its duration and training content.

The goals of the personnel reliability index as a performance evaluative measure have been described in terms of the following questions which it will answer:

"1. What is the current level of effectiveness of the maintenance personnel in a given rating, ship or squadron?

2. How does the maintenance personnel effectiveness level of a given rating, ship or squadron compare with that of other ratings, ships or squadrons?

3. Why is the maintenance effectiveness of a given rating, ship or squadron low or high?

4. Which specific job skills need improvement within a rating, ship or squadron?

5. In what maintenance areas is more training needed?" (p. 1, Ref. 3).

Although these goals appear limited to measurement of current effectiveness, it is legitimate to think of the index as a predictive device because of its output measure "which is interpretable as a probability of performance" (p. 65, Ref. 3). From that standpoint it is legitimate to ask how well the index satisfies the uses listed previously. The answer to this question will be discussed in the section on Anticipated Model Uses.

A secondary (although not unimportant) use of the personnel reliability index is to supply data which can be included in a simulation model predicting the maintenance efficiency of the system. Siegel and his coworkers are presently validating the index in this regard with two naval systems. Since the simulation model is closely tied to the personnel reliability index, it is reasonable to consider also the goals of the resultant human reliability prediction technique, which are for

"1. predicting the maintainability of future systems;

2. the provision of significant design verification information...

3. the development of preferred methods of maintenance and use of equipment by operational commands." (p. 1, Ref. 2)

ASSUMPTIONS

The technique builds upon the work of Whitlock (Ref. 4) who "investigated the relationship between observation and performance evaluation. Whitlock pointed out that (1) performance evaluation represents a response to the observations of performance; (2) observations associated with performance evaluation are observations of performance specimens; and (3) observations of performance specimens can be remembered over reasonable rating periods and reported accurately at the end of the rating period. A reasonable rating period for Whitlock was up to six months in duration.

"Whitlock defined a performance specimen as "an incident of relevant performance which is uncommonly effective or uncommonly ineffective... Regarding the definition of uncommon performances, Siegel and Pfeiffer (1966b) pointed out that this definition, in a sense, represents an adaptation of Flanagan's critical incident technique..." (p. 3, Ref. 1), which is widely accepted as a means of eliciting information.

A number of studies were performed by Siegel and his co-workers which eventually resulted in the personnel reliability index. These studies examined the nature of the scales underlying the index and the correlations between the index and other measures. For example, Siegel and his co-workers isolated various job activity factors descriptive of the Naval avionics job by means of multi-dimensional scaling analysis. These factors are listed in Table 13. Siegel and Pfeiffer (Ref. 1) found a correlation of .73 between peer estimates of personnel proficiency and peer estimates of job proficiency. There is consequently a substantial basis of research to support the index.

Certain assumptions are inherent in the index methodology and should be examined.

One major assumption is that the remembered effective and ineffective episodes represent a valid sample of the entire body of either the observed or the observable performance. The two performance categories (observed/observable) must be distinguished. Observable performance is the totality of the technician's behavior, if one were in position to view it. That part of his total (observable) performance which has been seen by others judging his performance represents the observed performance. Tne episodes reported as being uncommonly effective and uncommonly ineffective represent that part of the observed behavior which was recalled by the judge.[1]

1. Siegel adds (personal communication): "However, Siegel and Federman (Ref. 3) demonstrated that the data elicited are reliable (repeatable). Hence, whatever the judges remembered was similar for two different time periods. Moreover, the use of multiple judges in developing the data base tends to minimize these effects on the data."

TABLE 13. DEFINITIONS OF JOB ACTIVITIES

1. Electro-cognition--includes the following type of activities:

    a. maintenance and troubleshooting of electronic equipments
    b. use of electronic maintenance reference materials

2. Electro-repair--includes the following activity:

    a. equipment repair in the shop

3. Instruction-- includes the following activity:

    a. teaching others how to inspect, operate, and maintain electronic equipments

4. Electro-safety--includes the following activity:

    a. using safety precautions on self and equipment

5. Personnel relationships--includes the following activity:

    a. supervising the operation, inspection, and maintenance of electronic equipments

6. Electronic circuit analysis--includes the following type of activity:

    a. understanding the principles of electronic circuitry
    b. making out failure reports
    c. keeping records of maintenance usage data

7. Equipment operation--includes the following type of activity:

    a. operating equipment, electrical and electronics test equipment, and other electronic equipments

8. Using reference materials--includes the following type of activities:

    a. use of supporting reference materials
    b. making out reports

9. Equipment inspection--includes the following type of activity:

    a. supervising and performing inspections of electronic equipments

Since the technique depends on observer reports, its adequacy is dependent on the ability of the respondents to analyze the performance observed in terms of the job factors, and to structure the report of their observations.

A number of factors may create difficulty in securing accurate responses to the index:

(1) Lack of opportunity on the part of judges to observe all relevant aspects of t e technician's performance. However, use of multiple judges may mitigate this problem.

(2) Difficulty in ascertaining which aspects of the technician's behavior are indeed relevant to his overall performance; in other words, if the technician is consulting blueprints, for example, which of the 9 job factors does this behavior apply to? Siegel's most recent work, however, suggests that it is possible to make this assignment.

(3) Failure to remember these relevant aspects;

(4) Difficulty in differentiating between what is merely average or reasonably effective performance and that which is uncommonly (outstandingly) effective or ineffective.

Siegel points out, however, that if there has been no opportunity to observe, the judge so notes; that lack of an opportunity to see some performance should not influence the validity of the index measure since missed UE and UI behaviors would be distributed in the same manner as these seen; that it is relatively easy to differentiate UE and UI behaviors and that no judges in his data sample had difficulty with the concept.

If one has difficulty in accepting observer reports as a measurement technique, it must be pointed out that magnitude estimates of observed performance are no more subjective than those involved in any other psychophysical technique. Siegel points out, moreover, that observer reports have historically produced quantifiable, valid data.

Another assumption implicit in the technique is that the nine job factors listed in Table 13 adequately reflect the Naval avionics maintenance job. The degree of confidence one may have in these job factors is enhanced, because they were derived from a factor analytic methodology; and indeed they seem to encompass most of the activities involved in avionics maintenance, for that matter, in most maintenance jobs. Although they differ somewhat from the tasks, functions or stages which are commonly used to describe the maintenance process, e.g., diagnosis,

testing, removal, replacement, etc., there is no reason why these behaviorally derived factors (which, moreover, represent elements or skills underlying maintenance jobs) should correspond to the task categorizations found in non-behavioral models.

It is assumed in any event that these job factors adequately represent job or task performance. This implies in measuring performance effectiveness that the idiosyncratic characteristics of the individual task (e.g., the design of the equipment on which the task is being performed, the context of task performance) are relatively unimportant in influencing the quality of that performance, although these characteristics may well influence the time needed to perform the maintenance.

Because these are factors rather than tasks, they are phrased in rather general terms. This may actually be advantageous; because of their generality they can probably be applied to any maintenance job with relatively little change in their descriptions. On the other hand, one consequence that may result from the generality of these factors is a limitation on the specificity of the conclusions and recommendations that can be derived from their measurement, if these are to be applied to specific tasks. The problem that may arise when the data must be applied is precisely how one defines the relationship between the job factor and the job.

Since we have 9 factors, each of which gives us a score between 0.00 and 1.00, they must obviously be combined. Combination is by multiplication:

$R = r_1 \times r_2 \times r_n$, where

R = series reliability, and

r = reliability of each job activity.

However, the methodology does take into account conditions under which there are redundant and parallel operations. In its application to the digital simulation model, Siegel notes that "we are not limited to an "and" and "or" logic. Statements could conceivably be connected by conditional or biconditional symbols. These in turn can be expressed in terms of "and", "or", and negation." (p. 7, Ref. 2).

To return, however, to the original job factors: there is an implication that the performance reliability of each job component is independent

of other job components. Although this is a necessary consequence of the multi-dimensional scaling analysis, the factors as they exist in real life may function interdependently. Attempting to include the element of interdependence in the index would, however, overly complicate the measurement.

The index is represented by the following formula:

$$\frac{\Sigma UE}{\Sigma UE + \Sigma UI} \quad \text{where}$$

UE represents uncommonly effective behaviors and
UI represents uncommonly ineffective behaviors.

Thus, UE - UI represents the totality of observed performance. The index yields a value which varies between 0.00 and 1.00, which is nice because it can then be combined statistically with conventional measures of equipment reliability which also vary between 0.00 and 1.00.

The combination of personnel reliability probabilities with equipment reliability probabilities may present a problem (although only theoretically) because of what the respective probabilities represent. Equipment reliability represents (in most cases) equipment performance varying as a function of time, whereas, whatever a personnel reliability probability means, it does not represent that quality as a function of time. This is a problem we have encountered before with reliability-oriented predictive models. Can two probability measures be combined, even though they measure different qualities in differing ways? What will the combined measure mean? Again, this difficulty, if indeed it is a difficulty, is inherent in the differing nature of behavioral and equipment measures.

METHODOLOGICAL SCOPE

As presently developed, the technique is based on avionics maintenance activities. However, because we are dealing with maintenance factors of some generality, it should be possible to apply it to any type of maintenance.

Moreover, there would seem to be no reason why the technique should not be expanded to operator factors and activities. Since the technique seems to be particularly sensitive to skills and training, whereas the other operability prediction models are not, the application of the

personnel reliability methodology to operator factors would seem to have some point.

Through the description of its factors, the technique "models" the set of activities to which it is applied, but models it so generally that the limitations found in other techniques relative to continuous and decision-making activities would not seem to apply. Presumably a new set of operator factors would have to be derived via multi-dimensional scaling and new data collected via the critical incident method if one wished to apply the methodology to operator tasks.

MODEL PARAMETERS

The personnel reliability parameters are the 9 factors which were derived by multi-dimensional scaling. As is obvious from the descriptions of these factors (Table 13), they are rather general. However, the results in reference 3 (p. 37) indicate that they appear to be significantly differentiated among themselves. Moreover, there is considerable internal consistency within any single factor rating.

DATA

1. Input Data

As indicated earlier, the two input measures required are the frequency of uncommonly effective and uncommonly ineffective behaviors for each of the nine factors. Neither of these measures is particularly well defined for the respondent, but Siegel indicates that respondents had no trouble with them. (It might be interesting-- purely from a research standpoint-- to have respondents describe the events which they consider UE or UI, to see if, in the opinion of others, these really reflect significantly different behaviors.) In any event, the measures do discriminate among ships, ratings and individuals.

2. Data Sources

These consist of frequency counts gathered from supervisors. No external data store (e.g., experimental literature) is required. In fact, the index reliabilities serve as a data store themselves, directly translatable into task performance probabilities.

3. Output Data

This is the ratio, UE/UE + UI, which yields a numeric indicating a probability of effective performance for the technician in terms of the specific job factor. These individual job factors are then compounded to produce a total effectiveness value for the technician (see the discussion following). The ratio may be used in two ways: (1) as a measure of effective general performance (e.g., sailors are 85% effective in performing maintenance as a whole); (2) when applied to specific tasks, the index yields the probability of accomplishment of those tasks. The first output represents a sort of capability metric independent of specific tasks or equipments; it is purely a measurement output, not a predictive one.

PROCEDURES FOR APPLYING MODEL

1. Analytic Method

One must consider two analyses, the first relating to the procedure whereby responses for the personnel reliability metric are gathered, the second relating to the application of personnel reliability data to the digital simulation model.

The analysis required for securing responses to the personnel reliability metric is performed solely by the respondent. He decides what the criteria for ineffective and effective behaviors are,[1A] relates them to the tasks performed for given equipment, and then recalls how a particular individual in the past performed these behaviors.

Although the validity of the responses secured in this manner has been experimentally demonstrated by the Whitlock study referenced previously, the author confesses to a lingering uneasiness about the amount of error that may enter into such judgments when they are elicited from operational personnel in the operational environment. We understand that the digital simulation maintenance study presently being conducted by Siegel will include a validation of the judgments made by personnel in comparison with their actual performance on maintenance tasks performed in an operational setting. In any event, to ensure valid and reliable results with the technique, it is necessary to ensure that personnel making these judgments clearly understand the effectiveness criteria to be applied, the operational meaning of the job factors, etc.

1A. Based, of course, on the definitions of the 9 job factors.

In the application of personnel reliability data to the simulation model, it is necessary to decide which factors are demanded by each task. There should be little difficulty in making this judgment. However, since the simulation model operates on the basis of subtasks, and the time input data refer to subtasks, the reliability index values referring to tasks have to be broken down into sub-values for each of the component subtasks. In other words, if job factor X (index value of .83) describes task Y, and task Y is composed of subtasks 1, 2, 3 . . . . n, it will be necessary to subdivide .83 among the subtasks of task Y. This subdivision has already been performed in exercising the simulation model.

2. Synthesis Method

As indicated previously, the combination of reliability values for each factor yields a total personnel reliability value for the individual technician. Because of the multiplicative process "a depressed reliability coefficient for any job activity will result in a depressed reliability coefficient for the composite value" (p. 11, Ref. 3). This is a disadvantage which one finds in the other reliability-oriented techniques.

In the case of the personnel reliability index part of the problem arises from the fact that each job factor apparently has an equal weighting in terms of its effect upon maintenance effectiveness. It is difficult to see how certain job factors like failure to observe safety precautions, failure to complete failure reporting forms or inadequate personnel relationships could have as much an effect as failure to make an adjustment or wiring a circuit incorrectly. The former factors would make maintenance performance less efficient and might increase the time spent on maintenance, but their effects would be less serious than those of the latter factors.

This point may not hold because of the complex relationship between the job factor and the job. Ideally, however, the relative importance of the job factors would be reflected (perhaps by some sort of weighting index) in the process by means of which they are combined to form the total personnel reliability value. We realize, however, that the introduction of an additional weighting factor might overly complicate the technique.

Within the model synthesis is performed according to any of three modes: series, parallel and series-parallel situations. The series

combination mode is completely multiplicative. In the series situation all activities must be performed satisfactorily. "The assumption that all decisions in series must be performed satisfactorily implies that if a wrong decision is made by an operator, he will not realize that the decision was wrong until... the whole task is performed unsatisfactorily." (p. 12, Ref. 2) This assumption makes the author uneasy, however, Siegel's recent data suggest the assumption is tenable; moreover, it simplified the combinatorial process.

However, "It is often possible to improve the results by repeating a process or by calling on someone else to correct deficiencies..." (p. 12, Ref. 2).

A parallel situation exists when a task is performed satisfactorily if either one or another (or both) activities is performed satisfactorily or else the same job activity is performed by two men and acceptable performance of either man will constitute acceptable performance for the team. Here the probabilities are combined according to the formula $1-(1-P_1)(1-P_2)$. The series parallel situation represents a combination of the preceding.

3. Data Application

The UE/UI + UE events that make up the index do not refer to individual tasks but to a complex of activities which come closer to the concept of "function" than to that of the task. The index, once derived from the combination of these UE/UI events, is applied to or translated into something called a task. The task is interpreted in terms of the job factors involved in or required by performance of the task; then the index values associated with each of the factors in the task are combined to predict the estimated performance of the task. The index data therefore are ap ied to the task (which may vary in terms of complexity-- the size of the task unit is not specified), although the basis of the initial data is not the task itself. The author has pointed out previously that there may be difficulties in equating a job factor with a task.

ANTICIPATED MODEL USES

1. Prediction of System Effectiveness

The personnel reliability index itself does not predict system effectiveness. Initially it provides an evaluation of the performance of the

personnel about whom the data were gathered. However, the index values can be used for system-predictive purposes when they are translated into task-equivalent probabilities as described previously. In this application the set of index values can be used essentially as a data store for use in a simulation model which, when exercised, would predict the maintenance task performance of the system. This is the research presently being conducted by Siegel and his co-workers.

Therefore, with regard to the prediction of system effectiveness, it is possible to make the following statements:

(1) The index itself can be used for this purpose only when maintenance of electronic equipment is the effectiveness parameter of interest. Conceivably, in much the same way in which one applies data bank probabilities to the non-simulation operability models (e.g., THERP, TEPPS) one could (a) take a system description in terms of required maintenance tasks, (b) analytically extract the job factors involved in these tasks, (c) apply the index values already gathered to these factors and (d) recombine these factor-task values to produce task-equipment probabilities of performance. One might then be able to make a statement such as, System X has a predicted 85% probability of maintenance effectiveness. If one had a system requirement specified in terms of maintenance effectiveness, one could then compare the 85% value with the requirement.

However, maintainability requirements are generally phrased either in terms of mean time between failure (MTBF) or a maximum downtime or maintenance task time. All of these are formulated in terms of time, which the personnel reliability index does not provide.[2] Consequently, some other means must be adopted to transform the probability of maintenance effectiveness values supplied by the personnel reliability index into time values. This can be done by utilizing the personnel reliability values as inputs to a simulation model of maintenance performance. Again, this technique is currently under test by the developers of the personnel reliability index technique.

(2) In the form of a data input to a simulation model of maintenance effectiveness, the personnel reliability index would seem to be useful as a predictor of system (maintenance) effectiveness. The extent of the adequacy of this application waits upon the results of the study presently being conducted.

2. Siegel notes however that system requirements might be stated in terms of personnel reliability index requirements, which would eliminate the difficulty.

2. Design Analysis

We include under this heading several aspects:

a. The comparison of alternative system configurations to determine which should be selected;

b. The determination of redesign requirements.

Again it is necessary to look at the personnel reliability index in two ways: as a measure of performance and as a data input to a simulation model of the system.

(1) As a performance measure. As described on the previous page, one could apply the job factor values secured from previous studies to the analysis of the maintenance tasks in each of two projected system configurations and then compare the resultant task-equivalent probabilities, selecting the one with the higher probability. However, because of the generality of the job factors describing the maintenance tasks making up the system, it is possible that differences in various configurations will not easily show up, where these are differences in terms of equipment characteristics alone. Two distinctly different equipments could both require inspection or operation or electro-repair; hence their estimated performance probabilities will not differ. Only if the job factors involved in the two equipments are distinctly different will differences in system performance be manifested. Siegel correctly points out that the design differences would probably be reflected in the time for malfunction correction which would be handled in the second use of the index, as an input to the maintenance simulation model (see below).

Nor is it likely that the personnel reliability index values would lead to the determination of very specific redesign requirements for a system which is either in test or in operation, because the job factors inherent in the index are not intended to be particularly sensitive to equipment differences. This is not to say, however, that no redesign requirements could be derived from the application of the index to these systems, but these requirements would probably result more from an examination of the system design than from the index itself.

(2) As an input to the simulation model, it is expected that the simulation model will be more sensitive to design differences and thus permit a meaningful comparison of alternative maintenance configurations

or procedures. However, it is likely that this sensitivity will arise not especially because of the personnel reliability inputs to the model but because the model exercises the system, and this exercise permits configurational differences to show up more clearly.

3. - 4. Selection and Training

The personnel reliability index appears to have significant potential in the areas of selection and training. The generality of the job factors, which may be a disadvantage for design, now becomes highly advantageous for selection and training, because the factors are specifically oriented to personnel capabilities. This advantage applies more to training than to manpower selection. If the probability of performing a task, which is heavily loaded on electro-cognition and personnel relationships, is low, then presumably more training should be given in these areas. One must relate the deficient job factors to the particular knowledges and skills that must be learned, but this should present few difficulties. As far as selection is concerned, we run into some problems because we do not know what measurable aptitudes define the electro-cognition and personnel relationship factors. This, however, simply reflects the inadequate state of the art in manpower selection and aptitude classification.

VALIDATION/APPLICATION STUDIES

The personnel reliability index is one of the few models reviewed in this report which has been subjected to formal validation studies. Correlations have been obtained with intermediate criteria such as the TPCF, GCT, ARI, skill, time in service, etc. (see Ref. 1, p. 59), but these correlations were low (a multiple correlation of .40 was obtained), although apparently correlation coefficients of this size are as much as one may reasonably expect.

Validation of the personnel reliability index by comparison with actual maintenance performance has not yet been performed, but is to be determined as part of the simulation study by Siegel already referred to.

The index has not to this author's knowledge been utilized in actual system development, but this is probably because it is still in research status.

Despite the questions raised in the preceding pages, there is a great deal to hope for from this model. First, as was pointed out previously, it is the only technique which behaviorally attempts to measure and predict the technician's maintenance effectiveness. Secondly, the technique is comparatively simple to use in terms of gathering additional data on maintenance effectiveness. Third, it may help in the development of a maintenance simulation model which is something we have not had to date. We have also pointed out that perhaps the methodology has applicability to operator as well as maintenance job factors.

EVALUATIVE SUMMARY

Validity - Formal tests have been and are being performed; results show reasonable correspondence between predicted and observed values.

Reliability - Formal tests have been and are being performed; reliability appears high.

System Development Applicability

A. Comprehensiveness: Although presently applied only to avionics maintenance, the methodology would appear applicable to all types of maintenance.

B. Applicability: Given that maintenance efficiency is the system effectiveness parameter, the index both measures and predicts that effectiveness. In addition, it appears to have significant potential in the areas of selection and training.

C. Timing: Model can be applied at all stages of system development in which detailed maintenance task descriptions are provided.

Model Characteristics

A. Objectivity: Index depends on observer reports. However, method of eliciting these is explicit.

B. Structure: Conceptually well defined, despite some difficulty in equating the job factors with maintenance tasks.

REFERENCES

1. Siegel, A. I. and Pfeiffer, M. G. Post-Training Performance Criterion Development and Application. Personnel Psychophysics: Estimating Personnel Subsystem Reliability through Magnitude Estimation Methods. Applied Psychological Services, 1966.

2. Siegel, A. I. The Applied Psychological Service's Program Plan for Developing a Human Reliability Prediction Method. Paper presented at U. S. Navy Workshop on Human Reliability, Washington, D. C., July 1970.

3. Siegel, A. I. and Federman, P. J. Development of Performance Evaluative Measures. Report 7071-2, prepared for the Office of Naval Research, contract N00014-67-00107, September 1970.

4. Whitlock, G. W. Application of the Psychophysical Law to Performance Evaluation. J. Applied Psych., 1963, 47, 15-23.

SECTION IV

SUMMARY AND CONCLUSIONS

In this section we consider what we have learned from the preceding review of the various models.

Common Elements

Despite individual variations in models, a single life line, as it were, connects them all. That line is the procedure employed by the model-user in applying a model to make a prediction. It must, however, not be imagined that every step in this procedure is consciously, explicitly performed, or that each step is always discrete.

(1) Determination of the evaluative (predictive/descriptive) measure desired. This involves selection of a model which supplies that output measure. Whether determination of the desired output measure precedes model selection, or the model user selects a model and simply accepts its output measures, depends on the user.

(2) Analysis (usually by means of function/task analytic methods) of the system whose performance is being predicted, to determine the behavioral units (e.g., subtasks, tasks, functions) to which the prediction will be applied.

(3) Analysis of those behavioral units to determine which parameters must be considered in making the prediction.[1] This step is often performed implicitly, so that only a few most important parameters are considered.

(4) Assignment of input data to the behavioral units being predicted. This involves a series of steps:

---

1. Among the parameters that may be considered in terms of their possible impact on unit performance are: number of identical components from which the control to be activated or the display to be read must be selected; organization of the controls and displays; presence or absence of feedback information; response pacing; required accuracy of response; display exposure time; type and number of stimuli presented; function performed by the operator. This list is illustrative only.

(a) Determination of the data sources available for the prediction, which involves matching (explicitly or implicitly) certain parameters describing the behavioral units (i.e., those found in Step (3)) with the parameters implicit in the data source. Again, because of inertia, available data sources are often accepted without examination of their parameters.

(b) Selection from the data source of the data items needed to make the prediction (obviously not every item in the data source is applicable, because behavioral units vary in terms of the functions, equipment objects, etc. which they describe).

(c) Application of the selected data items to the behavioral units.

(5) Exercise of the human reliability model to derive the desired predictive output. This exercise may occur in two ways:

(a) Combination (usually by means of probability statistics) of predictive values for molecular behavioral units to derive predictions for more molar units containing two or more of these molecular units (analytic methods).

(b) Simulation of the behavioral operations described by the model (simulation methods).

(6) Combination of the derived terminal predictive value for the operator (maintenance man) subsystem with the terminal predictive value for the equipment subsystem to achieve an overall system reliability prediction. For example, to illustrate the process very simplistically, if operator performance in a system is estimated to be .98 and equipment performance in that system is estimated at .99, the resultant system reliability value is .98 x .99 = .97. It is our impression that although a system reliability prediction is the logical end goal of the process, this combination is not often performed, perhaps because of lack of confidence in the validity of the human reliability value.

Another common element that should be discussed are the behavioral models that are implicit in various methods. As we pointed out at the beginning of this report, only a limited number of these methods contain models (in the Chapanis sense) of man-machine system (MMS) operations. The most outstanding work in such model development has been performed

by Siegel and his co-workers, although each method implies a concept of how the MMS functions even if it does not explicitly describe that concept. And although it is unnecessary that a methodology include an explicit model of system operations, it is our impression that the more sophisticated methods do.

One can therefore look at these methods not only in terms of their direct application to the solution of practical MMS problems, but also in terms of what they reveal of Human Factors concept-building. This requires a somewhat more theoretical viewpoint.

The broad outlines of the MMS concepts we derive from an examination of the methods include the following propositions:

(1) A system requirement exists which is the "forcing function" for system operations. This requirement may be a minimum number of outputs, a maximum time in which responses must be made, a probability of detection, etc. Since the MMS is an artificial construction (i.e., not found naturally), the system developer specifies this requirement.

(2) The requirement sets a goal for the system and for the operator as part of that system; system operations (including those of the operator) are performed in order to achieve that goal. What is implied here is that idiosyncratic operator behaviors must be subordinated to the system requirement if the goal is to be achieved. The system requirement therefore constrains the individual's freedom of function and in fact molds his behavior by eliminating system-irrelevant responses. The implication is that as the system requirement demands greater precision, faster responses, etc. the freedom of the individual to do "his own thing" becomes less. The criteria of operator adequacy are those of the system: the operator is efficient only when he helps to achieve the system requirement.[2]

2. This may sound to some readers like the theorizing to be found in Orwell's 1984. It is possible (although the author does not) to defend the point of view that as society becomes more technologically oriented, requiring more and complex MMS, it is inevitable that the human will become increasingly constrained by MMS demands. One can also view the Human Factors discipline in a somewhat broader, sociological sense as an attempt to interpret the mechanisms of an increasingly technological (i.e., MMS-oriented) society.

(1) The effort to accomplish the MMS requirement creates certain stresses in the system as a whole (e.g., queuing) but particularly in the operator (who has more flexibility than his equipment), stresses which influence his behavior by increasing or decreasing response rate, increasing or decreasing his error production, elimination of non-essential responses, etc. In some models (e.g., Siegel's) stress is considered as an organizing, positive parameter as well as a counter-productive one. However, failure to achieve the system goal intensifies these stresses. [3]

(4) Adequacy of system design is conceptualized in terms of satisfaction of system requirements. Repeated failure of the operator to accomplish system goals implies a system inadequacy which must be remedied by some redesign activity, either of the individual, his training or his equipment mechanisms. Initial design of the system is performed by selecting those mechanisms (e.g., personnel skills, functions, tasks, equipment components) which will lead to maximum probability of goal accomplishment.

Since the purpose of system development is to increase this probability of goal accomplishment, a human reliability methodology is most effective when it can be utilized in the initial design of the system, to select those personnel whose behavior can be most readily manipulated, to determine what training is needed for the manipulation, to select that man-machine configuration which comes closest to meeting the system requirement, and to predict the operator's ultimate performance in relation to system goals. The reader will recognize here the various objectives with which the various models/methods were compared.

To help our summary further a matrix chart (Table 14) has been constructed to summarize the similarities and differences among the models. The categories around which the matrix has been organized will be explained in the remainder of the section, but the reader will see that they are similar to the categories used in describing the individual models.

3. It is interesting that in these models stress is determined exclusively by system factors; the internal genesis of stress is largely ignored.

Before proceeding to describe what Table 14 indicates, it should be explained that this chart is not an evaluative device; it is simply an attempt to summarize the various methods and their characteristics. Necessarily such a summary, simply because it categorizes (lumps) characteristics, tends to obscure details; consequently in the discussion below we have drawn on the individual model descriptions in the body of the report.

Simulation/Analytic

Our first category is the differentiation of the models into those that apply simulation and those that do not. As was pointed out in the section dealing with ground rules for data bank development, there are other ways of categorizing the methods reviewed — e.g., descriptive vs. predictive or predictive vs. evaluative — but we feel that the simulation/analytic dichotomy is the most significant. A simulation technique reproduces (via computer) the operations actually involved in a mission and secures its output measure as a result of those reproduced operations. An analytic method applies data to the operations required by a system without attempting to reproduce their functioning. These categorizations overlook, of course, major differences among both types of methods.

Although this is a highly significant difference, examination of Table 14 reveals that the two types of models also possess many elements in common, e.g., the uses to which they are applied, their task scope, the task analytic process, their output metric, etc.

Twelve of the methods reviewed are analytic techniques; six utilize simulation.

It is obvious from the nature of the simulation process that the methods making use of it are much more powerful than the analytic methods. The simulation process permits the determination of functional relationships which the non-simulation methods cannot provide. This is because the simulation partakes of the character almost of an experimental test. The

TABLE 14
SUMMARY OF MODEL CHARACTERISTICS

Descriptive Categories	Air Data Store	THERP	TEPPS	Pickrel/McDonald	Askren/Regulinski	Berry/Wulff	Throughput	DEI	Personnel Part Metric	Digital Simulation	TACDEN	Boolean Approach	HOS	ORACLE	Personnel Subsystem	ERUPT	Maintainability Prediction	Personnel Reliability
1. General Classification																		
Simulation									X	X	X	X	X	X				
Analytic	X	X	X	X	X	X	X	X	X						X	X	X	
2. Model Uses																		
A. Prediction	X	X	X	X	X	X	X		X	X		X		X	X	X	X	
B. Evaluation	X	X	X	X	X	X	X	X	X	X		X		X			X	
C. Design Comparison	X	X	X	X	X		X	X	X	X		X	X	X				
D. Design Analysis									X	X	X	X	X	X				
E. Selection/Training												X					X	
F. Personnel Standards			X															
3. Model Scope																		
A. All Tasks/All Systems		X		X	X	X	X		X	X	X	X	X	X				
B. System-Limited								X	X									
C. Discrete Tasks Only	X		X															
D. Maintenance Only															X	X	X	
4. Input Data Sources																		
A. All		X		X					X	X			X	X				
B. Experimental/Empirical only	X				X	X	X		X		X	X				X		
C. Subjective Only			X														X	
D. Other								X							X			
5. Input Data Detail																		
A. Very Detailed	X	X		X	X	X	X		X	X	X	X	X	X	X	X		
B. More Molar		X	X											X			X	
C. Not Applicable								X										

TABLE 14 (continued)

Descriptive Categories	Air Data Store	THERP	TEPPS	Pickrel/McDonald	Askren/Regulinski	Berry/Wulff	Throughput	DEI	Personnel Perf. Metric	Digital Simulation	TACDEN	Boolean Approach	HOS	ORACLE	Personnel Subsystem	ERUPT	Maintainability Prediction	Personnel Reliability
6. Behavioral Unit Employed																		
A. Subtask or S-R Unit	X	X		X	X			X	X	X	X	X	X	X		X		
B. Task		X	X	X	X									X				
C. Function															X		X	
D. Not Applicable						X	X	X										X
7. Analytic Method																		
A. Task Analysis	X	X	X	X		X	X			X	X	X	X	X		X		
B. Other Methods															X		X	
C. Task Analysis Not Needed					X			X	X									X
8. Use of Combinatorial Statistics																		
A. Yes	X	X	X	X	X									X	X	X	X	
B. No						X	X	X	X	X	X	X	X					X
9. Output Metric																		
A. Prob. Successful Perf.	X	X	X	X	X	X				X	X			X				X
B. Response Time	X	X	X	X						X	X		X	X		X		
C. Other							X	X	X			X			X			
10. Validation/Application Data																		
A. Formal Validation Tests	X							X		X	X						X	X
B. Partial Data Available		X	X										X					
C. None Available				X	X	X	X		X			X	X	X	X			

simulation also permits a detailed examination of how the operations were performed — a sort of "time-history".

If the simulation/analytic dichotomy were the only basis for selecting a predictive methodology, the simulation methods would win "hands down". The question arises, however, whether for a particular system development need the user requires such a powerful tool. It is impossible to tell because no one has ascertained the precise nature of this need.

(Note that we do not consider here the research interest in man-machine modeling, which is obviously important in its own right. We are concerned here only with the solution of system development problems.)

The situation is complicated by the fact that there is not one user but several. The governmental planner or anyone else involved in the very early planning stages of a system (e.g., SOR, ADO, etc.) might wish to consider the factors that a simulation model like Siegel's can handle best, i.e., where functional relationships are involved and the planner wishes to trade off various values of the relationship. For example, in planning the crew requirements for a new destroyer class like DX one might wish to explore the relationship between number of personnel required, work-rest cycle and work output. In the later stages of design a designer might wish simply to compare alternative design configurations or secure an estimate of the error anticipated with a given control panel. In this case a non-simulation method might provide him with a perfectly acceptable answer.

We could be wrong, of course. More powerful methods might be preferred by all users at all system development stages, but the question remains open.

Theoretically some very different models provide essentially equivalent outputs. If one ignores the functional relationship factor which marks the simulation methods, and the fact that some simulation methods handle certain parameters that non-simulation methods do not, both types of models provide many of the same outputs, i.e., probability of success and response time. (However, the simulation models provide a greater richness of output, e.g., standard deviations, variability over time.) Certainly the developers of the various models would claim that their models can do everything that any other model could do. In the absence of data based on a formal application test comparing the effectiveness of the various methods, the choice of method may well devolve on the specific needs of the user and the relative cost-effort involved in applying the different models.

These utilization factors were generally given a much lower weighting in the ranking of man-machine evaluative criteria.

Nevertheless, however much one might consider them peripheral compared to factors such as validity and reliability, they could be crucial in the acceptance by the user of a particular method. System developers prefer to work as little as possible to secure their answers and desire them as quickly as possible. Unfortunately, very little data are available concerning the length of time required to gather data for a particular model, the number of man-hours required, time required to exercise the model, etc. In general, as we shall see later when we consider validation/application, not much is known about the mundane problems of applying a model to specific system development questions.

Consequently the evaluations we can make of the various methods at this stage of our knowledge cannot be considered final.

Model Uses

Six possible model uses are listed in Table 14. Prediction involves an estimation or anticipation of the future performance of a system. Evaluation is an assessment of the present state or capability of a system. However, a prediction can also be used to evaluate a system in the design stage, given that some explicit or implicit standard for evaluation exists. For example, if a system in design has an anticipated (predicted) probability of successful performance of .99, and if .99 is considered satisfactory by the system developers, then the system has also been evaluated. As can be seen from Table 14, prediction is closely linked to evaluation, in the sense that the same method used for prediction can also be used to evaluate a system in operation.

It was necessary to differentiate between these two because certain methods can perform one or the other function, but not both. For example, the personnel performance metric of Brady is purely an evaluative method, since it can be applied only to a system in operation. On the other hand, the maintainability methods currently in use in DOD are strictly predictive; maintainability evaluations employ different methods (not considered in this report).

Most of the models can be used also to compare design configurations, since a prediction applied individually to two or more configurations automatically implies a comparison. It should be noted, however, that this use is probably more potential than actual. Most models can be used for this purpose, since a comparison can be made on any basis, provided the basis is relevant. However, we suspect that few models have been used to compare design configurations.

Most of the methods possess little or no capability in the areas of manpower selection and training (definitions of these were provided in the

individual model descriptions) despite the fact that claims for these capabilities are often made. In part this discrepancy may arise because the criteria applied in this report demand greater sensitivity to these factors in the models themselves rather than as deductions from model outputs. In other words, we feel that to be sensitive to selection and/or training, a model must indicate what capabilities should be selected or trained, rather than merely that more adequate selection or additional training is required. On that basis only a few of the models, i.e., the personnel reliability technique of Siegel and Wherry's HOS, seem to possess this sensitivity. It may be that the majority of the models available do not include parameters which are sensitive to these factors or it may be that a distinctly different type of model (such as the personnel reliability technique) is required.

We have similar difficulties in the category of design analysis, which we define as the capability of performing a detailed time-historical analysis of the system's operations leading to specific redesign recommendations. Manifestly, only the simulation methods permit such an analysis, because the analytic methods do not reproduce the individual operations involved in performing the mission. This may be considered one of the advantages of the simulation methods. It is felt that such a detailed scrutiny is required for a method to suggest redesign possibilities. This does not mean that the non-simulation methods cannot make such redesign suggestions; however, they are limited in this respect. Certain of the simulation methods (such as HOS and ORACLE) specialize in this; ORACLE, for example, produces no predictive figure of merit but is solely interpretative.

Only one of the methods reviewed, TEPPS, dealt with the determination of personnel standards by means of apportionment (its "derivative" use).

It is apparent therefore that almost all the models can perform predictive, evaluative and design comparison functions to some degree of efficiency. The question naturally arises, how well, but information on this point is not readily available, since many of these models have not been validated, and of those that were validated, the validations were not performed on the same test situation.

It would appear then that selection of an appropriate model would have to be based on factors other than their presumed capability to predict, evaluate or perform design comparisons. Each model has certain peculiarities that may render it more or less effective in a given situation. One of these peculiarities is the question of Model Scope.

Model Scope

Here we deal with the range of situations and behaviors which the method can handle. Although 11 of the 22 models presumably can deal with all tasks in all types of systems, this is actually not so. Most of these 11 models would have great difficulty dealing with maintenance operations, which is the reason we included a specific category of maintainability models. But a maintainability model can handle only maintainability. Other models are system limited. For example, DEI can be applied to the comparison of control panels only; HOS is limited to pilot models (at present); and the personnel performance metric to highly automated ground systems. Other models (AIR, TEPPS) admittedly can deal only with discrete tasks.

In selecting a model, therefore, it is necessary to ask: what kind of system is involved, what kind of tasks? The combination of model uses and model scope serves to bound the possible choices even more.

Several of the models are so closely related to others that the number of choices is further restricted. For example, there is not much to choose between TACDEN and Siegel's digital simulation model, and since the former was derived from the latter, the latter would be preferable. Similarly, the Pickrel/McDonald model merely elaborates on THERP and between the two one might well choose the latter, particularly in view of the much more frequent application of the latter. In other cases a model is merely in the formative state, such as HOS; or was never applied, like ERUPT or the personnel performance metric, and is consequently an unknown factor.

Input Data Sources

Here we were interested in comparing the various models in terms of the kinds of data sources they utilize. The categories here are (1) all, which means that any available source will be acceptable (although obviously certain sources are preferable to others); (2) experimental/empirical only, which we define as data from the experimental literature (e.g., AIR Data Store type) or from prototype or operational testing; (3) subjective only, i.e., expert judgments; (4) other, which may include reliability failure data or, as in the case of DEI, a scale applied to equipment characteristics.

From Table 14 we see that about a third of the models will accept data from any source, about a half will accept data only from experimental / empirical sources, while two accept only subjective estimates.

If we combine the "all" and "subjective only" categories, we see that approximately half the models make use of subjective data either as the preferred data source or as part of their data sources.

In view of the widespread use of subjective estimates, we consider, along with Blanchard, that it is highly desirable to investigate this method further, and our recommendations for needed research will include this suggestion. Only TEPPS includes a highly specific methodology for securing subjective estimates, although Swain indicates that he secures his estimates in a systematic manner (and we hope the others do too).

The point is that, in view of the fact that subjective judgments are widely used (presumably because other applicable data do not exist) and because in the case of TEPPS, even with a highly formal method of securing such judgments, inter-judge reliability was low, means must be found of making such judgments more reliable than they are presently.

It is not that, as Blanchard suggests (along with Burger et al. (Ref. 1) and Hanifan et al. (Ref. 2), subjective data are to be the preferred source, but that if models must make use of this source, we had better attempt to standardize and improve on these data.

Input Data Detail

It is interesting that most models require very detailed information, only TEPPS and the personnel reliability index requiring relatively molar data. This is undoubtedly linked to the fact that the behavioral unit to which the data are applied is in most cases, as we shall see below, the subtask, task element or S-R unit. This detailed data requirement has obvious implications for the development of any data bank. The "not-applicable" category refers to the fact that in the DEI the data are inherent in the design being evaluated and hence cannot be thought of as being "input" to the method.

Behavioral Unit Employed

As indicated above, the behavioral unit employed in most models is the subtask, task element or S-R unit. Because it is difficult in some cases to determine the precise behavioral level at which a model operates, we indicated for those cases, e.g., THERP, Askren/Regulinski, that they use both the subtask and the task. In only two cases (personnel reliability index and ERUPT) is the model concerned with anything as gross as a function. In a few cases (e.g., DEI) a behavioral unit is not involved.

Analytic Method

Almost all of the methods employed use some variant of function/task analysis, even when a model is derived from Operations Research concepts, i.e., ORACLE, or from maintainability concepts. All of the methods take

this analysis as a given, that is, they assume the user knows how to perform the analysis.

In consequence the nature of the parameters in relation to which the analysis is performed is rarely indicated. By this we mean that if one is to apply input data to a model, it is necessary to indicate whether such factors as display exposure time, accuracy requirements, time limitations, etc. (the so-called performance shaping factors of Swain) are relevant to the task being analyzed and must therefore be taken into account by the input data. In part this failure to specify input data parameters as part of the task analysis results from the fact that most data sources do not include data relative to many of these parameters. It is very probable that the model developer implicitly takes such parameters into account when gathering his data, but unless he specifies them formally as part of his model, it will be difficult to develop an appropriate data bank and for the user of the model to determine what data he needs.

Use of Combinatorial Statistics

The term "combinatorial statistics" refers to the use of probability statistics to derive a system figure of merit by combining output values for more molecular behavioral units to secure data for more molar behavioral units. Examination of Table 14 reveals that the use of such combinatorial statistics is largely confined to the analytic methods. Simulation methods do not have to "build up" data values by combination, since the simulated operations themselves produce the desired higher level outputs. Combination is necessary for the analytic methods. because they employ the subtask or task element as the basic behavioral unit.

In view of the difficulties encountered with such combination — the problems of dependence/independence, and conditional probabilities — the fact that simulation techniques do not require such combination represents a very definite plus factor for them.

Output Metric

There are two major outputs of most models, i.e., the probability of successful task or system performance and response time (time required to complete the task or mission). Understandably so, because these two represent the two measures most adequately describing the effectiveness of the system.

A number of what one might term "time-history" simulation models, i.e., HOS, ORACLE, Boolean technique, do not output a probability value

as a measure of system effectiveness. Indeed, such models disavow any desire to evaluate the adequacy of systems, although they can be used for this purpose. Such models are largely interpretative; one examines the time history of the simulation exercise to diagnose problems. Of course, a simulation model can also output an evaluative measure, e.g., Siegel's digital simulation model.

The ability to interpret a time history is of great value, but unless it is combined with a figure of merit measure, it has restricted usefulness.

Other models, e.g., DEI, personnel performance metric, the maintainability models, output other measures which are of specialised interest.

Validation/Application Data

Validation data may be derived from formal (experimentally controlled) tests; or validation results can be inferred from application of a method to a problem and seeing how well the problem is solved. We have referred to this second case as providing "partial data". Obviously a formal validation is to be preferred to an informal one or to application data. Or there may be no known efforts to validate or apply a model.

We have already pointed out elsewhere that where a model is derived from and finds its major application to system development problems, there may be some difficulties in setting up a formal validation test. This may account in part for the relative paucity of validation data for many of the models.

Nevertheless, the absence of validation or even application data for many models represents a serious deficiency in the government's support of such model development. Formal validation has been attempted only in the case of the AIR Data Store, Siegel's digital simulation model, personnel reliability index, DEI, Miller's TACDEN (which is actually a validation of Siegel's model) and the maintainability models. TEPPS was supposed to be validated, but the test conditions were such that one can think of this only as an application test. According to Swain, THERP has been validated, but this is essentially based on application experience, experience which, because of the security classification of the problems to which THERP has been applied, is not available to the general reader. The developers of ORACLE refer to good results in their application of the method, but supply no data.

All things considered, if one were to apply only the criterion of validation to the selection of a predictive method, Siegel's digital simulation

model would be selected without question. However, in view of the sparsity of such data for the other models, the validation criterion might be considered a somewhat unfair one to apply.

What this part of Table 14 shows is that if the government is to continue supporting the development of predictive models, it must insist on validation tests and supply the conditions under which an adequate validation can be performed. Otherwise it is essentially wasting its money.

The most effective validation is, as has been pointed out previously, a comparison of a prediction with data gathered from an operational exercise of the system or task to which the prediction has been applied. It may appear, as it has to a number of developers commenting on their model descriptions, that such a validation requirement is too stringent; that perhaps other validation measures, e.g., concurrent or construct validity, would do as well. However, these are only partial indications of validity and do not replace validation based on empirical comparisons. Again, it is up to the government to ensure that the validation is adequate to its purpose.

The conditions of such validation should, wherever possible, approximate those in which the model will be applied. By this is meant that all of the constraining conditions of actual model application, e.g., full range of tasks (not merely tasks selected as being most appropriate to the model), realistic time demands on supply of model outputs, etc., should be included in the validation. For it is entirely possible that a model will provide valid answers but be difficult to apply in the use situation. Users may say that they care nothing about a model except for its validity, but that is before they have to apply the model in their own shops. Consequently the validation situation should throw some light on the problems of application.

It would appear from a review of Table 14 as a whole that no one model satisfies all possible requirements and that a model will be effective depending on the use for which it is intended, the type of system and behavior for which it is specialized, the kind of measure it outputs, whether or not the user wishes to employ simulation, the level of detail which is required of the answer, etc. Consequently the selection of a model requires consideration of all these factors. There is no such thing as a general purpose model.

The fact that many of these models are specialized makes it very difficult to point out one model as being overwhelmingly superior to another. Certainly one model is superior to another for a particular application, but not when considered over the range of all possible uses.

Moreover, these uses are limited. It is apparent that while most models may do a pretty fair job of predicting and evaluating and possibly even comparing design configurations, they seem to be relatively ineffective as far as design analysis, manpower selection and training are concerned. It may be necessary to develop individual models or techniques for these uses.

All of the models are deficient in one respect or another. Obviously models should operate on the basis of distributions of input data, but no data bank presently available contains such distributions with the exception of the data used for maintainability predictions.

The assumptions on which many models operate, such as normality of distribution, relationships between stress and performance, etc. often lack verification.

The manner in which the measure of probability of successful task accomplishment is formulated (i.e., p = 1.0 - error rate) is overly simplistic and misleading.

Hence a great deal still remains to be done.

Although we have reviewed 22 models or techniques in this report, the reader must be aware that the number of these that are still viable is far fewer. A number of these models were not carried far enough to make them useful or are still in process of development to the point where they can be used.

(Why were they considered then? Because their concepts and approaches are still valuable as indicating ways of looking at the man-machine modeling problem. For example, certain of the models employ behavioral elements, while others do not. We have already mentioned the dichotomy between simulation and non-simulation models. Some models emphasize independence concepts, while other utilize dependency/independency relationships. Some models make use of at least some behavioral parameters in their assumptions, while others avoid them. Some models are evaluative, while others are strictly interpretative. These models also have a good deal to offer us in terms of their implications for input data banks to support them. Moreover, someone may be interested enough to pick them up and develop them further.)

Models that have not been fully developed (at least in the sense of being validated) are: Pickrel/McDonald, ERUPT; Boolean approach; Berry/Wulff; Throughput. Models still being developed are: HOS; personnel subsystem model. It is fascinating to ask oneself why those models that never reached fruition remained half-developed. There is a parallel here to the many military weapon systems that were cancelled before they reached the production or operational stage.

Conceivably some models may have been aborted because they were developed not in response to a specified need and in a systematic fashion, but in response to an unspecified research need or an idiosyncratic interest in a particular area by the investigator.

Perhaps the greatest value this review may have is in terms of suggesting an orderly way in which models should be developed. An outline of such an orderly procedure would include the following steps:

(1) Determination of the needs for and uses of the models.

(2) Determination of the parameters required by the model.

(3) Determination of the required characteristics of input data for the models.

(4) Development of the model.

(5) Formal validation testing of the models.

(6) Formal application testing of the models.

Of these steps the most important (because everything else depends on it) is the first one.

REFERENCES

1. Burger, W. J. et al. Validity of Expert Judgments of Performance Time. Human Factors, 1970, 12(5), 503-510.

2. Hanifan, D. T. and Knowles, W. B. Human Performance in System Effectiveness Modeling: Issues, Approaches and Critique. Dunlap and Associates, Inc., Santa Monica, California, December 1968.

SECTION V

DATA BANK DEVELOPMENT GROUND RULES

If one is the developer or user of a human reliability predictive model/technique, one of the basic questions he must ask himself is: what kind of a data bank do I need in order to supply input data to the model/technique? If one wishes to develop a data bank, it is necessary to ask: what kind of data (i.e., content and format) should that bank contain?

The preceding paragraph assumes that a data bank is needed as a source of behavioral input data to exercise or apply the technique selected. Only a few of the models reviewed in this report- personnel reliability index and DEI- can perform without the use of input data. These input data can be a standardized set derived from the experimental literature, like the AIR Data Store; or it can be the results of special mockup or prototype testing; or data gathered operationally on systems similar to the one for which a prediction is to be made; or it may be subjective estimates. All of these types of data can be utilized as data banks, and consequently the questions posed below apply to all of them.

The purpose of this section is to present certain questions the data bank developer/user should ask himself. The answers he receives will help to determine the characteristics of the data he will develop and/or use. The reason for including this section in a report devoted to the analysis of predictive models is that the data bank is a tool which can only be applied to a particular technique to satisfy some user requirement. A data bank (of whatever type) assumes some method of using the data it contains. Consequently the data bank configuration must be related to the types of models we have reviewed. Another way of saying the same thing is that a particular method or technique implies a particular data bank configuration.

Of course, one can fail to consider a possible model application and user need in developing a data bank, but under these circumstances it is likely that any use of the bank will be quite inefficient. Alternatively one could assume simply that the data will be applied merely with the aid of probability statistics, but even this simple assumption implies a model of the THERP-type. The concept of a data bank assumes or requires some consideration of contents, format and level of detail.

As a tool, the data bank serves some need on the part of the user. The more closely the data bank characteristics correspond to this need, the more cost-effective the data bank will be.

It is possible, of course, to think in terms of a so-called "universal" data bank which would then have the following characteristics:

(1) Both error probability and response time data (formulated in terms of known distributions and variances) for

(a) all human behaviors (motor, perceptual, cognitive)

(b) all types of tasks (e.g., continuous, discrete, operating and maintenance)

(2) data at various levels of behavioral units (e.g., function, task, task element)

(3) data at various equipment levels

(a) the component characteristic (e.g., number of scales on a display, length of switch arm)

(b) component type (e.g., switch, meter, knob, CRT)

(c) equipment assembly (e.g., control panel)

(d) subsystem and system

(4) data accounting for differences in what Swain calls "performance shaping factors", e.g., stress, motivation, skill level; differences in aptitude of various types; differences in training of personnel

(5) data organized in terms of significant "affecting (independent) variables" like exposure time, order of stimulus presentation, stimulus input rate

(6) all of the above interrelated by combination.

A primary question the developer must ask himself then is: am I attempting to develop a "universal" data bank or will I be satisfied with a bank having a more restricted scope? If he decides that he will develop the universal data bank, then further consideration of the points raised in this section is largely unnecessary, and certainly such a data bank would satisfy the requirements of all the models reviewed.

However, the task of developing a universal data bank is a very onerous one and we are not likely to see one completed in the very near future. A further question that bears on the advisability of attempting to develop such a bank is, does the model user need it? It is possible that because of the nature of the questions the user asks that some "degraded" form of data bank will be acceptable. In any event, if the question about the universal data bank is answered negatively, it is necessary to ask the further questions described below.

We have emphasized the "need" for a data bank having particular characteristics. There are two types of need: (1) the requirement of a particular class or individual model/technique for particular data; (2) the need represented by the particular use to be made of both the model and the data bank. The first need stems from the second; if the

user does not require a particular model/technique (if it will not solve the system development and use problems he has), then obviously the kind of data required by that model becomes unnecessary. Ultimately therefore one must revert to the potential use of the model and data bank to ascertain its requirements.

We assume that the models reviewed in this report sample approximately the various types of data needed and the way in which these data are manipulated. Even if these models do not represent all present and future models (since it is always possible that some have not been unearthed and others will be developed), they represent the various ways in which one can think about data usage. Therefore the characteristics of these models specify decision criteria which can be applied to the problem of selecting a data bank configuration.

The first thing we note about these models is that they can be categorized - grossly - in two ways: (1) simulation techniques; (2) non-simulation techniques. The first category reproduces successively in a computer the behavioral operations involved in performing the mission or task in real life. The output of the reproduction represents the prediction of effectiveness. The second category does not reproduce these operations successively but rather assigns a unitary estimate of expected performance to each behavioral unit and then combines these estimates in accordance with a concept of how individual behaviors combine in real life to accomplish the task.

The fact that a model falls into one or the other category has significant data bank implications. Generally (not always) a simulation model requires more detailed data than does a non-simulation model, because it must reproduce the individual behavioral actions of the personnel performing the mission. For example, Wherry's HOS requires the locations of controls and displays and the values these controls and displays may assume, eye movement and arm reach data, etc. Although non-simulation models like the AIR Data Store may require very molecular data, in general simulation models require a somewhat more detailed data level than do non-simulation models.

Consequently it is necessary for the data bank developer to decide whether the technique he will apply the data to will involve simulation or non-simulation. If he can identify the individual technique which will make use of his data, so much the better; but he may be unwilling to do this, because it is always possible that a new technique variation will be developed which he would prefer to use. However, that new technique is likely to be either simulation or non-simulation. (There are other classification dichotemies that can be applied to human reliability models, but the simulation/non-simulation difference is the most fundamental.)

Of course, the data bank developer could say, I will make my bank as detailed as I can; but this is likely to be unprofitable because it leads directly to an attempt to develop a universal data bank.

The choice between simulation and non-simulation techniques as the vehicle for the application of one's data is a matter of relative advantage. In general simulation techniques are more powerful than non-simulation, because

(1) They output functions describing parametric relationships, e.g., the relationship between performer speed and accuracy, thus permitting one to select a particular input value required for a given output; non-simulation techniques provide only single-value outputs.

(2) Since they simulate successive replications of operating performances, simulation techniques provide a history over time that can be examined for diagnostic purposes; non-simulation techniques provide much less diagnostic information.

(3) Moreover, simulation techniques require no combinatorial process as do non-simulation methods, which is an advantage because of the problems involved in assigning conditional probabilities. In consequence, if the data bank developer assumes that his data will be used with a non-simulation model, he must also ask whether the data parameters are such that they can be readily combined.

On the other hand, simulation techniques impose an additional cost in the sense of requiring more detailed information. In consequence, the choice of model should be based on whether the user needs for his purposes anything as powerful as a simulation technique. If, for example, one did not need to determine functional relationships between parametric values, or to diagnose system operations, one might be satisfied with less powerful techniques requiring less detailed information.

It is not that non-simulation techniques make use of different types or sources of data other than operational or laboratory testing or the results of controlled experiments or subjective judgments. It is simply that the data level is less detailed. An illustrative example is the kind of data required by TEPPS as opposed to the kind of data needed by Siegel's digital simulation models.

We have been talking about input data required by the various models. At the gross distinction between simulation/non-simulation, one can speak only of level of detail. However, each model, regardless of the category to which it belongs, has certain peculiar needs for input data. Therefore each of the models to which it is considered that the data bank might be applied should be examined to determine if it has special input data requirements that must be accomodated by the data bank.

A second question the data bank developer should ask is: what is the desired output measure to be provided by the model? Two types of output are commonly provided by models: error probability and time. There is an intimate relationship between the type of input data required of the data bank and the output measures the model provides. If a model outputs error probabilities, then the data input to the model must also be in the form of error probabilities. The same is true of time. In general, time data are easier to secure for the data bank because fewer performance replications are required.

Another interactive factor to be considered is the type of system to which a model will be applied. Some systems are more heavily time-dependent than others and will therefore require primarily time data.

One of the categories used in the review of the preceding techniques was methodological scope, by which was meant the range of behaviors and functions dealt with by the model. The data bank developer must therefore ask: how much will the data bank cover: (1) perceptual, motor, cognitive behaviors; (2) continuous/discrete tasks; (3) maintenance functions as well as operations. All the models reviewed pretend to deal equally well with behaviors and tasks of types (1) and (2), but some of them deal with these more adequately than do others. Since it is easier to collect data on perceptual and motor behaviors involved in discrete tasks, the data bank developer must ask whether the additional effort required to expand his scope is needed in terms of the use to which his data will be put. To gather data on cognitive behaviors and continous tasks may require special controlled and expensive experiments.

Again, the type of system to which the model and data will be applied is important here. If the data bank developer has in mind a class of systems with special characteristics, e.g., involving largely discrete tasks, then he can gear his data bank development efforts to that system-class.

A great deal has been made in this section of the intended use of the model and data bank. This is related to the anticipated uses of the models we have reviewed. These uses can be summarized as follows:

(1) prediction of system effectiveness;
(2) comparison of system configurations;
(3) redesign to correct deficiencies;
(4) manpower selection and training.

Let us leave aside category (4), for which no model reviewed was very useful. Most of the models predict system effectiveness (1) in one form or another. The models reviewed do differ, however, in their capacity to deal with uses (2) and (3). These differences depend on the sensitivity of the model to equipment factors and this sensitivity in turn depends on the extent to which equipment factors are conceptualized as influencing input error probabilities and response times.

If a model is selected which emphasizes equipment details (e.g., HOS), then obviously the data bank developer who will use this model to apply his data must incorporate equipment parameters into his data bank.

The question then becomes, what does the ultimate user wish to do with his model and his data bank? If it is considered important that the data bank be used for comparison of system configurations and for redesign, then the data bank developer should collect data reflecting the impact of relatively molecular equipment characteristics on performance (because it is these which most often differentiate among system configurations).

This involves a cost, however. To collect performance data reflecting equipment parameters it is necessary to set up data collection situations in which the effect of equipment parameters can be controlled. Practically speaking, this requires a laboratory type situation or recourse to the experimental literature.

If one makes the decision to make one's performance data bank sensitive to equipment parameters, it is necessary for the developer to specify what those parameters are, because only in this way can data be correctly categorized. The AIR Data Store provides some examples of potential parameters: number of components of a given type; arrangement of these components of a given type; arrangement of these components; individual component characteristics, such as joystick length; exposure time (for displays), etc. One could of course allow one's data collection situation to determine the equipment parameters (this most often happens in operational testing), but this would imply an assumption that the data collection situation is representative of all the system-types to which one expects to apply the data. For example, if one were to collect data in a Terrier launch situation, the data secured would be most useful in application to Terrier type systems, and of somewhat lesser use to other types of systems.

Finally, the developer should consider what Swain calls "performance shaping factors" (PSF). This relates to the assumptions and parameters inherent in the various models, because these assumptions and parameters deal largely with PSF. The models reviewed do differ in terms of the PSF they include (or for that matter do not include, because a number of models are extremely reserved about these), although a common PSF in many of these models is stress. All of these models would include various types of PSF if data were available on them, e.g., amount of learning, skill level, motivation.

The question the data bank developer must ask himself is: how important is it to include PSF as part of his data bank? Manifestly, the more such PSF data he inputs to a model, the more precise will be the answers provided by the model, for whatever use he wishes to make of the model. Does the eventual user of the model need the increased precision supplied by PSF? Are any of the system-types to which the data

will be applied especially sensitive to PSF? Should a model requiring fewest such PSF be selected in preference to one requiring more of these factors? There is a cost factor here too, because to collect performance data as these relate to PSF may involve a more extensive collection effort and controls than would otherwise be required.

What all of the preceding discussion comes down to is that the data bank developer must relate his development efforts to the anticipated needs and uses of his data. This means consideration of models and possible system development applications. We repeat that the data bank is only a tool to be applied with some model/technique to satisfy some user requirement. One can think of the model/data bank combination as an interesting research problem, but its practical value - which must at one stage or another be addressed - lies in the use that is made of it.

The question arises whether we (we considered as potential data bank developers) know what users need and want from models and data banks. Since these users are various: designers, system developer managers; government planners, etc., their needs may differ. One could of course adopt the point of view that the data bank or model developer already knows the user's needs or will tell (impose upon) the user what he can do with these tools, but the chances are that under these circumstances the user will ignore the tools supplied to him. We have had some unfortunate experiences along this line in the past with other types of human factors techniques provided to system developers.

This does not mean that all data bank development work need be postponed until we determine user requirements for this tool. It does, however, require that the developer attempt to anticipate those uses in very practical terms.

SECTION VI

RECOMMENDATIONS

The recommendations developed on the basis of the model reviews can be described in several ways. First, some of these recommendations are for research, whereas others are for action to be based on research. Secondly, some of these suggestions are for immediate and short term implementation, whereas others can be accomplished only on a continuing longer term basis. Each of the recommendations made will be characterized in this way.

A. Determination of User Parameters

This recommendation is for short term (e.g., 9-12 months) research which should be implemented immediately.

The author makes the following assumption: that any human factors technique, predictive model or methodology (particularly those developed under government contract) should be directed primarily at satisfying a known requirement or need of the system development process. In other words, any government-sponsored research should have as its ultimate aim the development of more effective systems.

We emphasize the phrase "known requirement." Hardly anyone would quarrel with the above premise, but in order to implement that premise it is necessary to determine what that system development requirement is, which is often not too easy. Often there is a substantial difference between the aims, processes and methods of system development as these are conceived of by governmental agencies and the aims, processes and methods actually included in system development at much lower levels. The problem is intensified by the fact that different types of people (for convenience sake call them "users") are involved in system development: governmental planners, SPO directors, contract managers, design engineers, personnel who will operate the developed system, etc. Each of these may have some special answer he wants from the application of a particular model or technique. For example, one of the comments made a number of times by respondents to the evaluation criteria questionnaire was that the evaluation of a particular model might vary, depending on the role a respondent would play in system development.

Actually it is impossible to evaluate the various models meaningfully unless one thinks of them in terms of the particular system development needs for which they are best fitted. If one can believe many model developers, each model will satisfy a wide range of system development needs. Actually, each model has individual features (both positive and negative) which makes that model better or less well adapted to a particular system requirement. For example, TEPPS may be particularly effective

in apportioning a system effectiveness requirement among various tasks; THERP may be particularly effective in resolving problems or questions of effectiveness during development; Siegel's digital simulation model may be most effective in supplying trade-off functions among parameters when the developer wishes to select the most desirable set of values of these parameters. And so it goes for almost all the models reviewed. No model is a "general-purpose" model; and the one a user might wish to select would depend on his particular need. Some methods are more descriptive than others; some are more predictive than others; others differ in terms of number of outputs they provide or the particular stage of system development at which they can be applied; some demand a particular type of data, while others do not, etc.

Because of this aspect it is necessary to look at the various models in terms of a number of system development factors which include:

(1) Stage of system development at which the model can most effectively be applied;

(2) Nature of the system development problem to be resolved. For example, one might not wish to apply an extremely high-powered model when the problem could be more easily and just as effectively solved with a lower-powered model;

(3) Amount of input data required to exercise the model;

(4) Manpower and time cost required to exercise the model.

Other factors which can only be determined by examining the use situation may well influence the choice of a particular technique.

Many of the differences in opinion seen in the proceedings of the first Human Reliability Workshop (Ref. 1) reflected a lack of knowledge on the part of all participants of the relative importance of the system development factors mentioned above. If, for example, a technique is denigrated because it will supply answers only at a less than optimal level of precision, this is because we do not know whether such partial answers will satisfy a particular set of user problems. If the goal of model development is ultimately to provide answers which will in fact solve system development problems, then it is necessary to know what techniques and answers users will accept as capable of solving their problems. It is possible to develop techniques which in theory or even under controlled validation conditions are satisfactory but which, when given to their users, do not supply desired answers.

It appears then that the following questions need to be answered:

(1) Who are the various types of people wh need to make use of a predictive model?

(2) How do the uses that each of these types would make of the models differ? What in fac are these uses? How important is each use? (This report postulated a number of potential uses for the various models, but these were logically rather than empirically based.)

(3) At what stage in system development must users have the information to be provided by the model?

(4) How precise or detailed must the information be (e.g., in terms of four figures, such as .9988, or more gross data)?

(5) In what form should the information be (a probability of task accomplishment, a time estimate, a problem diagnosis, a recommendation for redesign, etc.)?

(6) What type of parameters must the information contain (e.g., equipment details, training requirements, personnel factors) and precisely what should be the details of these parameters (e.g., if equipment recommendations are required, what equipment factors are important to include in the answer supplied)?

It must be emphasized that the answers to the above questions do not yet exist, despite the confidence of some model developers that their models are perfectly adapted to a variety of system development requirements. It is also possible that model developers do not fully know what the needs are to which their models should respond. Upon occasion this author indicated that one or the other model did not handle a particular type of task or requirement. Developers in responding to the preliminary drafts of their model descriptions would reply to the above comment that the model could deal with the situation, <u>if there were a need to do so</u>, thus suggesting that (a) they did not realize all the potential demands that might be levied on their techniques; and (b) there is a need to explore these demands more fully.

If one had answers to the above questions it would be possible to determine what requirements should be levied on models and on data banks.

Such a determination probably would not eliminate any of the more effective models, since it is quite possible that all of these should be made available to system developers, because each attacks a particular set of system development questions in a somewhat different way. It would then be necessary to determine for which set of uses and users and

system development factors a particular method was most applicable. This is the procedure utilized in MIL-HDBK 472 (ref. 2) in which the particular application of the models included is specified.

To buttress the importance of the proposed study the following should be considered. Utility criteria (e.g., ease of use, availability of data, etc.) were generally given a lower weighting by the majority of respondents to the evaluation criteria questionnaire. Most respondents would probably echo one man who replied in his cover letter "All I care about is whether a method works". However, it would be desirable to consider the experience gained with the four maintainability prediction models described in MIL-HDBK 472 (Ref. 2) which presents the four methods and allows the user to select the one he wishes applied. Although Method I is, from the data presented, the best predictor of system down time, it is apparently rarely used because its mathematical complexity defies anyone without an advanced degree in mathematics. Method IV is so general that it cannot practically be applied. The point is that of the four methods presented, only two are really available, because the others contain disabilities which directly pertain to user limitations on the models. Had user requirements and constraints been fully considered when the research for methods I and IV was performed, it is very likely that these methods would not have been developed, or would have been developed in a different way.

It is highly desirable that the operability performance models should avoid the fate of many human factors techniques which, despite much research, money and effort are spurned by the system development personnel these were designed to help (Ref. 3, 4).

It is therefore recommended that a short term study be instituted to determine user needs with regard to the prediction of human performance. The following questions should be attacked by that study:

(1) Who are the various types of people who would make use of these models? How much use would they make of them? To what uses would they apply them? What information would they wish to secure from the models? In how detailed a form? Containing what parameters? At what stage of system development? What limitations would users place on the models (e.g., in terms of manpower needed, type of manpower, etc.)?

(2) Of the various types of models available, which ones appear to be most suitable to answer particular questions? (Since it is obviously not feasible to have all types of users try out each model in turn on their individual problems, a substitute method might be to present descriptions of the model procedures (although not in the depth presented in this report,

of course), together with sample model outputs, and have users indicate for what problems these model-types and their outputs would be most useful.) This could permit government planners to anticipate user reaction to these models when a human reliability prediction requirement was levied on a contract.

The methodology suggested for this investigation is a combination of interview and questionnaire based on the research performed previously by the author and his colleagues (Refs. 3, 4) for the Office of Naval Research and the Air Force and determined empirically to be effective in providing desired answers.

It should be noted that this methodology differs from the typical interview/questionaire procedure in the following aspects:

(1) The methodology requires the development of sample problems based on those ordinarily found in system development, thus lending a high degree of realism and validity to the questions asked.

(2) Respondents are required by the nature of these questions to simulate (with paper and pencil only, of course) the processes they would ordinarily employ to solve these problems.

(3) The questions contain materials actually representative of the problems to be solved. For example, in the studies (refs. 3, 4) cited above, actual system descriptions, functional flow diagrams, etc. were employed. In the study proposed the questions would contain extracts from model procedures, details of system development problems, alternative data formats, etc.

(4) Problems should include typical system development constraints, such as the need to supply information quickly, cost constraints, etc.

Once developed (and obviously considerable care is needed in its development), the test instrument should be presented to representatives of identified classes of potential users, such as SPO representatives, contract managers, design engineers, etc. This will require sampling a number of organizations both in government and industry. The test instrument should be administered in the form of a highly intensive, structured interview, although there may be occasions where this procedure may have to be modified.

In considering the outputs of such a study, it is necessary to consider one of the major outputs of the Navy's Human Reliability program. As indicated by Momyama in reference 1, one of those outputs will be a manual or guide to the application of various human performance predictive

techniques. Among the several resultants of the suggested research would be guidelines to the type of format in which models should be made available to potential users.

So, for example, if it were decided to provide a handbook for operability prediction (similar to MIL-HDBK 472 for maintainability prediction), the information needed by users to make maximum use of these models would be specified.

In addition, the information gained from the study could be used to develop planning guidelines for the entire model development/application and Human Reliability program. It would also indicate the range of system development needs which the Human Reliability program should satisfy. This might be important in convincing personnel who do not realize the importance of that program.

It is not suggested that further research and development of already available models or development of new ones should be postponed while the above information is gathered. The brevity of the study effort will not force a postponement of any planned model or data bank efforts, since it is assumed that these efforts will be of value regardless of what is found as a result of the proposed study. Rather, the study outputs should be directed toward making more effective use of concurrent research.

B. Studies of Subjective Judgment Data

This author concurs with others like Blanchard and Swain that further investigations should be performed to improve the validity and utility of subjective ("expert") judgments. The reason for this recommendation is not that this type of data is to be preferred to any other, but that since so much use is made of it (in the absence of definitive data banks), much more should be done to structure the way in which such data are gathered and used.

This study is a longer-term effort, which would probably extend over 12-24 months.

The study should be directed to answering the following questions: In terms of application to various classes of predictive models

(1) What is the most effective (from the standpoint of validity, reliability, time and cost factors) method (e.g., paired comparison technique) that can be developed to secure subjective judgements?

(2) How effective are subjective judgments in relation to various types of required data (e.g., task times, performance reliabilities)?

(3) How much detail is it reasonable to ask of judgment data and what parameters can be included? For example, how possible is it to include some of Swain's performance shaping factors (Ref. 6)?

(4) What is the highest inter-judge reliability that we can achieve with various methods and how can this be accomplished?

(5) What type of background and how much experience would be required to make meaningful judgments?

Obviously there are a number of conditions that should be included in the study:

(1) alternative methods of securing judgments;

(2) alternative information parameters about which judgments are to be made, such as equipment details, personnel factors, operational conditions, etc.;

(3) relative degrees of personnel background with the type of performance being judged;

(4) relative degrees of detail (e.g., number of information categories) to be required of the judgments.

It is essential, we think, that in addition to inter-judge reliability some measure of the validity of the judgments made should be determined. There are, of course, various ways of establishing that validity, but hopefully the proposed investigation would involve some sort of comparison between the judgments made and actual task performance. This might require selecting a particular Navy system as a reference vehicle about which judgments were to be made.

C. Development of Design-Oriented Models and/or Data Banks

We have noted in previous sections of this report that many of the models reviewed supply comparatively little information about the performance correlates of design factors or supply little design-specific guidance to engineers. Although such models may indicate where a problem exists, they provide little information concerning the precise source of the problem. We consider that it is not enough to be able to say that, for example, an operator is overloaded at point X in the mission

without being able to specify the nature of the (equipment) design features producing that overloading (if equipment design features are in fact the cause). Nor do many of the models provide any initial design guidance in the development of the system.

In saying this it must be recognized that many of the presently available models (as their developers pointed out in personal communication) were not developed for the above purposes. For this reason the government might consider sponsoring the development of models specifically directed at providing this design guidance.

There are two ways in which such design guidance might be provided:

(1) Development of a model on the order of method III in MIL-HDBK 472, which predicts maintenance downtime as a function of three checklists dealing with design, support and personal features of the system;

(2) Development of a data bank based on the experimental human performance literature specifically directed at describing the performance correlates of design features; this could either be an independent data bank or incorporated as part of the NELC data bank which is in process of development.

Both these efforts would involve considerable time (e.g., 24 months or more).

The development of a model which would predict performance reliability and task time as a function of checklist judgments would require

(1) the isolation of the design features which appear to influence operator performance;

(2) the development of appropriate checklists;

(3) the application of these checklists to a variety of systems in the operational environment;

(4) the gathering of human performance reliability data on those systems; and

(5) finally the performance of multiple regression analyses which would lead to a predictive equation. Nor should we ignore the necessary validation of the technique.

The advantage of such a model would be that it would permit design engineers to predict very early in system development the human performance

to be anticipated from that design. Then, if that reliability appeared inadequate, it could be traced back to the individual design factor (as described in the checklist) responsible for the inadequacy. At the present time, although human engineering judgments are made during development, they are not tied to performance predictions, nor can performance predictions (from presently available techniques) be easily linked to human engineering aspects of design.

Somewhat less time would be required for the development of a design-specific data bank. It is apparent from the investigations reported in reference 5 that a considerable amount of data exists in the literature that could be applied to this effort. It would be course be necessary to determine in advance the specific equipment parameters on the basis of which the data bank would be organized, but it is assumed that study A (user parameters) would provide this information. Reference 5 also suggested a number of alternative data bank formats, the adequacy of which could also be investigated in the proposed study of user parameters. The relevant discussion from reference 5 is appended at the conclusion of this section.

D. Experimental Validation of Presently Available Models

A much longer range recommendation and one which requires more sustained action of the part of governmental sponsoring bodies is to emphasize the validation of the most promising of the models available. It was pointed out previously that a number of the models reviewed had never been validated and that even for those models which had received some attention in this regard, the amount of validation data seemed insufficient. It is suggested, therefore, that those agencies responsible for the development of models place additional emphasis on their validation, particularly by comparing model predictions with actual task performance.

Even when a model has been formally validated, it often has not been formally applied, by which is meant that no study has been made of the success with which it has been applied to various system development problems. (Incidentally, this might be one additional aspect to be examined in the proposed study of user parameters.) As was shown by Method I of MIL-HDBK 472, a model may be effective in prediction, but unsuitable (for one reason or another) in actual application. Investigations of model application could lead to improvements in models that have predictive power but are difficult to apply.

The investigations suggested are of a continuing nature and preferably conducted by the governmental sponsoring agency itself. Hence a single

specific study is not recommended. It is of course possible to make certain suggestions as to how such studies should be conducted. In addition to supporting formal validation studies of models under development, the sponsoring agency should periodically survey those users who have employed a particular model to determine the degree of satisfaction with the model, the particular uses for which the model was found most satisfactory, any user difficulties encountered, etc. This type of information should be fed back to the individual model developers and to workers in the field generally.

E. Other Studies

So many variables affect the predictive efficiency of man-machine models that is difficult and perhaps unjustified to pick out certain ones to be emphasized in experimental studies. It should be noted, however, that special difficulties were experienced by a number of non-simulation models with regard to conditional probabilities and the effect of feedback loops. Any reasonable method of determining conditional probabilities would be highly desirable; the paper by Williams (Ref. 7) illustrates how difficult the problem is. Similarly, it is possible that a major reason for overly pessimistic estimates of performance is that it is difficult to account for feedback factors in behavioral models. Ways of including these factors more specifically in models would be quite beneficial.

Summary

The preceding list of research/action recommendations does not by any means include all possible research suggestions. The further development of any presently available (or for that matter, any future) models will certainly suggest further studies to be performed that are peculiar to that model. In the preceding section we have concentrated on suggestions that would have broad range applicability.

In this list of suggestions we have emphasized that research which in our opinion would have the greatest impact on the overall effort to develop useful models for predictive purposes. Our point of view is that if models are to be more than research devices, they must be responsive to system development requirements and that, therefore, lacking detailed knowledge of those requirements, we must determine those requirements and allow them to guide further model development. It is for that reason that study A has been presented in somewhat more detail than the other recommended research.

REFERENCES

1. Jenkins, J. P. (Ed.), Proceedings of the U. S. Navy Human Reliability Workshop, NAVSHIPS 0967-412-4010, 22-23 July 1970, published February 1971.

2. Maintainability Prediction, MIL-HDBK 472, Dept. of Defense, Washington, D.C., 24 May 1966.

3. Meister, D. and Sullivan, D. J. A Further Study of the Use of Human Factors Information by Designers. Bunker Ramo Corporation, Canoga Park, Cal., March 16, 1967.

4. Meister, D. Human Factors: Theory and Practice. New York: Wiley, 1971.

5. Meister, D. and Mills, R. G. Development of a Human Performance Reliability Data System, Phase I. Report AMRL-TR-71-87, Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, July 1971.

6. Swain, A. D. Development of a Human Error Rate Data Bank. In, Proceedings of the U. S. Navy Human Reliability Workshop, NAVSHIPS 0967-412-4010, February 1971.

7. Williams, H. L. Dependent Models for Estimating Human Performance Reliability. In, Symposium on Man-Machine Effectiveness Analysis, Human Factors Society, Los Angeles, Cal., 15 June 1967.

ALTERNATIVE HPR DATA BANK FORMATS

The question we address in this section is what the HPR data bank format should consist of.

A data bank is not simply a data bank, although that impression has developed over the years. It is possible to distinguish five types of banks and two types of users of these banks.

The first type of data bank is what can be termed a probability statement of task performance. A sample item might be: the probability of throwing a double-pole, double-throw switch correctly is 0.9968. Note that this statement says nothing about the characteristics of that switch (other than its designation) and does not apply a probability statement to those characteristics. Table 15 (taken from Blanchard et al.) reflects such a data bank.

A second type of data bank would consist of probability statements associated with specific equipment characteristics. For example, the probability of correctly operating a joystick control of stick length 6" - 9" is 0.9963; the probability of correctly operating a joystick with 5- to 10-pound control resistance is 0.9999, etc. Note that there is no single performance probability associated with the task, only with equipment characteristics, although there is no reason why the two could not be combined. The classic example of such a data bank is the American Institute for Research Data Store (Munger et al.), a page from which is shown in Table 16.

A third type of data bank item could consist of the raw performance data values associated with particular parameters. Table 17 presents an illustration of such a data bank format. Note that the data shown in the table are not presented in a probabilistic fashion, although the error data could presumably be transformed into probabilities. Presumably data would be selected to illustrate the desirability of selecting one or the other design characteristic. For example, in the item dealing with TV resolution, it would seem reasonable to the designer that if one wished near perfect observer response, between 7.8 and 11.5 scan lines would be required, with symbols 10.2 minutes in size.

A fourth type of data bank item could consist of quantitative, nonprobabilistic statements related to specific equipment characteristics. For example, a sample item might be: display format X will produce 1.658 times more effective performance than display format Y (X and Y differing in specified ways). The statement can be quantitative or qualitative; one could use an arbitrary set of scale values to represent relative performance, or rating scale, or a ranking. The statement could be relativistic, e.g., ranking various attributes or it could be absolute, e.g., format X is good as indicated by a rating of 6.8 on a scale of 9. Table 18 presents a sample set of data bank items of this type.

TABLE 15

Data Bank Format I

Stimulus Activity	Probability of Correct Performance
1. Turn rotary selector switch and observe CRT signal quality.	0.9972
2. Observe several dials qualitatively for correct readout.	0.9973
3. Observe radar scope and mark target position with grease pencil.	0.9989
4. Turn rotary selector switch to specific position.	0.9996
5. Observe several quantitative and qualitative indicators and determine if equipment is operating correctly.	0.9639
6. Track rapidly moving radar target with 2 unidimensional controls.	0.9709
7. Find scheduled maintenance procedures in maintenance manual.	0.9968

TABLE 16

Data Bank Format II

JOYSTICK

(May move in many planes.)

BASE TIME = 1.93

Time added	Reliability	
		1. Stick length
1.50	0.9963	a. 6-9"
0	0.9967	b. 12-18"
1.50	0.9963	c. 21-27"
		2. Extent of stick movement (extent of movement from one extreme to the other in a single plane).
0	0.9981	a. 5-20°
.20	0.9975	b. 30-40°
.50	0.9960	c. 40-60°
		3. Control resistance
0	0.9999	a. 5-10 lbs.
.50	0.9992	b. 10-30 lbs.
		4. Support of operating member
0	0.9990	a. Present
1.00	0.9950	b. Absent
		5. Time delay (time lag between movement of control and movement of display).
0	0.9967	a. .3 sec.
.50	0.9963	b. .6-1.5 sec.
3.00	0.9957	c. 3.0 sec.

TABLE 17

Data Bank Format III

Equipment Type: Digital Switches

Characteristics	Mean Positioning Time (seconds)	% Reading Errors
1. Standard rotary selector switch	3.66	48
2. Digital pushbutton switch	6.44	0
3. Thumbwheel	5.88	0

Equipment Type: TV

Characteristics	Subtense Symbol Angle (minutes)	Percent Correct Response: No. scan lines per symbol height 4.6	6.3	7.8	13.5
1. No. of raster scan lines per symbol height.	4.4	66	76	70	80
	6.0	73	91	91	95
2. Symbol subtense angles.	10.2	66	87	97	99

TABLE 18

Data Bank Format IV

Equipment Type: Large Screen Display

Characteristics	Performance Relationships
1. Vertical versus horizontal format.	66% more time is spent scanning vertical format than is spent on horizontal array.
2. Effect of coding display.	Mean time to locate coded update information is approximately 65% less than for uncoded updates. As number of stimulus elements presented increases from 36 to 90, time to locate coded updates increases 100%, time to locate uncoded updates increases 150%.
3. Effect of number of stimuli.	Response time increases linearly with number of stimulus elements presented.
4. Effectiveness of large screen display over small screen display.	Performance with large screen display is 15% faster than with small screen; no difference in error found.

A fifth type of data bank format and one which is personally most appealing on purely heuristic grounds would combine all the characteristics of the preceding formats. Such a format would provide to the user all the data available in whatever form it could be provided, whether or not the data could be formulated probabilistically. Thus, probabilistic values would be associated with certain tasks and task characteristics, where such values were available; raw performance data for other task parameters would be supplied when probabilistic information could not be supplied, etc. Table 19 presents an illustration of such a combined data item.

The two types of customers who might make use of the various data banks are, beside the human factors specialist, reliability engineers and design engineers. Historically the concept of the human performance reliability data bank was developed out of the reliability engineering tradition which has emphasized prediction — hence, the need for probabilistic statements. The design engineer, however, is not so much concerned with prediction as with the selection of one design concept or characteristic rather than another. What this means is that he considers a number of alternative design characteristics and decides that one of these will give him more effective performance. From that standpoint, therefore, he does not require probabilistic statements because his choices are all relative.

The first type of data bank item shown in Table 15 is not likely to be much use to a design engineer because it does not specify equipment characteristics, which is what he is interested in. (However, we suggested previously that predictive data might be useful to the system developer in comparing alternative system configurations. The first type of data bank could be used for this purpose). Moreover, the data presented does not imply relationships among equipments.

Data bank formats I and II differ, moreover, in the ease with which they can be secured. The first type of data can be secured from almost any kind of testing in a nonlaboratory environment and requires no special control situation. The second type of data bank must be derived from a highly controlled test situation in which the operator's performance can be partialed out to reflect the individual equipment characteristics he is responding to. This can be done only in a laboratory situation where the experimenter has specifically set up controlled situations that contrast two or more different equipment characteristics. The experimental situation is therefore directed at the individual characteristics being compared rather than at the equipment as a whole. This is precisely what a laboratory situation is equipped to do.

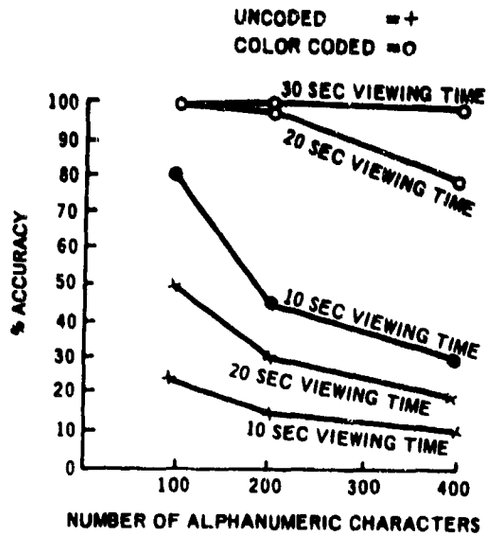
However, this is very difficult to do in the nonlaboratory situation, since in this situation the operator responds to the entire equipment, rather than to an isolated equipment characteristics.

TABLE 19

Data Bank Format V

Equipment Type: CRT Displays

Characteristics	Performance Relationships
1. Probability of correctly performing reading and updating functions.	0.9743*
2. With alphanumeric symbols.	0.9889*
3. With geometric symbols.	0.9654*
4. As a function of resolution:	
6 scan lines	0.7543*
8 scan lines	0.8644*
10 scan lines	0.9044*
12 scan lines	0.9756*
5. The effect of density and display exposure time on accuracy.	



*Note: These probabilistic values are purely hypothetical.

TABLE 19

Data Bank Format V (continued)

6. Improvement in observer performance when displays are coded.

Original Displays	Code Type	Observer Function	% Accuracy Improvement	% Response Time Improvement
Alphanumerics	Color	Locating	44	--
Alphanumerics	Color	Counting	86	72
Alphanumerics	Size	Update	50	65
Map	Conspicuity (border)	Information Assimilation and extraction	97 and 57	--
Alphanumeric and Shape	Color	Search and Count	15-53	5-25
Alphanumeric	Size	Update	49	

Although one can use the second type of data bank for design decisions, the engineer does not make use of the absolute value of the probability statement for this purpose. If one joystick length gives 0.9968 performance, whereas a second length gives 0.8968 performance, then the designer implicitly or explicitly ranks the two characteristics (lengths) and selects the one with the higher probability. It would make no difference to the designer if the absolute performance probabilities were different as long as the two lengths retained their relative performance standing, nor would it make any difference to the designer if the two characteristics were simply ranked 1-2, although he would probably want supporting data to back up the ranking.

Data for the third and fourth types of data banks (see Tables 17 and 18) can also be secured from laboratory studies. However, they are easier to develop than either the two previous banks because they do not require that their data be transformed into probabilistic values. An examination of the masses of experimental data to be found in the general behavioral literature reveals that much of it can be used for data bank formats III and IV, but the same thing cannot be said about data bank formats I and II. Much experimental data cannot be adapted to probabilistic statements because they deal with such measures as reaction time, response duration. trials to learn. etc.

All things considered, the third and fourth types of data banks are easier for the design engineer to use, they will provide more back-up information (since they are not confined to probabilistic statements) and one can build up a larger data store, because data from the literature that could not be used for the other more restrictive types of data banks could be used for this one.

In the present state of data availability and considering that one of the primary users of the HPR data system should be the design engineer, it seems only reasonable that all the data bank items should be included in the HPR data system. This is the format shown in Table 19. This format takes maximum advantage of all possible data sources by including probabilistic statements, where these can be made, and nonprobabilistic performance data where probabilities are not available. Where possible the data should be transformed into probability statements so that they can be used for predictive purposes. However, all other relevant data relating human performance to equipment/task parameters could be supplied in a format which is most suitable for design purposes.

APPENDIX A

LIST OF RESPONDENTS TO THE MAN-MACHINE MODEL EVALUATION CRITERIA QUESTIONNAIRE

1. Altman, J. W., Syneetics Corp., Allison Park, Pa.
2. Arima, J. K. U. S. Naval Postgraduate School, Monterey, Calif.
3. Askren, W. B. Human Resources Lab., Wright-Patterson AFB, Ohio.
4. Bennett, C. State University, Manhatten, Kansas.
5. Carr, R. M. Raytheon Co., Portsmouth, R. I.
6. Chapanis, A. John Hopkins University, Baltimore, Md.
7. Christensen, J. M. Human Engineering Division, AMRL, Wright-Patterson AFB, Ohio.
8. Coburn, R. Navy Electronics Laboratory Center, San Diego, Calif.
9. Collins J. Office of the Chief of Naval Operations, Washington, D. C.
10. Erickson, R. A. Naval Weapons Center, China Lake, Calif.
11. Folley, J. D. Applied Science Associates, Valencia, Pa.
12. French, R. S. Naval Undersea R & D Center, San Diego, California.
13. Graine, G. N. Naval Ship Systems Command, Washington, D. C.
14. Hagen, W. Human Resources Lab., Williams AFB, Arizona.
15. Hanifan, D. T., Harbinger Corp., Santa Monica, Calif.
16. Hampton, D. Raytheon Co., Portsmouth, R. I.
17. Harris, D. Anacapa Sciences, Santa Barbara, Calif.
18. Hiss, D. Raytheon Co., Portsmouth, R. I.
19. Howard, R. Naval Ship Systems Command, Washington, D. C.
20. Jenkins, J. P. Naval Ship Systems Command, Washington, D. C.
21. Jones, D. Martin-Marietta Co., Orlando, Fla.
22. Knowles, W. San Fernando Valley State College, Northridge, Calif.
23. Krendel, E. S. Univ. of Pennsylvania, Philadelphia, Pa.
24. Lamb, G. Naval Underwater Sound Center, New London, Conn.
25. LaSala, K. Naval Ship Systems Command, Washington, D. C.
26. Mackie, R. R. Human Factors Research, Goleta, Calif.
27. Miller, J. W. Natl. Oceanic & Atmospheric Admin., Rockville, Md.
28. Mills. R. G. Human Engineering Division, AMRL, Wright-Patterson AFB, Ohio.
29. Moore, H. G. Naval Materiel Command, Washington, D. C.
30. Ozkaptan, H. Navy Personnel R & D Lab., Washington, D. C.
31. Parsons, H. M. New York City.
32. Pew, R. W., University of Michigan, Ann Arbor, Michigan.
33. Regan, J. J. Naval Training Device Center, Orlando, Florida.
34. Rigney, J. University Southern California, Los Angeles, Calif.
35. Rizy, E. Raytheon Co., Portsmouth, R. I.
36. Siegel, A. I. Applied Psychological Services, Wayne, Pa.
37. Sjoholm, A. A. Bureau of Naval Personnel, Washington, D. C.
38. Swain, A. D. Sandia Laboratories, Albuquerque, N. M.
39. Topmiller, D. A. Human Engineering Division, AMRL, Wright-Patterson AFB, Ohio.

40. Townsend, J. C. Catholic University, Washington, D. C.
41. Webster, R. G. Sandia Laboratories, Albuquerque, N. M.
42. Woodson, W. E. Man-Factors, San Diego, Calif.

A number of responses from others were received which could not be used because the author's instructions were not followed. He extends his appreciation to these respondents also.

APPENDIX D

REFLECTIONS ON THE STATE OF THE ART OF HUMAN RELIABILITY MODELING

INTRODUCTION

In this Appendix the author will endeavor to summarize his impressions of our status with regard to the development and use of methods for quantitatively assessing and predicting human performance effectiveness. In this he follows previous summaries written by Meister (1964, Ref. 6), Freitag (1966, Ref. 3), Altman (1968, Ref. 1) and Swain (1969, Ref. 12). Of these, the most comprehensive and significant is that of Altman; consequently, we will update Altman's paper by quoting from it those sections that are most appropriate and adding our own comments based on conclusions reached in reviewing models in this report. Certain topics have also been added to this discussion which were not covered in Altman's summary.

The reader may ask whether this discussion does not merely repeat the Conclusions section of this report. However, the Conclusions section dealt solely with the models reviewed; this section relates to the entire field of human performance quantification (not that the author pretends to be omniscient) and includes certain topics not explored in full in previous sections.

Like Altman "I have chosen to use idealized characteristics of quantitative methods as criteria against which to evaluate progress" (p. 1).

MODEL VALIDITY

Altman found "little in the way of definitive study of the relationship between quantitative performance estimates and more immediate measures of job performance. Virtually every study is, or should be, couched by the authors in the most guarded terms" (p. 12). This statement remains as true today as it was in 1968.

We have found that for almost all of the models/techniques reviewed there is simply insufficient evidence with which to assess their validity. For some models, e.g., AIR Data Store, Siegel's digital simulation model, formal tests of predictive efficiency have been made, whereas for others (e.g., THERP, ORACLE) predictive efficiency has been claimed in

applying models to system development but on this point data are not presented. Even when a formal test of validity has been performed, and indicates promise for the method, the validity test was not performed in the context of system development, so that questions about validity in use still remain to be answered.

In part this unsatisfactory situation exists because there appears to be some objection on the part of model developers to the notion that the most effective test of validity is comparison of a prediction with operational performance. Most frequently validity comparisons are based on such methods as validation by comparison with expert judgments, with other test measures, etc.

Although comparisons of predictions with operational performance are sometimes difficult to perform, the author feels that only such a validity test will completely satisfy the validity criterion. It should be the responsibility of the agency sponsoring the model or method to insist on such a test; otherwise the agency cannot determine the value of its research efforts.

RELIABILITY

Altman indicates (p. 11) that "No complete replications study... has come to my attention... The only time quantitative estimates of reliability have been reported seems to be when a fallback to ratings or judgments has been involved" (e.g., Irwin, Levitz and Freed, Ref. 4). With regard to the models/techniques reviewed in this report, there is no formal evidence concerning reliability, except for the semi-abortive user test of TEPPS (Smith et al., Ref. 11). Either the developer alone has employed his method (which is the usual case) or, if the method has been used by others, data on the adequacy of the replication-use is unavailable. Second only to the requirement to demonstrate validity, the sponsoring agency should give high priority to formal efforts to demonstrate reliability. A method cannot be considered useful until it has been established that others besides its developers can perform model operations.

OBJECTIVITY

We echo Altman when he says that model developers are trying "to free their administration from obvious bias on the part of the administrator"

(p. 11). However, there are still subjective elements in most methods, which implies nothing more than that these methods are still incompletely developed. Among these elements to which subjective judgments are applied are (1) the determination of the behavioral unit (e.g., subtask, task) to which the predictive data will be applied; (2) the determination of the parameters for which relevant data are to be selected; (3) subjective estimates of performance. Subjectivity makes it difficult to secure reliability; and also makes the method more difficult to use, because the user may not be quite clear about the procedural steps he should employ. What is disturbing about subjectivity is not that it exists (because one would expect it to exist in prototype methods), but that the subjective elements are not identified in the model descriptions, so that one can attempt to reduce their subjectivity as much as possible. Model developers are not as systematic in their descriptions of their methods as one would wish them to be.

EASE OF USE

Altman indicates "It ought to be obvious how easy available quantification techniques are to use, but this is one of the least reported aspects" (p. 9). The author can only concur with this statement. From the results of the criteria evaluation questionnaire (see Section II) and from comments made at the 1970 Human Reliability Workshop meeting (Ref. 5), the attitude of some model developers seems to be that this aspect is of miniscule importance. Compared to validity and reliability, they are quite correct; we have pointed out elsewhere, however, that if a method is difficult to use or involves an excessive "cost" in terms of time and effort, it is unlikely to be employed, regardless of its predictive efficiency.

One of the requirements that should be levied on the model developer is quantitative information on the time, effort and resources involved in applying his model, even if we recognize that these will be greater during model development than when the model is used operationally. Very few of the model descriptions contain such data. Since ease of use information can be gathered at the same time as are reliability data, reliability testing should include ease of use aspects.

STAGE OF SYSTEM DEVELOPMENT

"Theoretically at least, a quantitative performance technique should have utility throughout a good share of the developmental sequence. There is virtually no organized information on (this)" (Altman, p. 8).

Most models claim they are applicable throughout the various stages of system development, but the author doubts this claim, simply because they have never been applied at these stages and information on this point is lacking. As has been pointed out elsewhere, we do not know what the demands of the various developmental stages are because we have never inquired formally into the problem. Moreover, one cannot speak of applicability as if it were a binary function; even though all methods may be applicable at a given developmental stage, they may not be equally applicable to that stage, if only because model requirements for data vary substantially and data availability varies with development stages.

STAGE OF MODEL DEVELOPMENT

It is obvious that no method can be considered fully developed ("completed", so to speak), nor would any model developer make this claim about his creation. Areas of ambiguity and unresolved questions are to be found in every method (as may be seen by referring to the model descriptions themselves).

DESIGN SENSITIVITY

Altman indicates (p. 7) some doubt as to the sensitivity of models to design-performance relationships, and we must echo that doubt. Most models reviewed in this report are apparently capable of pointing out where a system deficiency exists, but not of relating that deficiency to the responsible equipment design characteristic. The determination of equipment design features to be modified seems to be more a function of human engineering analysis following upon application of the model than of model outputs themselves.

This situation may in part result from greater emphasis on the quantitative predictive aspects of modeling than on the implications of the achieved predictions. It is interesting to note that some of the models that do seem to output equipment design relationshis (e. g., HOS, ORACLE)

do not output a quantitative assessment of performance capability, but are more specialized for qualitative analysis of system events.

METHODOLOGICAL SCOPE

It has been observed that most model developers seem to claim more for their methods than the methods, upon close examination, seem to be able to handle. In general, most models seem to have much less difficulty predicting for discrete tasks than for continuous and complex cognitive ones. In the case of continuous tasks, this deficiency reflects our difficulty in modeling feedback effects; in the case of cognitive tasks behavioral science as a whole lacks knowledge about cognition.

DATA STORES

With the exception of the AIR Data Store and TEPPS, most data sources employed by modelers are informal and must be gathered specifically for the particular system problem to which the model is applied. All data sources are examined to select most relevant data. This presents special problems for the model user. It is our feeling that the average engineer wishing to make use of a model will be unlikely to engage in strenuous efforts to find relevant data in the literature or through experimentation; if he uses a model he wishes to have the appropriate data store given to him with the model. If this is not available, he will, more likely than not, reject the model. Even if he does not reject the model, it is unlikely that his data gathering efforts will be adequate to the requirements of the model. This in fact is the main reason for the push behind the effort to develop a formal human performance data bank.

We have already pointed out the requirements for development of an appropriate data bank: specification of the parameters included in the bank; the model(s) to which the data can most validly be applied; and detailed procedures for using the data bank. In this connection it is difficult to determine from published reports of the models which parameters are relevant to the data required by the model.

It might also be considered that although most workers in the field think of the data bank, as if there were to be only one, it is at least conceivable that several data banks, including different parameters, may be needed to satisfy different model requirements.

TASK ANALYSIS METHODOLOGY

Almost all of the models require some form of task analysis to determine the behavioral unit to which its predictive data are to be applied. Equally, those models that make use of task analysis assume that the user is familiar with the method. In consequence, the analytic operations are usually unspecified, and the level of detail of the behavioral unit, which is critical for model application, is rather vague, in some models involving several levels ranging from the subtask to the detailed function.

Some modelers feel that this behavioral unit level is unimportant, claiming that the model will fit any unit desired. Regardless of the accuracy of this claim, failure to define the behavioral unit in specifics makes it difficult to determine what requirements the model levies on its data bank.

This difficult may reflect merely a lack of specificity in the descriptions that modelers provide; if so, this could be easily remedied. On the other hand, it may reflect the absence of a recognized task taxonomy, which Altman (p. 8) also noted. In this connection, mention should be made of continuing efforts by Fleishman and his co-workers (Ref. 2) to develop a more efficient human performance taxonomy. As of this date a taxonomy accepted by most workers in the field does not exist.

SIGNIFICANT METHODOLOGICAL PROBLEMS

A number of significant methodological problems noted by Altman continue to hamper further development of an effective human performance predictive model. These include:

(1) The problem of subjective estimates and data banks in general; this has already been treated in some detail.

(2) The problem of independence vs. interdependence of behaviors. Manifestly most behaviors are interdependent but it would be much easier to deal with them (i. e., to combine the predictive probabilities of each task) as if they were independent. As an example of this problem, let us consider task X which is composed of subtasks x_a, x_b, x_c.... x_n. If one were to measure x_a, x_b, x_c.... x_n individually, and then combine the results of their performances, would the resultant performance value differ very much from a measurement of X as a whole? This is the

problem to which Mills (Ref. 10) has addressed himself; the results of his studies will be discussed later.

A similar problem is that of conditional dependency. An example of this problem is the effect of task X, performed concurrently with or at an earlier point in the mission, on task Y. The author began this project by thinking of this as a modeling problem, i. e. , that the nature of conditional dependencies would pose more difficulty for some models than for others. However, he has since changed his mind, because it appears that the probability theory is available which can handle the quantitative treatment of this dependency, provided the data are available.

Thus, the binary effect of X as occurring or not occurring upon the occurrence or non-occurrence of Y can be handled readily, as Williams (Ref. 13) has shown. However, the problem one has to face is that different probabilities of accomplishment of X will result in different probabilities of accomplishment of Y. This is a problem of data availability, rather than of modeling per se. When, therefore, certain models assert that they can deal with conditional dependencies, they are probably correct, provided that the requisite data are provided to them.

The fact of the matter, however, is that we do not have the requisite data. Consequently it cannot be asserted that any model can effectively deal with the problem, except in abstract terms. However, this is not really the modeler's responsibility. As Altman, p. 11, points out "Two factors need to be kept in mind about dependent probabilities vis-a-vis future progress in quantifying human performance evaluations. First, we don't know how much dependency characterizes different circumstances or behaviors. This means that we currently do not know where to concentrate our efforts or even how important it is to concern ourselves with the whole issue. Second, there is, to my knowledge, no existing body of data which can be drawn upon currently to support even a limited dependent probability model of performance."

(3) Another set of problems revolves around the relative importance of time and error as dependent performance variables. Certain models ignore error almost completely, or treat it as a factor which merely tends to delay the completion of the mission. Such models assume that, given that the operator is reasonably well trained and has sufficient time, he will inevitably succeed. Presumably he will recognize when he has made an error and will be able to correct it, thus merely lengthening

mission completion time. Other models utilize error probability as a performance metric, which assumes that there is some finite likelihood that the task and mission will not be completed successfully. Selection of one type of model or the other will provide significantly different outputs, although one could combine the two measures in terms of a probability of completing the mission successfully within a specified (required) time. The relative roles of time and error therefore need further exploration.

Even if one restricted oneself to error alone, the methodological problems one faces do not disappear. It is apparent that one can categorize errors in various ways depending on their consequences. There are errors that have an effect on equipment performance, which we term human-initiated failures. Then there are errors that have no equipment consequences. There are errors that are correctable and those that are not. There are errors that have an effect on mission performance and those that do not. As Altman, p. 10, points out, "It is quite characteristic of human performance that errors are detected and corrected with varying probability, time between commission and correction, and consequences. Most techniques either ignore this important issue or are relatively ineffectual at coping with it realistically." Further research efforts along this line would be most helpful.

SATISFACTION OF USER NEEDS

It is probably characteristic of most specialists in a particular subject matter that they feel they know what the user-audience to which their techniques are directed need and want. In the case of Human Factors, methods are often developed for use by engineers, or for application in their designs, which have never been checked against the engineers' need, willingness or ability to use these methods. As a result, many such methods are more honored in the breach than in the observance (Meister Ref. 8).

In this respect, human performance predictive techniques do not vary significantly. Certainly no model reviewed in this report has been developed after an investigation of user needs which specified the general requirements of that model. The fact is that we do not know what the user (e.g., the system developer, the governmental planner, the design engineer, the reliability and human engineer) needs because he has never been asked.

We suspect that most models differ primarily in terms of the particular system development needs (e.g., data precision, stage of system development applicability, scope of behaviors/tasks to be predicted, level of detail) to which they are sensitive. Even if these models were all valid (as determined by their ability to predict operational performance), they would still probably differ in terms of their efficiency, defined as response to or satisfaction of user needs.

ON-GOING RESEARCH

The reader should also be aware of the following relevant research which has either just been completed or is still on-going. Again, the list may not be complete.

(1) Research on prediction of the frequency of human-initiated failures.

The author and his colleagues have for some years been conducting research to develop a method for predicting the incidence of human-initiated failures (HIF). These are equipment failures resulting from an operator or maintenance technician error. This research which has been sponsored by the Rome Air Development Center, has its origin in the fact that predictions of equipment reliability made during system development are often quite erroneous. It is hypothesized that the factor responsible for such predictive inaccuracies is the incidence of equipment failure in which there is some human involvement.

In the first study of the series (Meister et al., Ref. 7), it was found that 13% of all equipment failures analyzed over a 5 months period resulted

from operator or maintenance induced error, 5% of maintenance actions were taken because of false reports of equipment malfunctions, and 18% of all remaining equipment failures had some human involvement. The frequency of these failures appeared to vary as a function of the type of component involved. The effect of HIF was to cause a significant reduction in mean-time-between-failures.

As a result of the first study it was hypothesized that HIF frequency was determined primarily by the frequency and duration of personnel interaction with equipment possessing certain design characteristics. A study is presently being performed to verify this hypothesis and to develop a usable predictive technique, based on multiple regression analysis.

(2) Development of a Human Performance Reliability Data System.

The author and his colleagues have also been involved in an effort (Meister and Mills, Ref. 9) to establish the parameters of a Human Performance Reliability (HPR) data system based on the available experimental literature. The heart of the system is a taxonomic structure for classifying both general behavioral and man-machine specific studies. Studies are classified in terms of the behavioral function performed, the stimuli presented and the equipment used to respond, environmental, subject and task characteristics. The end product of the classification is a descriptor, e.g., A2/C3/F14/H1/J2/K1 M11/P7, which is assigned to the data and which is used as the means of retrieving those data.

Data are retrieved by first encoding a question asked of the HPR system. This is done by translating the question into descriptor categories. The HPR system then operates on the basis of "and/or" logic to sort progressively through the various categories to achieve the closest possible match with the entry descriptor. By this matching process, the precise answer to the question asked can be retrieved, always assuming that the data bank contains appropriate data.

To test this concept a preliminary HPR data bank of 140 studies was developed. Actual system development projects were used to construct sample questions that were asked of the HPR system. The data retrieved were examined in terms of their relevance to the questions asked. An average of 74% relevant recall was achieved.

A FINAL NOTE

The author would like to conclude this review of human reliability models by emphasizing how important these models are not only to the solution of the practical problems of man-machine system development but also to the discipline of Human Factors as a whole. The distinctive character of Human Factors (as differentiated from experimental psychology) is that it deals with systems (in system terms) and with operational tasks which psychological studies (and psychological theories) do not. The models reviewed (at least the operability models) represent attempts to organize man-machine behaviors in a meaningful, predictive manner. Nothing like them are to be found among the many psychological models that have been developed, most of which deal with molecular stimulus-response elements. It appears to the author, therefore, that further progress in Human Factors theory and application must depend upon the construction and testing of models such as these. Hence the importance of these models which transcends their stated purposes.

REFERENCES

1. Altman, J. W. Progress in Quantifying Human Performance, paper presented at the Electronics Industries Association Systems Effectiveness Workshop, Chicago, Ill., 18-20 September 1968.

2. Fleishman, E. A. and Stephenson, R. W. Development of a Taxonomy of Human Performance: A Review of the Third Year's Progress. Report AIR-726-9/70-TPR3, American Institutes for Research, Washington, D. C., September 1970.

3. Freitag, M. Quantification of Equipment Operability: I. Review of the Recent Literature and Recommendations for Future Work. Technical Report 940, Navy Electronics Laboratory, San Diego, California, 2 June 1966.

4. Irwin, I. A., Levitz, J. J. and Freed, A. M. Human Reliability in the Performance of Maintenance. Aerojet-General Corporation, Sacramento, California, May 1964.

5. Jenkins, J. P. (ed.) Proceedings of U. S. Navy Human Reliability Workshop, 22-23 July 1970, Report NAVSHIPS 0967-412-4010, Dept. of the Navy, Washington, D. C., February 1971.

6. Meister, D. Methods of Predicting Human Reliability in Man-Machine Systems. Human Factors, 1964, 6(6), 621-646.

7. Meister, D. et al. The Effect of Operator Performance Variables on Airborne Electronic Equipment Reliability. RADC-TR-70-140, Rome Air Development Center, Griffiss AFB, Ohio, July 1970.

8. Meister, D. Human Factors: Theory and Practice. New York: Wiley, 1971.

9. Meister, D. and Mills, R. G. Development of a Human Performance Reliability Data System, Phase 1. AMRL-TR-71-87. Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, August 1971.

10. Mills, R. G. et al. Sequential Task Performance: Task Module Relationships, Reliabilities and Times. Human Engineering Division, Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio (unpublished paper).

11. Smith, R. L. et al., Technique for Establishing Personnel Performance Standards (TEPPS), Results of Navy User Test, Report PTB-70-5, Vol. III, Bureau of Naval Personnel, December 1969.

12. Swain, A. D. Overview and Status of Human Factors Reliability Analysis. Proceedings of the Eighth Reliability and Maintainability Conference, Denver, Colorado, July 7-9, 1969, pp. 251-254.

13. Williams, N. G. Dependent Models for Estimating Human Performance Reliability. Proceedings os the Man-Machine Effectiveness Analysis Symposium, Los Angeles Chapter, Human Factors Society, 15 June 1967.

APPENDIX C

STUDIES OF THE INDEPENDENCE/DEPENDENCE VARIABLE

INTRODUCTION

In this section we describe two studies of the independence/dependence variable, the first performed by Mr. Robert G. Mills (and co-workers) of the Human Engineering Division of the Aerospace Medical Research Laboratory, the second by Mr. Gerald C. Lamb, of the U. S. Naval Underwater Systems Center. We have had occasion to comment upon this variable in relation to the combinatorial statistics employed by the analysis methods.

Subtasks and tasks may be independent of each other (i.e., the performance of one is not affected by the performance of the other), in which case their individual performance times and reliabilities can be combined multiplicatively to derive a performance value for some more molar behavioral unit which comprises 2 or more of these subtasks/tasks. Alternatively, these subtasks/tasks relate to each other dependently (i.e., the performance of one is affected by the performance of the other), in which case the nature of that dependency must be ascertained and explicitly accounted for in the probability statistics used to secure a higher order system effectiveness value. Manifestly many of the problems faced by the analytic models (and to a lesser extent by the simulation models) would disappear if the assumption of independence held generally, and in fact we have had occasion to note that analytic modelers often assume subtask/task independence merely to simplify their combinatorial problems.

The question of whether tasks exhibit performance independence or dependence is one which is not susceptible to a theoretical solution; it can only be answered by studies of the Mills/Lamb type. While their results are highly indicative, the reader must be warned that the problem is very complex and no one or two studies are sufficient to resolve it. A complicating factor is that, as Mills very astutely points out, there may be several varieties of subtask/task dependency. Consequently a single study may deal with only one type of dependency.

SEQUENTIAL TASK PERFORMANCE: TASK MODULE RELATIONSHIPS, RELIABILITIES AND TIMES (Mills, Hatfield and Bachert)

Mills et al. distinguish between what they call behavioral dependency and operational dependency. Operational dependency is illustrated by

"a series of instrument readings which may relate to some overall system performance measure, but the value obtained from reading one instrument has no effect whatsoever on the value that would be obtained from reading any other instrument". (All quotations are from Mills, et al.) Operational dependency is found "when alternative responses or courses of action... each of which is contingent upon a previous (task's) performance outcome or merely its completion (e.g., dialing a telephone)" Behavioral dependency is less clearcut. "Merely the fact that task modules (the term used by Mills et al. to describe behavioral units) must be performed in combination... might have an effect on HPR (human performance reliability) or task time which would be different from that expected based on performance of each task module separately".

What Mills et al. are saying (or so we interpret them) is that some dependency (operational) derives from the nature of the system configuration, its operating procedures, etc. Manifestly in the telephone example they cite each dialing action can be performed only after a preceding action has been completed.

Other types of task dependency may be performed concurrently, like reading a CRT and listening to a message, and there may consequently be some sensory interaction. Tasks may also be performed sequentially, although the start of one may not depend on the completion of the other. For example, since I am writing this while on an aircraft, when I finish writing this sentence, I will pick up my Martini, but there is no necessary dependence here; I could just as well drink as I write. The question is whether the temporal relationship of two logically independent actions influences the performance of either one. Must actions be logically interrelated before they exercise an effect on each other, if at all? If unrelated tasks 1 through 10 are performed sequentially, does the performance of task 1 have an effect on task 2 or on task 10? Can one think of temporal dependency as well as functional dependency? And if each exists, how is it affected by the nature of the task (e.g., perceptual, motor, cognitive) and the nature of the parameters influencing that task (e.g., speed and accuracy requirements, etc.)?

So the problem is not an easy one, because one must consider many factors that may influence human performance. The two studies described in this Appendix can therefore be considered only a first, although an important, step. (It might be thought that one could secure answers to these questions from already available experimental literature. Unfortunately, because the researchers represented in that literature were not

dealing with operationally meaningful tasks- for the most part, anyway- the data they provide is not satisfactory. For example, the extensive literature on serial effects in learning nonsence syllables provides few clues. Hence we need studies such as the ones reported in this Appendix.)

To study the problem we need subtasks which are essentially independent of each other but which can be synthesized into a more complex task. Performance on the latter, more complex task must then be predicted from a combination of the independent performance estimates obtained from the former subtasks. If performance on the more complex tasks can be estimated from performance of the independent elements of that task, then obviously the independence assumption holds.

The description of the study presented below is taken verbatim from the author's preliminary report. An incomplete version of the study is reported also in the U. S. Navy Human Reliability Workshop Proceedings (op cit).

Subjects

5 university students made up the subject sample. No special qualifications were required of them.

Apparatus

The apparatus used included a display/control device comprised of analog meters, digital read-out display modules, indicator/switch modules, and a numeric keyboard. Also included was a device for displaying single lines of a computer printout on a trial-by-trial basis and two books of tables.

The upper display panel contained 5 analog meters and 5 digital read-out modules. The analog meter located in the lower left corner of the panel was designated the Primary Meter and was labeled "X". The remaining 4 meters were used to display extraneous information which was not used in any of the experimental tasks.

3 of the 5 digital display modules were collectively labeled "Y". These modules were located in the lower right corner of the panel. The "Y" value was presented by lighting the display mechanism of the appropriate module. The lower panel of the display/control device consisted

of a key for each digit 0-9 inclusive and a key for the decimal point. A separate key was not available for the negative sign; therefore, the zero key was used for both zero and minus.

The 8 indicator/switch modules on the lower panel presented instructions and were also used as controls. An instruction was presented by lighting the indicator containing the appropriate legend instruction.

The device for displaying computer printout was controlled by the subject who, when instructed to do so by the appropriate indicator, advanced the printout to a new line of information by pressing a push-button on the side of the display device. A line of information was presented in a window located on the top of the device.

Also used were two books of "Z" tables. Z values were obtained from the tables using X and Y values as table coordinates. Z values for X = 0-69 were contained in one book and values for X = 70-149 in the second book. Y values ranged from -99 to +99 in both books.

The subject sat in a 3-sided experimental booth facing the display/control device. The booth was in a room separate from all programming and driving equipment.

The occurrence of all instructions and stimulus values was controlled automatically using continuous, punched paper tape. All subject control responses and input values were also recorded automatically on punched paper tape.

Tasks

Subjects performed each of 6 tasks, called by the authors "reference tasks". These were:

1. For reference task 1 - Meter Reference Task: The subject was instructed to read a value X displayed on the Primary Meter and input the value on the numeric keyboard. Coincident with the READ X instruction, a print-out timer was activated. Coincident with an END OF KEYBOARD INPUT (EOK) response by the subject, after inputting X, the time taken for the operation was recorded and the timer reset. Times were recorded to 0.1 second. The EOK response was made by depressing an indicator/switch module and is henceforth considered

implicit in any input operation. HPR is determined by later processing in which the subject's X input is compared with the correct X displayed. Although the variables compared for HPR differ, the time and HPR recording operations are also, henceforth, considered coincident with any input operation across the remaining tasks below. X ranged from 0 to 149.

2. For reference task 2 - Digital Read-Out Reference Task: This task module is identical to reference task 1 with the exception that a value Y is read from the digital diaplay. X ranged fro -99 to +99.

3. For reference task 3 - Table Look-Up Reference Task: This task consisted of two parts. First, the subject read X and Y from a computer print-out which displayed these values simultaneously. The instruction READ X, Y also instructed the sutject to advance the print-out listing one line to obtain the X, Y values for that trial. The X, Y values read were then input by the subject. The second part of the trial instructed the subject to look up a Z value using the tables. The tables were entered with the X, Y values read during the first part of the trial. Z ranged from -10.0 to + 10.0.

4. For reference task 4 - Compute Reference Task: This task module is similar in procedure to reference task 3 except X, Y and Z were obtained from the print-out and then input. The subject then computed a value Q using one of three formulas selected randomly. The three formulas used were: $Q = (X+Y)/Z$, $Q = Y \cdot Z/X$, $Q = Z/Y \cdot X$. Q ranged from -2480 to +2480.

5. For reference task 5 - Explicit Combined Reference Task: This task consisted of four parts. The subject sequentially obtained X, Y, and Z as in reference task 3. Q was computed as in the second part of reference task 4. In this task each operation occurs and is measured explicitly.

6. For reference task 6 - Implicit Combined Reference Task: The subject performed the same operations as in reference task 5 with the exception that he was not given procedural instructions to read X, read Y, and look-up Z. These operations are implicit and thus, are assumed to have been performed in order to compute Q. Thus, for this task module the subject was simply instructed to COMPUTE Q. HPR was determined as noted above; however, only subject and correct Q values could be compared.

Reference tasks 1-4 are considered to represent independent task modules. Reference tasks 5 and 6 represent combined tasks in which the independent task modules have been synthesized into a single task module 1.

In the case of reference tasks 3 and 4, however, independence is difficult to assure because the tasks involve operations performed under other task modules by using X, Y, and Z. By displaying these latter values simultaneously via a different display mode (i.e., computer print-out) and by requiring separate input on each trial, an attempt has been made to obtain HPR and time measures which can be attributed only to the table look-up (i.e., entering X and Y into the tables, extracting, and inputting Z) and compute (entering X, Y, and Z into the formula, calculating and inputting Q) aspects of reference tasks 3 and 4. In other words for the look-up module perfect reading of X and Y must be assumed because imperfect reading is being estimated in reference tasks 1 and 2. The same is true for the compute module except for the inclusion of Z which is being independently estimated in reference task 3.

The separate inputs of these tasks served the additional purpose of reducing the likelihood of reading errors because the values were available simultaneously on the print-out and input errors could be cross-checked during analysis to remove their effects on either the look-up or compute task modules. Thus, using this approach it was hoped that the operational dependency of the look-up and compute task modules upon the task modules of reading and inputting X and Y and, for reference task 4, looking up Z could be removed.

Procedure

Subjects performed task modules in blocks of 50 trials each. The eventual total number of trials selected was 400 (8 blocks) on each of the reference tasks 1-4. An additional 50 trials (9 blocks) on reference tasks 5 and 6 were completed for reasons to be indicated below. The start of each trial was separated from the previous trial by a random inter-trial interval ranging in one-second intervals from 5 to 15 seconds. The use of variable inter-trial interval decreased subject anticipation of the start of each trial.

Subjects completed blocks in consecutive, daily sessions of one block per reference task per session. It took very little time to complete a

block for reference tasks 1-3 (e.g., one block for reference task 1 could be completed in 15-20 minutes) and considerable time to complete a block for reference tasks 4-6 (e.g., one block for reference task 5 took approximately 1 hour to complete). The length of time for each daily session was equated by having subjects perform on two tasks per session. Each pair of tasks to be completed during a given session was randomly comprised of one of the reference tasks 1-3 and one of the reference tasks 4-6. This procedure resulted in daily sessions ranging from 1 1/2 - 2 hours for each subject.

Upon completing 8 blocks of each reference task, subjects completed an additional block of reference tasks 5 and 6 during the last two sessions. This procedure was used because it was considered advantageous to terminate each subject with data collection on the combined task modules for eventual prediction purposes. No use was made of this procedure in the analysis, however.

In their first session subjects were given a set of written instructions detailing the operations to be performed on each of the reference tasks. Following the written instructions, subjects were allowed to ask questions and they worked through a practice session of 5 trials for each reference task. They were also told that time and errors were being recorded and that they should work at a comfortable pace emphasizing accuracy of performance.

Results and Conclusions for Task Module Performance Time

Results are available only for performance time. Performance reliability data are available, but as of this date of this report had not been analyzed.

Effects of Learning

The effects of learning across trials were estimated by comparing the medians of distributions of task performance times obtained for each task module and block of 50 trials collapsed across subjects. Individual subject distributions were representative of group distribution. In this and remaining analyses only the performance times for the second parts of reference tasks 3 and 4 are considered.

Performance times tended to decrease from the first to the last block for all task modules. However, the decrease was substantial only for those modules involving computation. Thus, the average decrease in median task time from block 1 to block 8 for the non-computation task modules of reference tasks 1-4 and 5 was .4, 10 seconds as compared with 21, 18 seconds for the computation task modules of reference tasks 4-6.

Of necessary interest here however, is whether task times stabilized after an initial number of blocks. This is important because in order to make comparisons across task modules the effects of learning must be assumed to be negligible. An analysis of variance for each task module was performed using individual subject median times and a repeated measures design across blocks. Using successive analyses it was determined that significant differences ($p < .01$) among medians across blocks were not present if blocks 1-4 were excluded from the data for each of the task modules.

From this it can be concluded that learning had stabilized after 200 trials regardless of task module or subject. As a result, later analyses presented below rely on distributions of task times that have been collapsed over subjects and blocks 5-8 for reference tasks 1-4 or 5-9 for reference tasks 5 and 6.

Effects of Combining Task Modules Explicitly

Figure 17 presents histograms of the empirical distributions obtained for the independent task modules of reference tasks 1-4. Figure 18 presents histograms of the empirical distributions obtained for the explicitly combined task modules of reference task 5. Figure 19 presents the histogram of the empirical distribution obtained for the implicit reference task 6. Each figure shows the probability of a given task time for each task module and the flow of operations is depicted across the top of each figure. Also shown are the statistical parameters, sample size (N), median (mdn), mean ($\bar{X}$), and standard deviation (SD) of each distribution. The figures have been positioned so that an easy vertical comparison between distribution of an independent and explicit consized task module can be made.

N in these figures is based on 5 subjects and 4 blocks of 50 trials each for reference task 1-4 (i.e., N= 1000) and 5 blocks for reference tasks 5 and 6 (i.e., N=1250). Actual Ns are somewhat less than these values because of an occasional loss of a time value due to equipment difficulties.

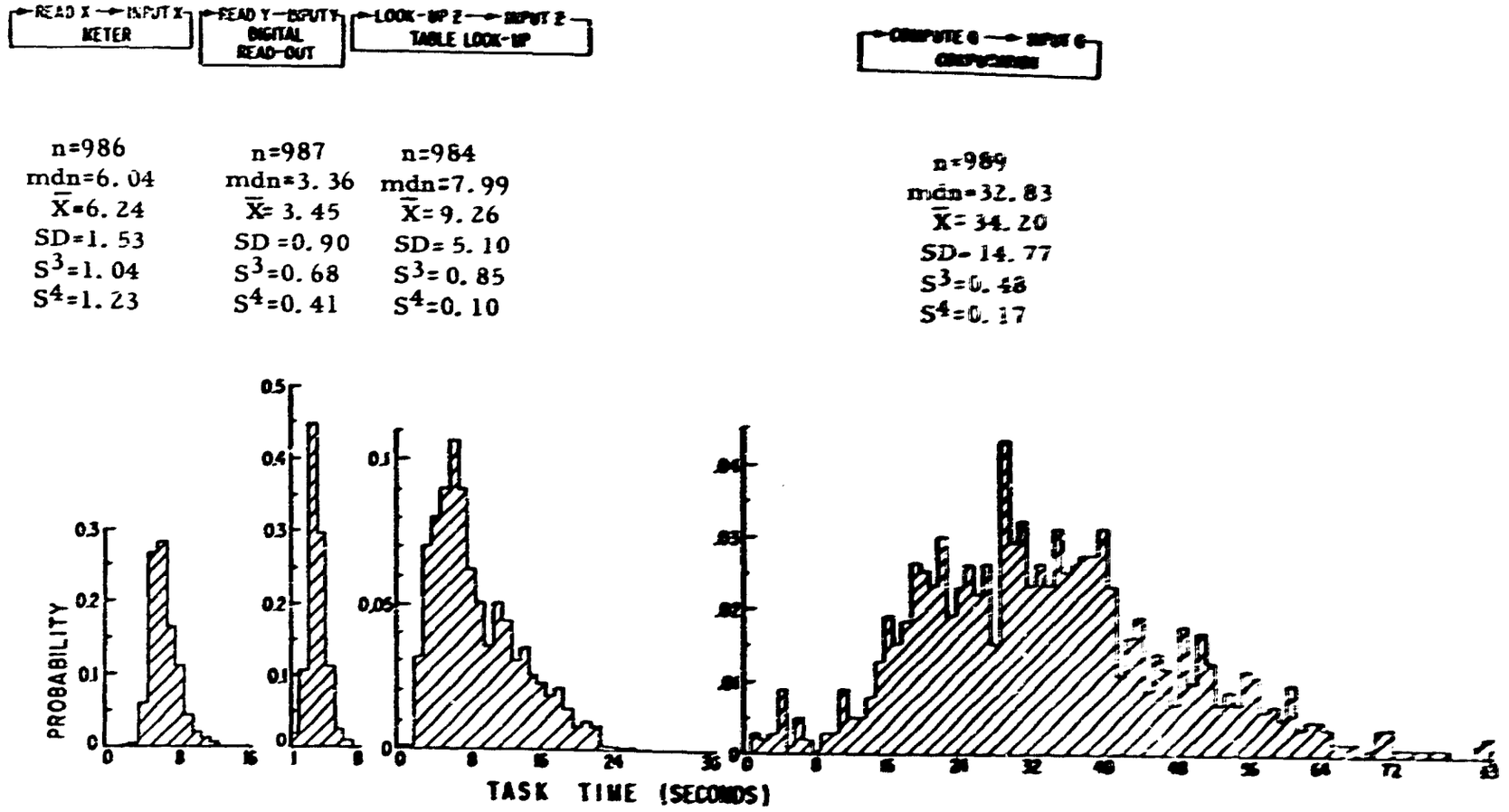
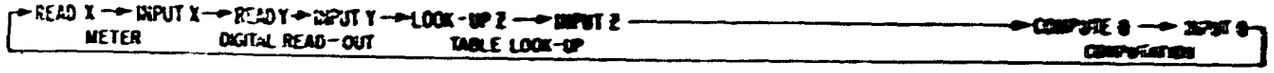


FIGURE 17. Responses to Independent Tasks



n = 1226	n = 1230	n=1225	n=1236
mdn=6.31	mdn=3.33	mdn=8.36	mdn=30.04
$\bar{X}$ = 6.67	$\bar{X}$ = 3.47	$\bar{X}$= 10.80	$\bar{X}$ =31.19
SD=2.09	SD=1.04	SD=7.05	SD=14.97
S^3 = 1.12	S^3 =1.24	S^3=1.35	S^3=0.50
S^4 = 1.84	S^4 = 1.98	S^4=1.28	S^4=0.15

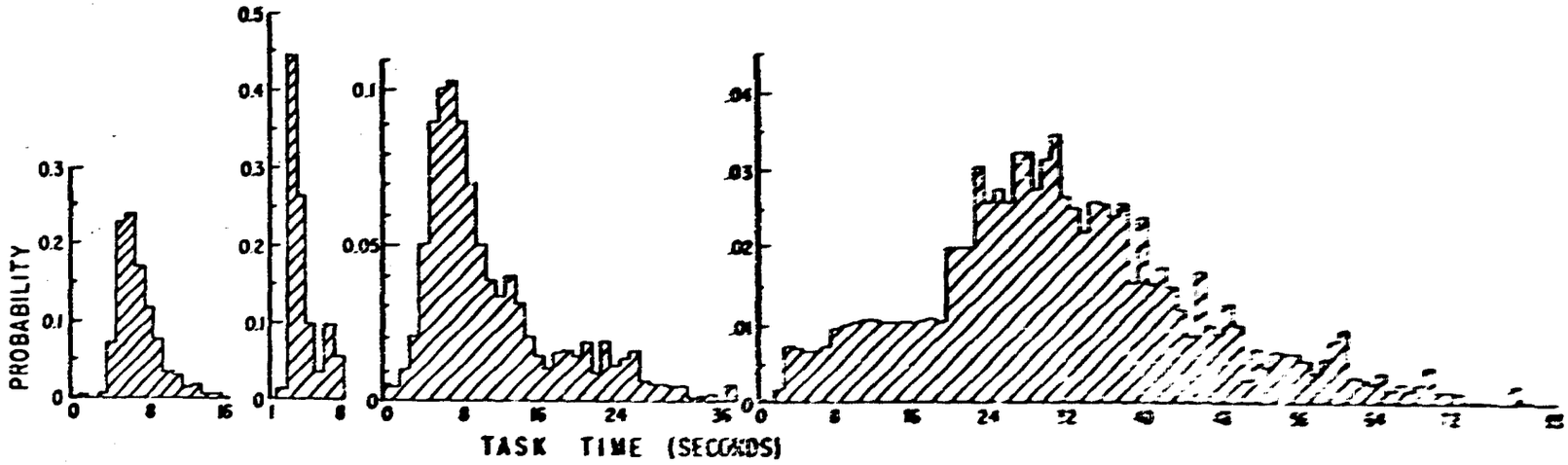


FIGURE 18. Responses to the Combined Explicit Task

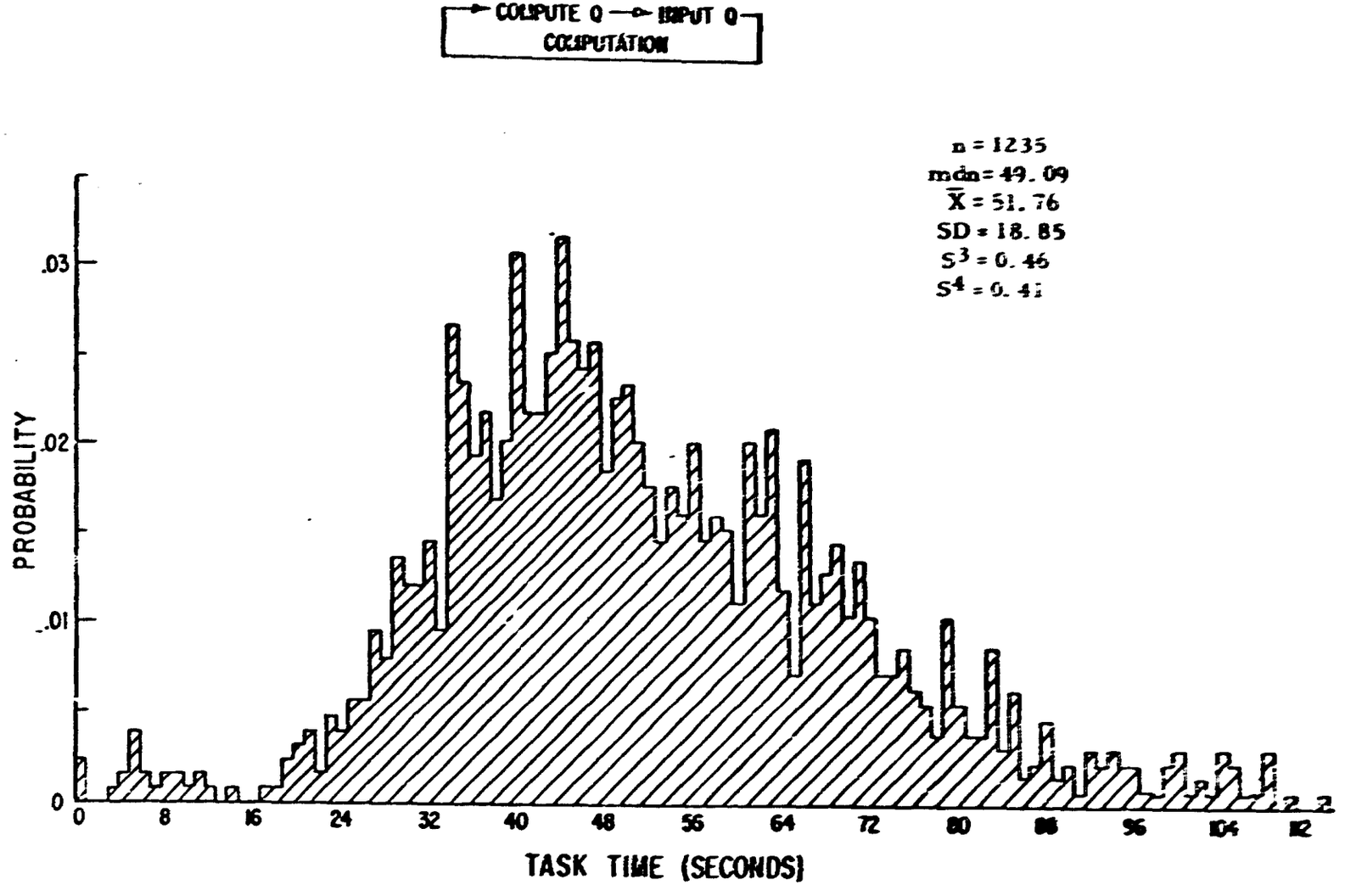


FIGURE 19. Responses to the Combined Implicit Task

In comparing the distribution of differing task modules in Figures 17, 18 and 19 it is clear that task times vary considerably depending upon task module. In addition none of the distributions appear to be normal. Determination of third and fourth moments indicates that all are positively skewed and leptokurtic. The impact of deviations from normality will be dealt with further below.

In comparing Figure 17 with Figure 18 substantial differences are not apparent between the performance times of an independent task module and the same module combined explicitly. A t-test performed on the difference between the independent and explicit means for each task module indicated only the table look-up module achieved statistical significance ($t = 2.55$, $p < .01$, $df = \infty$). With the large N involved, these tests are quite conservative (nearly any difference will be significant) and the observation that these distributions are not normal make the test suspect. Also the probability of obtaining the larger times of the explicit table look-up module is quite small. Therefore, it is unlikely that this difference is of practical importance in the present study.

However, there are subtle but systematic differences between these distributions which are worth noting particularly as they may apply in latter studies. In the first place the variability of task times increase when a task module is combined explicitly. For example, the standard deviation obtained for the independent table look-up module increased by 38.2% when it was combined explicitly.

Secondly, with the exception of the computational task modules, the third and fourth moments also increased somewhat from independent to explicit combined task modules. This fact in conjunction with the increase in variability of task module times suggests that the distributions of explicitly combined task modules have also increased in positive skewness and have become less peaked. The computation task module showed little change in its statistical parameters except for median and mean, from independent to combined explicit reference tasks.

In order to determine the degree of interrelationships among explicitly combined task module's performance times, bivariate correlations were obtained between times within trials. These correlations ranged from 0.37 between the digital read-out and the table look-up task module times to 0.19 between the meter and computation module times; they are quite small and indicate only slight dependencies among the times.

The true magnitude of these dependencies is reduced further when they are evaluated relative to correlations obtained among independent task modules within the same trial number (this represents artificial combination based only on having the same trial number). The value of these later correlations would be expected to be near zero. However, they actually ranged from 0.33 to 0.13 indicating that the times of the independent task modules were also correlated slightly. These unexpected correlations are probably due to a combination of two things, residual effects of learning (i.e., times tend to decrease across trials regardless of task module) and the fact that one fifth of each sample was obtained from the same subject. Thus, trials and intrasubject variability become correlating factors.

These factors would also be present in the correlations obtained among the combined explicit task modules. The true degree of dependency among explicit combined task modules is reflected by the differences between their correlations and those obtained among the indepentent task modules. These differences ranged from 0.14 to -0.02; with the exception of this latter difference , all differences were in a positive direction indicating a slight increase in dependency among explicit combined task modules over those treated independently.

Based on the above analyses there is some evidence that task module performance time is differentially affected depending upon whether the task modules are combined or not. However, there are also suggestions that the effects of combining may be slight implying little overall impact on prediction based on the use of the summation rule for synthesizing task module performance times. The issue of prediction is addressed more closely immediately below.

Prediction of Combined Task Module Performance Time

A Mont Carlo computer simulation program was used to examine the predictive capability of independent task module performance times. Principal interest will be in attempts to predict the performance time of the implicit combined reference task. The general procedure used was as follows.

1. A probability density function (PDF) was derived from the empirical distribution obtained for each independent task module. Three types of PDFs were fit to the empirical time data. The first was the

Weibull and its parameters were derived by employing the linear regression method of fit following the lead of Askren and Regulinski. The second was the Weighted-Weibull. This PDF derives Weibull parameters from data fit by linear regression and which have been weighted by their probabilities of occurrence. The third PDF employed was the Gaussian normal using the empirically derived mean and standard deviation as inputs to a computerized Gaussian random number generator. Although represented as histograms in Figures 17-19, all distributions were assumed to be continuous in these derivations.

2. A random sample of n= 1000 simulated task module times was drawn from each PDF.

3. Using the summation-rule, the times sampled first for each module were summed across modules. The resulting time became the first simulated time in the predicted distribution of times for the implicit combined task. Each successive sampled time was also summed across modules until a predicted distribution with n = 1000 was obtained.

A set of 50 samples each with n = 1000 was drawn for each independent module, thus creating 50 predicted samples each with n = 1000 for the implicit module. Comparisons were also available between the empirical distribution for each independent module and its derived PDFs.

4. The predicted implicit performance time distributions were compared to those obtained empirically.

5. The above procedure was also applied to the explicit task modules. This allowed further comparison with the independent modules as well as, with the implicit module.

The results of these tests indicate that the normal assumption does not generally yield the best fit for independent task modules. Only for the meter task module does the normal appear as if it is best. The normal is able to predict the empirical means quite well but offers no clear choice relative to the other PDF types when predicting the empirical standard deviations.

The poorer predictive capability achieved under the normal assumption is indicated further in attempting to predict implicit task module performance time. In this case the normal fit clearly ranks third behind the fits achieved by the Weighted-Weibull and the Weibull, respectively.

Attempts to predict the performance time of each explicit task module from independent task module performance time was best under the normality assumption in only one case. The normal yielded the best fit for the meter task module. The Weighted-Weibull yielded the best fit for the digital read-out module and the Weibull yielded the best fit for the table look-up and compute modules.

These results do not completely rule out the normality assumption for synthesizing task modules because the error in predicting implicit task performance may have resulted from a failure to account for dependencies among modules. The validity of the summation-rule for combining performance times was examined by attempting to predict, as closely as possible, the implicit task time. This was accomplished by simulation using a mixture of the better PDFs and synthesizing under the summation-rule.

The results of these simulations then provide evidence that the assumption of independence regarding task module performance time is valid. This statement is based on the fact that close approximation to a distribution of empirical task times can be achieved by summing the independently derived performance times of its task modules. This conclusion is also supported by the fact that substantial correlations were not found among the performance times of the explicit task modules.

However, it should be recalled that slight dependency effects of combining task modules explicitly were obtained. Presumably, these effects could account for much of the remaining error in the mixed simulation. Furthermore, it is possible that this error might become significant as task modules are added or increased in complexity, thus magnifying the effects of dependency. Examination of the data indicates that the magnitude of prediction error for independent module performance times does not tend to increase or decrease systematically as a function of task module operation. This suggests the error in prediction of synthesized task modules will accumulate as a function of the number of task modules only and not their complexity. This of course would hold only if the PDF fit were accurate. Determination of the magnitude of this constant error, if it exists, and the synthesis model to account for it should however, wait until additional data are collected.

The simulation results also indicate that the normality assumption should be rejected as a general assumption in synthesizing task performance times. This conclusion is based on the fact that prediction of

task performance time is not as accurate using normally distributed times as alternative PDFs. This statement is also supported by the comparisons of independent and explicit task module performance which indicates that distribution of times are not normal.

This is not to say that task module times will never be normally distributed nor that the normal PDF cannot be used. Only that on most occasions the normal PDF does not produce PDF the best fit to the data. From a practical point of view the error produced by violating a non-normality assumption may not be large enough for one's purpose to warrant using a better fitting PDF. The general impact using a normal PDF remains to be evaluated.

A TEST OF A BASIC ASSUMPTION OF HUMAN PERFORMANCE MODELLING (Lamb)

Lamb's study tested the independence/dependence variable in an operational setting in which the task was symptom detection and fault location of sonar system failure.

The sonar equipment used was an operational sonar simulator at the Naval Underwater Systems Center. The failure symptoms were concentrated in the passive receiving console. This equipment is in fleet use with many man-years experience in its operation and maintenance.

The faults chosen were common component failures each with an obvious effect on the face of the sonar console (Table 20). The faults were chosen so that operators using the maintenance troubleshooting manual for the system could isolate them. They were not unusual or unrepairable types of faults.

Subjects

Twelve Sonar Technicians familiar with the sonar or similar systems served as subjects. Their experience with the equipment ranged from six months to four years. Their normal duties included system maintenance.

TABLE 20. FAULTS USED FOR EXPERIMENT

Fault No.	Symptom	Cause
1	No recording on Bearing Time Recorder	Tube failure
2	No paper and stylus movement	Blown fuse
3	No movement on manual bearing indicator	Blown fuse
4	No audio from speaker	Tube failure
5	Right/left indicator	Tube failure
6	No stylus movement	Drive chain off

TABLE 21. EXPERIMENTAL DESIGN

Subject No.	Number of failures and specific faults		
	3	2	1
1	1, 2, 3	4, 5	6
2	1, 2, 4	3, 6	5
3	1, 2, 5	4, 6	3
4	1, 2, 6	4, 5	3
5	1, 3, 4	2, 5	6
6	1, 3, 5	2, 6	4
7	1, 3, 6	2, 4	5
8	1, 4, 5	2, 6	3
9	1, 4, 6	3, 5	2
10	1, 5, 6	2, 3	4
11	2, 3, 4	5, 6	1
12	2, 3, 5	1, 6	4

Experimental Design

Each subject participated in three trials; one trial had one fault, another two faults, and the final trial had three faults (Table 21). No particular fault occurred more than once for any subject. The order of the trials was randomized over subjects to minimize the effects of practice. All possible combinations of the six faults were used to generate the trials.

Each subject had all three trials in one session. The subject was read the instructions, given the maintenance manual, and allowed to isolate the fault using any procedure he desired. The subject could at any point indicate that he needed and was going to use test equipment and was told the results of the test by one of the experimenters. Actual test equipment was not used in order to save the time associated with the manual tasks of setting up the tests since the experiment was concerned with fault isolation.

The time to isolate a fault was used as the dependent variable. Thus each major maintenance action was recorded and timed. In addition, protocols were recorded for questions about conformity of the experimental problems to normal ship-board faults and maintenance procedures.

Results

Quantitative

The basic measures were time to isolate each fault in a given trial, and total time to complete each trial (where a trial had 1, 2, and 3 faults). Total time to solve three faults ranged from 5 to 60 minutes, for two faults from 2 to 48 minutes, and for one fault from 1 to 44 minutes. Time to isolate <u>each</u> fault is more indicative of the operator's ability with multiple faults and should vary if independence is not met. The average time to solve <u>each</u> fault was calculated for sets of:

1 fault	11.47 min.,
2 faults	10.23 min., and
3 faults	8.65 min.

These data suggest that multiple faults are being more quickly solved than single faults. However, a statistical comparison did not bear this out ($t = 0.62$, $p > .05$).

There is extreme variability in the time to isolate any single fault over both subjects and conditions. In order to remove some of this variability, specific faults were analyzed to see if the time to isolate them increased or decreased when combined with other faults. There was a tendency for times to decrease when combined with other faults. The average times were:

alone	0 (base line)
with one other	-1.83
with two others	-3.53

Again, a statistical test between the extreme values (alone and with two others) showed no significant effect ($t = .910$, $p > .05$). Therefore, the quantitative data showed independence for fault isolation at the task level when multiple faults were present.

Qualitative

Since the quantitative data had not led to rejection of the null hypothesis (independence of maintenance actions), the qualitative data were examined to see if they confirmed the results of the quantitative data.

The experimenter's taped observations did, in fact, support heavily the idea of independence of fault isolation. Subjects were about evenly divided as to whether they detected all or only one symptom at a time. However, they all proceeded to isolate faults sequentially, beginning entirely anew with the isolation of each fault. This even included closing the manual and reopening it. Several subjects even commented, "Finished with that one, time for the next one." There was only one exception: One subject isolated two faults simultaneously because they were adjacent blown fuses. This was an experimental consequence of using all possible combinations of faults.

It is clear from both the quantitative and qualitative data that at the task level there is independence of maintenance actions of the type used in this study.