722689

NAVSHIPS
0967–412–4010

PROCEEDINGS OF U.S. NAVY HUMAN RELIABILITY WORKSHOP

22–23 JULY 1971
WASHINGTON, D. C.

SPONSORED BY

NAVAL SHIP SYSTEMS COMMAND
OFFICE OF NAVAL KESEARCH
NAVAL AIR DEVELOPMENT CENTER

D D C
MAY 5 1[illegible]

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

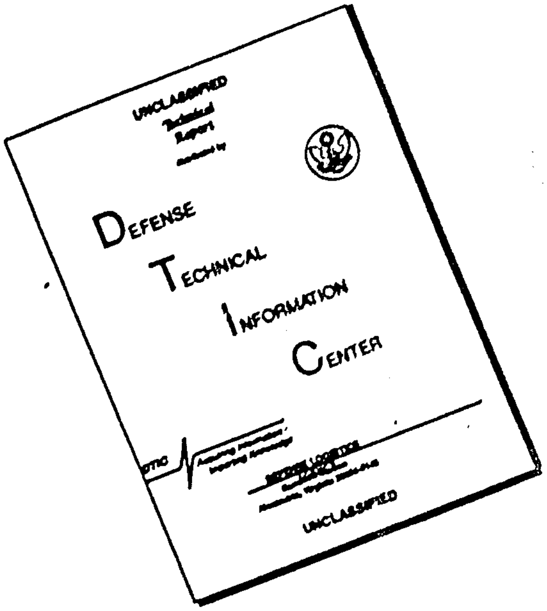
FEBRUARY 1971

DEPARTMENT OF THE NAVY
SHIP SYSTEMS COMMAND

NATIONAL TECHNICAL
INFORMATION SERVICE

DEPARTMENT OF THE NAVY
WASHINGTON,D.C. 20360

DISCLAIMER NOTICE



UNCLASSIFIED
Technical
Report
[illegible]

DEFENSE
TECHNICAL
INFORMATION
CENTER

DTIC [illegible]

[illegible]

UNCLASSIFIED

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

PROCEEDINGS OF U.S. NAVY

HUMAN RELIABILITY WORKSHOP

22-23 July 1970

Washington, D. C.

Sponsored by

Naval Ship Systems Command
Office of Naval Research
Naval Air Development Center

James P. Jenkins
Editor

February 1971

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

SUMMARY

The purpose of the Navy Human Reliability Program, under ADO 43-13X, is to develop and test human reliability and availability models and techniques for incorporation in system reliability, availability and effectiveness prediction. The workshop was a necessary step in achieving the cohesive program the goal requires (P. 2 - 23).

To meet the purpose the program must consider the methods, criteria, operational needs and problems in human reliability analysis (P. 24 - 84). Yet, a clear cut approach is not necessarily evident because of the complex nature of the human reliability problem itself. Several studies are presented which amplify this conclusion (P. 85 - 207).

The Navy's program in human reliability did not begin from a zero point in research. Two Navy laboratories and two contractors had begun a multi-phased study to meet the program objectives, prior to the workshop (P. 208 - 296). Naval Underwater Systems Center, New London, is attempting to relate human reliability values from maintenance tasks with the equipment reliability of the maintained system (P. 208 - 234). A contractor, Applied Psychological Services, is applying a previously developed human reliability prediction technique to two electronic systems' maintenance subsystem to refine the technique and validate it so as to provide engineering and system designers with a means of accurately computing human reliability early in the design phase (P. 235 - 274). Naval Electronics Laboratory Center and another contractor, Human Factors Research, Inc., are establishing a data bank and automatic data extraction system for command and control systems, including the Naval Tactical Data System.

The discussions following each paper and at the conclusion of the presentations do not always flow in a concise manner, nor were they expected to be cut and dried. Rather, the sense of the participants' give and take in the discussions reflects recognition of the several problems highlighted in the papers, such as reliability models and their adequacy, the focus of human task levels at the macro or micro stage, the nature of a human performance data bank and requirements to assess and compare current human reliability models.

The discussants agreed that an evaluation of the various models mentioned in the workshop was absolutely necessary to ensure that the total program rests on a firm foundation. The Program Manager concurred and funds were set aside to support this new research task. The attendees also were willing to continue to act as an advisory group to the Program Manager and future annual meetings would be held (P. 297 - 324).

HUMAN RELIABILITY WORKSHOP ATTENDEES

Mr. Donald Aldrich
Naval Underwater Systems Center
New London Laboratory
New London, Connecticut 06320

Dr. James W. Altman
Data Graphics Inc.
4790 William Flynn Highway
Alison Park, Pennsylvania 15101

Dr. Robert E. Blanchard
Integrated Sciences Corporation
3122 Santa Monica Blvd.
Santa Monica, California 90404

Mr. Richard Coburn
Human Factors Technical Division
Naval Electronics Laboratory Center
San Diego, California 92152

CDR. Michael Connery USN
Office of the Chief of Naval Operations
OP 0701E
Pentagon
Washington, D.C. 20360

Dr. Victor Fields
Office of Naval Research
Code 458
Washington, D.C. 20360

Dr. William Harris
Human Factors Research, Inc.
Santa Barbara Research Park
Goleta, California 93017

Mr. Robert L. Howard
Naval Ship Systems Command
Code 00V131
Department of the Navy
Washington, D.C. 20360

Mr. James P. Jenkins (Workshop Chairman)
Naval Ship Systems Command
Code 00V13
Department of the Navy
Washington, D.C. 20360

Dr. Jerry C. Lamb
Naval Underwater Systems Center
New London Laboratory
New London, Connecticut 06320

Dr. Robert Mackie
Human Factors Research, Inc.
Santa Barbara Research Park
Goleta, California 93017

LCDR Paul Chatelier USN
Naval Air Development Center
Human Factors Department
Westminister, Pennsylvania

Dr. David Meister
Bunker-Ramo Corporation
8433 Fallbrook Avenue
Canoga Park, California 91304

Mr. Robert Mills
Aerospace Medical Research Laboratories
Wright-Patterson Air Force Base, Ohio 45433

Mr. Thomas A. Momiyama
Naval Air Systems Command
Code 3034B
Washington, D.C. 20360

Professor Thadeous Regulinski
School of Engineering
Air Force Institute of Technology
Wright-Patterson Air Force Base, Ohio 45433

Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Street
Wayne Pennsylvania 19078

Dr. Alan Swain
Sandia Laboratories
Division 1644
Alburquerque, New Mexico 87115

Dr. Martin Tolcott
Office of Naval Research
Code 455
Washington, D.C. 20360

PURPOSE OF U.S. NAVY HUMAN RELIABILITY WORKSHOP

James P. Jenkins

Sonar Systems Office
Naval Ship Systems Command

PURPOSE OF U.S. NAVY HUMAN RELIABILITY WORKSHOP

INTRODUCTION

A perusal of past conferences, symposia, and papers discussing human reliability reveals the constant cry - let's have an organized and systematic, service-sponsored human reliability program. This workshop says, the Navy has the watch and the orders are clear. The word has been passed from Chief of Naval Operations. The results of this gathering will set the course and speed.

Each participant has been specially invited because of his ideas and contributions to the development and application of human reliability, whether in concepts, models, techniques or applications.

If such talent exists, why then a workshop? What special purpose or function can it serve in the light of prior research and in view of the several past symposia on the subject. Simply this: the Department of the Navy, has a multi-year, interdisciplinary, service-wide program to develop and test a human reliability model, including data bank generation, for inclusion in system reliability, availability and effectiveness prediction. Past research and ideas are important, but obviously the Navy's current program was not known at the time research was begun. Therefore, the purpose of the workshop is to bring to bear what is known and what is required in human reliability to meet the Navy's objectives. The results of this workshop will then form the structure, the framework of research for the next several years.

The human reliability program is a part of a total human engineering advance development program. Its purpose as described in the ADO is: "An advanced system to achieve effective integration of the human operator into the weapon/ support system will be demonstrated. This development includes a sub-system of function allocation, effectiveness assessment, reliability evaluation, and interfaces related to maintainability and maintenance. A related development in aviation training success prediction and adaptive schedules is to be tested." Cdr. Connery of the Office of CNO and Mr. Momiyama of NAVAIR will present a further exposition of the ADO.

The workshop's genesis began in early 1970 where the ADO's principal developing agency and technical director, NAVAIR 303, in the person of Mr. Momiyama, and with concurrence of Cdr. Connery, requested me to be the project director of the human reliability program. The System Effectiveness Branch, Sonar Systems Office, my organization, had started human reliability research in 1966 and continued it within funding constraints. A review of the literature, current research and Navy needs was done and the following conclusions were reached:

- Validation of certain HR methods and modes was necessary

- The integration of HR models with equipment reliability models to predict overall system reliability was necessary.

- The most critical Navy weapon subsystem affected by HR was command-control.

- A data bank was essential.

- This workshop, for reasons already described, was necessary.

Navy laboratories were queried and the Naval Underwater Systems Center, New London (formerly called Underwater Sound Laboratory) and Naval Electronics Laboratory Center responded. Two contractors, Applied Psychological Services and Human Factors Research, were also responsive. The particular programs each has just begun will be described. The workshop will examine the programs and make recommendations for each. If change is necessary changes will be made. Other programs in and out of the Navy will be done on the basis of responses. The schedule and budget requirements of the HR program will be given. This may modify some recommendations.

The agenda has been arranged such that papers are presented first and the three programs in HR by Navy and contractors are given next. So, on the one hand we shall define HR requirements and on the other we shall have described responses to these requirements. From the verbal melting process the structure and form of the Navy's HR program for the next several years will emerge.

OVERVIEW OF PROJECT W43-13X
HUMAN FACTORS ENGINEERING TECHNOLOGY
PRESENTED TO THE NAVY HUMAN RELIABILITY PREDICTION WORKSHOP

Thomas S. Momiyama
Naval Air Systems Command

OVERVIEW OF PROJECT W43-13X
HUMAN FACTORS ENGINEERING TECHNOLOGY
PRESENTED TO THE NAVY HUMAN RELIABILITY PREDICTION WORKSHOP

(Viewgraph 1)

Jim Jenkins told me to give you an overview of the human factors project, a part of which is the effort addressed by this workshop.

(Viewgraph 2)

Before I go into my discussion, perhaps I should show you how our Working Party is organized. Here is the Project Organization Chart.

Mike Connery has just told you the OPNAV side of the story. The big bosses in the Pentagon tell us what they want to see as an end-product for the money they put up. Our job in the Systems Commands, of course, is to provide the task of "how to get there from here." Requirements documents are sometimes specific and sometimes not so specific.

(Viewgraph 3)

ADO 43-13X is specific in that it wants five areas of human factors technology developed. These are:

Human Reliability Prediction System, the very subject of the meeting

Automated Man/Machine Function Allocation System

Human Factors Test and Evaluation System

Integrated Job Aids System

Aviation Training Success Prediction and Adaptive Schedule Test

Here I have begun to categorize the human reliability prediction effort. The specificity of the ADO, however, seems to become lost quickly about here. These five areas, save one or two items, indicate that we are talking about some basic problems of human factors.

Logical questions might be: Where does Project W43-13 belong in the Navy RDT&E program? What role does it play in human factors inputs into Navy systems development? What specific Navy systems are we talking about? And possibly several others.

(Viewgraph 4)

The title of the project implies some of the answers. It is a development of technologies or methodologies we seek, rather than hardware as normally is the case for Systems Command projects. It is also human engineering work, as contrasted to life support or personnel work, per se. This we consider is significant because human engineering effort has <u>in the past</u> been <u>always an adjunct</u> to various hardware systems development and not a concerted and direct effort.

Thus, ADO states its objective as: "To develop an advanced system (an obviously broad meaning of the term system) to achieve effective integration of the human operator into the weapon and support system.

(Viewgraph 5)

ADO Themes

I may be able to do a little better "pigeon-holing" if I dwelled a bit on several explicit themes of the ADO and interjected some relevant notes.

One theme, is the need to establish human engineering effort beyond the life support requirements and personnel management and planning as I have just mentioned. The important historical event on this matter was the establishment in FY-69 of this human engineering project (W43-13) separating it from Project 43-07 called Manpower Productivity and Management. This manpower project was established in FY-67 with the express purpose of improving levels of personnel performance and readiness in naval weapons systems operation. The development of new techniques in man-machine trade-off was one of the several manpower-management development objectives. Now, that portion is made an independent effort to be pursued by Systems Commands, or by engineers who develop hardwares. The manpower management project is ongoing by the way under Bureau of Medicine and Surgery and Bureau of Naval Personnel to provide that important development. So we are in one of the two major, generic, human factors efforts in the Navy.

Although I said, "Human engineering beyond personnel management," I by no means imply "exclusive of these considerations." Since any improvement in man-machine trade-off will naturally impact on training and planning of the force posture, awareness or even active balancing of system-sophistication against training pay-off will be our major concern.

This brings us to another theme of the ADO: that is, the lack of adequate training and technical manuals. They are inadequate, often because a system is designed with a very unfair assumption of zero human defect or human adaptability of almost infinity. We are going to take a hard look at the human factors accommodations in the information presentation as one of the five ADO goals, as well as alleviating post-design quick-fix treatments. I will discuss the last point on extending the current capability in the next viewgraph.

(Viewgraph 6)

Extending Current Capability

How does this broad human factors approach tie into the myriad of specific subsystems human engineering efforts - such as F-14 aircraft cockpit layout, and AGILE missile handling techniques? There is a mandatory Human Factors section in all system/subsystem technical development plan documents such as PTA and TDP. In other words, there *is* a requirement to provide an appropriate human factors implementation in the complete development cycle of a particular hardware.

Now, the guideline for this implementation is the military specification MIL-H-46855: *Human Engineering Requirements for Military Systems, Equipment, and Facilities*. This Mil Spec "touches on" a broad spectrum of human engineering needs for system development. The problem is that it does not, because of its intended scope, go much beyond statements of the needs, such as "Human engineering effort shall include ..... systems engineering to identify, define, and allocate ..... functions to man, equipment, or man and equipment." The Mil Spec refers to a military standard (MIL-STD-1472) entitled Human Engineering Design Criteria. This STD provides some specific guidelines. But these criteria, such as anthropometric measurement, are static; and having complied with them, one is not always assured of adequate man-machine interfacing in the dynamic, operational situations - again until post-design observations and analyses are made, and this is often too late. Thus, the Mil Spec is a "Thou shalt not ignore human engineering" document; it is not a "How to do it" text. This is the current capability.

Human factors technologies developed in our project, then will contribute to a collection of generalized and realistic techniques for meeting the human factors requirements, i.e., "How to do it." We shall develop and demonstrate the feasibility of analytical and simulation techniques for effecting function allocation as well as quantitatively assessing and accommodating human performance in system operation. An important contribution would be the various data banks of human performance parameters, lack of which limits the usefulness of many advanced methodologies developed to date.

(Viewgraph 7)

I might mention that Project W43-13 is an R&D Category 6.3 or an Advanced Development project. This means that we are concerned with: technical and cost feasibilities, military usefulness, and experimental system development. We are talking about finding specific Navy use for new human factors technologies such: Man-machine system digital simulation, reliability models, and computerized function allocation techniques. And Mike expects us to take these developments right down to the threshold of Category 6.4., Engineering

Development, or to the point of showing marked improvement on human engineering of Navy systems development.

(Viewgraph 8)

A note of significance may be made on the general trends in human factors consideration. Scanning through various DOD policy-making documents such as Joint Research and Development Objective Document, Navy Strategic Study for 1976 - 90, Marine Corps Mid- and Long-Range Objectives, etc., I see that human factors and social sciences applications have generally captured a section as big as major warfare plans. Specifically, these documents urge, among other things: systems approach to human factors problems, improved human performance measurement, and development of methodologies making use of mathematical models and simulation techniques. This may sound like a common-sense notion, but explicit messages in documents at this level give us a lot of momentum in heretofore "No. 2" effort in R&D. All I am saying is that what we are pursuing in this project is on the right track, or at least is the DOD way.

(Viewgraph 9)

Technical Development Plan

Based on our interpretations of the ADO and hierarchiacal mandates there are several basic philosophies or developmental characteristics we would like to maintain.

(Viewgraph 10)

Here is the block diagram of the project development. At the top are various stages of Navy system development and factors requiring human engineering inputs. In the center of the diagram are five ADO-required development area, (integrated job aids and information present technology are essentially same development) underlined by the total system effectiveness evaluation and personnel management inputs. Solid-lined blocks are the development within this project while chained blocks are developed elsewhere.

Two technologies, human reliability prediction and function allocation techniques, are mandatory during early stages of system design, to ensure optimal use of human capabilities as an operator and maintainer. Human performance and reliability standards, of course, are first needed to accomplish adequate and timely function allocation.

Human factors test and evaluation methods relate quality and quantity of human performance to total systems effectiveness. They determine the degree to which the early human factors efforts have met system operational requirements.

The integrated job aids and aviation training success prediction system are fleet support items. Job aids concepts of information presentation techniques improve the efficiency of manpower and skill utilization, as well as contributing toward the efficient standardization of documentation and reduced training requirements. Besides enabling iterative analysis and restructuring of aviation training, the training prediction system will provide data on pilot reliability, related aircraft and training-device design factors.

I will discuss briefly the general approach and anticipated product of each of the five technology development efforts.

(Viewgraph 11)

Human Reliability Prediction Development

Since this is the subject of your gathering here, let me just state what we are looking for in the end. Results of the development will be in the form of handbooks containing system design recommendations, techniques of prediction, methods of evaluation, data banks, and training requirements.

(Viewgraph 12)

Automated Man-Machine Function Allocation

An automated model for determining optimal air crew station requirements for use by design engineers will be developed and demonstrated. The mathematical model and the computer graphic techniques developed in the Boeing/JANAIR cockpit design program (BOEMAN) are used as a baseline of the function allocation model. Automated operational sequence diagram and other operator decision task flow charts, and computerized files of man-machine criteria and human performance data will be developed.

(Viewgraph 13)

Human Factors Test and Evaluation Techniques Development

Both micro and macro analyses and simulations of human operator and maintainer functions will be developed as "tools" to evaluate human factors contributions to system effectiveness. The micro systems include: human reliability and aircraft cockpit models. The macro simulations include: The Naval Aircraft Maintenance Effectiveness Simulation (NAMES), which is currently being completed and modular extensions of this model into the areas of air traffic control operations and shipboard aircraft servicing.

(Viewgraph 14)

Explanation of NAMES

This human factors evaluation technique will be demonstrated by comparative simulation runs of the current fleet techniques and new systems and techniques with common parametric outputs. For example, the effectiveness of job aids are presently being tested in the NA'!ES model.

(Viewgraph 15)

Integrated Job Aids Development

The integrated job aids subsystem is a data preparation guidance and evaluation system made up of tools, guides, and standards for information presentation. An example is the AMSAS project, which is an application of an Air Force-developed, proceduralized maintenance guide concept in the Navy environment. The new maintenance instruction is based on principles of perception, learning, and short-term memory.

(Viewgraph 16)

Here is a sample outcome of the recent AMSAS field shipboard tests. It shows how a human-engineered job aid could improve on the current capability.

(Viewgraph 16A)

Here is why we need such an effort. This viewgraph shows a typical fleet Work Center personnel assignment and qualification.

(Viewgraph 17)

Aviation Training Success Prediction System

A computerized data management system is being developed which will far exceed the data processing functions of the current Pensacola Student Prediction system or Basic Aviation Training Information System (BATIS). Besides training management functions and syllabus evaluation, aircraft and training device design factors will be extracted for engineers' use.

So, this is the overview. Again our endeavor is to help Human Factors effort pick up enough speed to match other technological progress, and above all, give engineers something they can bite into and like it too.

W43-13: HUMAN FACTORS ENGINEERING TECHNOLOGY

ADO THEMES

- BEYOND LIFE SUPPORT AND PERSONNEL MANAGEMENT
- POST HOC ARGUMENT AND LACK OF TRAINING AND TECHNICAL MANUALS
- BEYOND MIL-H-46855, AND TIME AND MOTION, WORK STUDY, ETC.

W43-13: HUMAN FACTORS ENGINEERING TECHNOLOGY

ADO GOALS

- HUMAN RELIABILITY PREDICTION
- AUTOMATED MAN/MACHINE FUNCTION ALLOCATION
- HUMAN FACTORS TEST AND EVALUATION
- INTEGRATED JOB AIDS
- AVIATION TRAINING SUCCESS PREDICTION

DISCUSSION OF MOMIYAMA's PRESENTATION

Momiyama - We always get criticized for dreaming, and the point I make: "Yes, we are dreaming, but at the same time we're working on how to get there from here, and we will pick up or give out a lot of things along the way."

Meister - To what extent is criticism like that going to modify the statement of the goals? To what extent will criticism that comes from the table, from the floor, to what extent will such critiques, criticisms or comments affect or modify the ADO objectives?

Momiyama - I don't think the ADO itself is sacred, I mean that the objective is sacred, but the technical approach to it is our approach and that's the reason why we are holding this kind of a symposium or workshop and I would like to see more of this as we go along.

Meister - Well, then, I don't think it's really a criticism. The point that is to be made is simply, I think, that "You've got two stages." First, you have to determine what you have to stick it into a computer. There is a tendency to jump over the immediate problem, "How do you really do this?" Methodologically this is poor, but the role of computerization is certainly something that nobody can object to.

Swain - Only if the computer becomes the tail that wags the dog again, do I object.

Mills - We have a specific problem ourselves and it is more immediate because we have a computer right on our floor. This thing has got to be utilized in a hurry. As in developing our data system, I have continually said that it has to occur in steps and I didn't even want to mention the word computer in particularly the first step, and probably not in the next two. However, this is just impossible because if we don't show some sort of utilization of this machine we could lose it at the time we really are ready for it. So the only thing you can do is try your darndest that you don't let this kind of thing shape the kind of project you're working on.

Siegel - I don't think you can answer the question until you know what they have in mind. In that program for example, you talk about functional allocation as the primary step, you talk about functional flow analysis. I could certainly see that with the graphic technique they're talking about how one would develop all sorts of functional flow analyses against time lines through a computer and come out with awesome conceptions of functional flows, 1, 2, 3, go home early. Such a development process could be worthwhile. On the other hand I certainly do agree with you that you're sure not going to automate something until you know how to do it yourself. Maybe you fellows know more than I, but I sure don't know how one functionally allocates.

Swain - I know how you do it, that is I know how we do it at Sandia.

Siegel - I know we do it, I just don't know how to do it.

Meister - If you guys really know how to do it, then tell the rest of us. However, I would like to point out one thing. There is a very close relationship as I see it between functional allocation and human reliability prediction because if you know quantitatively how well people can perform, then when you use these figures to compare a number of alternative design configurations, and I'm using the term design configurations in a very broad sense, then your allocation automatically falls out. But I think it's a mistake to think of function allocations as being something distinctly different from the human performance reliability aspects, as a technique which can proceed independently of the business of providing values to stick into certain functions.

Tolcott - I'll second that and emphasize that you can't begin tc allocate your functions, whether automatically or manually, unless the reliability data has been developed as a prerequisite. Not only are the two problems related, but you have to take the reliability problem first and solve that one before you can move onto the next one.

Coburn - Quite right. As you recall in the flow chart that Tom had up there, the reliability measures went into the function allocation criteria area and preceded other measures.

Meister - When you pull function allocation out as a distinct entity, a distinct function which must be formed, there is a tendency to segregate it from the rest of the items which impact on it. So, I'm sure, for example, if you would let a contract to any of us for developing function allocation models, we would merrily go on our way developing function allocation models, with probably minimal consideration of the human reliability aspects, simply because the requirement would be structured in this way; and obviously there wouldn't be enough money to include these other primary aspects as well. That's just life, I suppose.

Blanchard - Well, function allocation is an inherent part of design and I think part of our problem with flow diagrams is that the tool is not adequate to the problem. Function design is a complex undertaking in which we must look at all the human capabilities we have to work with and consider the interactions and constraints which exist and available trade-offs. I'm not sure that we have enough information about human behavior in complex systems to be able to use a computer effectively for function design at the present time.

Meister - I don't even like the term function allocation, although it's hallowed by tradition. If you work with what design engineers really do, as system development really proceeds, functions are not really allocated in the sense of, "you, man, you look" or, "you perceive," "you decide," or, machine, "you perceive" or "you decide." What design engineers do typically is to take the entire system requirement, develop a design configuration which includes both a human and a machine interaction, several of these probably, and it is only after they do that, do they want to decide which of these gross configurations they want to go with. The comparison of these configurations is the real function allocation. It is only at this point that you really get into the business of allocation.

Siegel - You see, Dave, this is really our fault. In the tables, which we put out in terms of what man can do better than machines and what machines can do better, there are only broad generalities. Man is better on short term memory, machine is better on long term memory, man can hear better over wider bandwidth all the broadest, non-usable data. None of it is ever geared to a specific weapon system or even an airplane or a sonar system or a command and control center. The engineer or the human factors man, too, looking at this says this is nice motherhood but certainly not applicable to what I'm doing here, and he is immediately lost in his own morass of background data. I think that it's possible to develop these guidelines, but I think they'll have to be specific to systems; otherwise, they're just going to get lost.

Meister - I don't think they can be specific to a particular system like the XYZ system, but they can be to specific classes of tasks, for example, yes. I don't know anybody that uses the traditional Fitts lists which were fine as a starting point when Fitts developed them but have no practical significance.

Swain - I do, I teach human engineering, it's great for teaching at the engineer level.

Meister - Well let me say this, Alan. I think it is most unfortunate that human factors people habitually repeat the same nonsense, not only to engineers and I don't mind telling engineers nonsense if you have to get across a point, but human factors people also tell it to themselves, as if the Fitts list meant things, as if this function allocation existed as a distinct stage in the development of systems when, in fact, it doesn't work that way. All you have to do is to look at how systems are designed to see how ridiculous this whole thing is. That's why I pin-point human reliability prediction as being the critical point, because the kinds of questions that an engineer will ask you are not, should I allocate this function to a man of this capability? He'll ask you, "If I stick a man in to look at this scope in this particular task, with this particular equipment configuration, Number 1, Can he do it? and Number 2, just how efficient is he going to be doing it?"

Swain - You don't understand what I'm saying. I'd like to clear this up. Like everyone else I've expanded these two lists; that is, what man does better than machines and vice versa. We use this information in our course at Sandia Laboratories to get across to the new engineer who's just coming from graduate school that there are things which people do very well and that they shouldn't automatically assume that they should automate everything. At least a lot of engineers I know, when they're fresh out of school, this is what they tend to assume so we use it simply as a guideline to teach them the limitations of design.

Meister - O.K. as a tutorial aid, perhaps, but as I said before, I think that human factors people are the victim of their own mythologizing tendencies, which is unfortunate.

Momiyama - Also there is our management. We have to use "words" that they understand.

Tolcott - You don't want to promise to achieve something that you can't achieve without laying out a series of prerequisite steps to attain that goal. Otherwise, you're in worse trouble.

Jenkins - I think that many of the comments that are probably very well known to the people who have to work from day to day with the problem, and it is certainly not the attempt of the ADO to foster erroneous types of data or concepts.

CRITERIA FOR DEVELOPMENT OF A HUMAN RELIABILITY METHODOLOGY

D. Meister
Human Factors Department
Bunker-Ramo Corporation
Westlake Village, California

CRITERIA FOR DEVELOPMENT OF A HUMAN RELIABILITY METHODOLOGY

INTRODUCTION

The author of this paper was given the assignment to "consider the three or four most important human reliability research problems the Navy has and the specific steps required to solve them".

It is undesirable to begin an assignment by quibbling about the terms in which it is phrased. There are, however, more than three or four problems critical to human reliability (HR) research. Moreover, almost certainly these problems are not peculiar to the Navy but are to be found in any man-machine system, in any military or civilian organization. Hence their solution would benefit not only the Navy, but all the military services; and most particularly the Human factors discipline itself. Some of these problems have been discussed previously in an excellent paper by Altman (1968), and in reviews by Freitag (1966) and Swain (1969).

The extent of these problems is such that it will take more than a single paper to specify the steps required to lead to their solution.

To approach the problem systematically, the discussion should center about answers to the following questions:

1. How do we define HR?
2. What do we want an effective HR technique to do for its users?
3. What system development and system operation questions should the technique answer?

4. What are the requirements (criteria) that an effective HR technique must satisfy?

5. What are the elements of such a technique?

6. What are the problems of developing an effective HR technique?

7. What are some of the ways of approaching the solution of these problems?

It may appear as if the answers to the above questions (or at least all but the last) are obvious to those who have been working in the area. Even so, it will be useful to review the answers to these questions.

A. How Do We Define HR?

I define HR as the application of performance data to the prediction of operator and technician performance in the context of the factors influencing that performance; the purpose of the prediction is the solution of system development and operational use problems. (Incidentally, when I use the term "operator" in the following discussion, I explicitly include maintenance technicians, although I will discuss the prediction of maintenance performance separately.)

The point of defining HR in this way is to emphasize that it is a tool for use by system development specialists and by personnel. It may also be a research instrument, and certainly research is required to develop HR; but the goal of that development must be to make it satisfy the needs of users, needs which will be discussed later. Any HR technique to be meaningful must be more than a research methodology and must be capable of being employed by others besides the researcher himself. It cannot be said that present HR techniques, at least as far as the author is aware of them, satisfy the definition.

B. What Do We Want An Effective HR Technique To Do For Users?

To solve the various problems encountered by system developers and system personnel, the HR technique should possess *design*, *prediction* and *measurement* capabilities.

In its *design* capability it should be able to

1. Aid in collection of functional responsibilities, thereby suggesting the manner in which a man-machine configuration should be designed.

2. Aid in the selection of the most effective configuration.

I assume that one of the bases for assigning responsibility to the human for implementing system functions is the known capability of the operator or maintenance man to perform that function. Very simply, this is function allocation. The HR technique must supply a quantitative value for the anticipated performance of the human operating within a man-machine configuration; the system developer must be able to say of that man-machine configuration: the operator can accomplish his task at a specified level of proficiency; and that level of proficiency will or will not satisfy system requirements. If the expected performance of the operator cannot satisfy system requirements, the human cannot obviously be assigned the function within the man-machine configuration as conceptualized and some other configuration must be sought. For example, in a hypothetical command/control system N messages per unit time must be received and transmitted; the question the system developer wants to know is, will the human operator be able to receive/transmit that number of messages within allowable error?

Similarly (at a perhaps more detailed level of design), when the system developer has created a number of alternative man-machine configurations, each of which will satisfy system requirements, the HR technique must permit him to select the most effective configuration in terms of anticipated operator performance (obviously there will be other considerations on the basis of which he will make his final choice, e.g., cost, but we are considering here only the human performance parameter). In other words, the HR technique must allow one to design a quantitative human performance value to each configuration and to compare these values.

To make use of the technique for the design purposes specified means that HR must be usable at very early developmental stages when details of the system configuration are at best vague, and performance data on system elements will not be readily available. Presumably a HR Data Bank will be available for application to the elements of the projected new system; the fact that the HR technique will be used for very early function allocations means that Data Bank must be able to deal with relatively gross "top level" functions phrased in terms like stimulus discrimination, monitoring, decision-making as well as relatively molecular task elements like "to read a meter". Not only does it imply predictions at various system levels from gross function to elemental stimulus-response combinations, it also suggests the necessity for combining or at least interrelating system elements at various levels of detail. The use of the technique in system development also presupposes that the predictions must be associated not only with an equipment type (e.g., types of controls or displays or internal components) but also with attributes or dimensions of those equipment components which the designer might wish to select (e.g., scales on a meter or the manner in which internal components are arranged).

Since the system one is constructing includes not only equipment but personnel as well (numbers and types of manpower needed, the procedures they should employ in running the system, the determination of training content and duration, the specification of work-rest cycles, etc.) the HR technique must be able to make predictions involving these parameters as well.

The HR technique must therefore supply system development answers at various levels of interrogation. (The following is in order of increasing complexity.)

Equipment

a. Component attribute (e.g., number and arrangement of controls, scale characteristics, location of test points);

b. Component (e.g., meter, joystick, potentiometer);

c. Equipment assembly (e.g., control panel, power supply, amplifier);

d. Equipment type (e.g., console, tape deck);

e. Subsystem (two or more interrelated equipments);

f. System (the total of all equipment considered to perform a system mission).

Behaviors

a. Task element or single stimulus-response combination (e.g., adjust potentiometer, throw switch);

b. Task (e.g., calibrate voltage, where the task consists of two or more task elements);

c. Procedure (e.g., perform pre-flight checkout, load weapon), where the procedure consists of two or more tasks;

d. Function (e.g., take off and land aircraft, navigate) where the function consists of two or more procedures.

It must be able to predict the performance of all the above with regard to specified conditions such as the operator being trained or untrained, whether the behavior is performed by one man or two, etc.

It should be noted that for the sake of completeness I have included the task element (e.g., the single discrete control activation) as one of the system levels for which answers must be provided. I do so regretfully, because this level adds measurably to the required complexity of the HR system. Actually I do not believe that many of the system development/use questions asked deal with molecular task elements; the preponderence of these questions relate to tasks and procedures, but there are a few occasions when questions are asked about task elements. Hence the need to include these elements in the HR system.

It is my impression that some HR techniques like that of AIR (Payne and Altman, 1962) use the task element as a basic building block for the task. There is some point to this in terms of precisely defining what the task is. The same calibration task may, for example, consist of a different number of task elements, depending on the equipment configuration in which calibration must be accomplished. In configuration A it may consist of three task elements because the equipment involved in the calibration are two switches and a meter; in configuration B it may involve only a single switch and an indicator. Unless the calibration task is defined in terms of number and type of task elements, one could get widely different performance values for what is supposedly the same task.

With regard to its predictive capability, the HR technique must be able to

1. Predict the operational performance of one or more personnel performing a variety of behavioral functions in relation to specified equipment configurations, at various levels of system complexity.

2. Indicate the contribution to or relationship of operator and maintenance technician performance to the overall system output.

What the above means is that at any time in system development the HR user must be able to determine that when the system is operationally activated its personnel will perform at such and such levels of performance. This permits the developer to determine whether that performance will be acceptable or not (compared with system requirements) and, by reference to the system configuration with which that performance is being accomplished, which factors might be responsible for any anticipated inadequate performance. If, for example, after making a prediction for the total system (equipment, personnel, procedures, etc.) it appears that that predicted performance will not satisfy system requirements, the developer must be able to determine whether the fault lies with personnel performance or some other system element.

To determine the contribution of the predicted operator/technician performance to overall system output, the personnel prediction must be capable of being integrated or compared with performance predictions of the other system elements, the most important one being equipment functioning. The HR predictive metric must therefore be compatible with techniques that predict equipment performance and must in fact be capable of being combined with the latter. Since the technique which predicts equipment performance is the reliability technique, HR must be compatible with equipment reliability methodology.

Ultimately it will be necessary to consider the relationship between HR predictions and system availability predictions, but this represents a greater degree of sophistication than we need presently hope for.

With regard to HR's <u>measurement</u> capability, it must be able to

1. Provide a methodology for measurement of on-going performance of system personnel in the operational situation such that the data gathered in this way can be integrated into already available HR Data Banks.

2. Assimilate new data from a variety of sources.

Until now we have been talking about system development needs which HR must satisfy. With the measurement capability we pass to the needs of operational military commands. The operational user will wish to know:

a. how well are system personnel meeting system requirements;

b. if system personnel are not performing as desired, what are the factors responsible;

c. if a modification in system operation is made, what will be the effect on operator performance;

d. how does the on-going performance of personnel compare with what was predicted.

(a), (b) and (c) are the system developer's questions extended to operational usage; (d) represents the need to validate the prediction.

Obviously any HR technique implies a certain measurement methodology. It cannot merely assume a body of presently available data and deal only with the application of those data. In the first place, no presently available Data Bank is adequate; and will therefore require expansion (which implies certain measurement operations). In the second place, no HR prediction is worth a penny unless it is validated; and the prediction cannot be validated unless the measurement operations required by the validation are compatible with the measurement operations implicit in the HR technique. Altman (1968) has applied the term "homomorphism" to this requirement.

Since present data banks are unsatisfactory, the HR technique must be prepared to accept data from a variety of sources. Among these sources is the general behavioral literature, describing research performed primarily in psychological laboratories. Realistically, until such time as the military services make their facilities available for the gathering of on-going operational performance data, the major untapped source of data to expand the data bank must be the general behavioral literature. We at Bunker-Ramo have a contract (F33615-70-C-1518) with the Human Engineering Division of the Aerospace Medical Research Laboratory under the sponsorship of Bob Mills to try to convert that literature into a form which can describe man-machine tasks. Much of this discussion has been taken from a paper prepared for this study.

One other possibility exists. The author is not a simulation specialist, but it is conceivable that, given a bank of presently available data included in a computer, one could ask the computer to operate upon those data (perhaps using Monte Carlo techniques) to generate more data. However, to do this it would be first necessary to ensure that the initial data included in the computer was of sufficient magnitude and of recognized validity so that one could have confidence in the resultant "simulated" data. Moreover, certain combinatorial rules would first have to be developed.

C. What Questions Should The Technique Answer?

These questions have been adumbrated by the previous section, but we are now in position to specify them in greater detail.

1. What is the operator's capability to perform various functions under various modifying task and environmental conditions?

If the system developer knows this, he has at least a rough screening device for determining whether functions should or should not be allocated to the operator/technician. Note that we specify "under various task and environmental conditions". In the previous section I used the example of the operator's ability to receive/transmit N messages per unit time. The question was raised with reference to function allocation. Obviously the question is difficult to answer if we think only of the general function of message reception/transmission (i.e., listening to, acknowledging, reading, typing, etc. messages). What about message duration? Message format? These are what I call modifying task conditions, i.e., conditions which influence the performance of the basic function. Environmental conditions refer to the physical environment, e.g., lighting, noise, etc.

Although it is possible to supply an answer to the question above without considering these modifying conditions, the answer will be very gross, if not misleading. For any usable precision it will be necessary to include in the HR system a large number of parameters dealing with the specific of the function being performed. This overly complicates the problems of developing an HR methodology, but seems unavoidable.

2. What is the effect of various types of equipment and equipment attributes on the operator's performance of specific functions under specific task/environment conditions?

Suppose it has been decided to employ a large screen display for depicting tactical or strategic information, e.g., aircraft available, missiles launched. How large must the alphanumerics be to secure minimally acceptable resolution? How large should the screen be? Amount of ambient lighting, etc.? To answer these questions the developer must be able to predict operator performance as a function of different resolutions, sizes of symbols, size of screen, different amounts of ambient lighting, etc. Note also that the operator performance predicted as a function of these parameters must be tied in with the specific task being performed, because the parameter values change also as a function of the particular perceptual task being performed, e.g., localization, updating, etc. What this implies is that not only must HR be able to provide performance answers for parameters in general (e.g., different sizes of characters) but must also interrelate these answers with values for different kinds of tasks. The manner in which this interrelationship between tasks and equipment attributes or parameters can be accomplished is one of the major problems the HR developer must cope with.

3. <u>What physical and physiological limitations does the operator impose on equipment design and function/task performance? What environmental factors influence design and performance</u>?

Obviously there are physical (i.e., anthropometric) constraints on equipment design. There are physiological limitations (e.g., tolerance of acceleration, vibration as in sea sickness) on function/task performance. Special environmental conditions excite these physiological limitations. Performance values for these conditions must be applied where they are relevant to the system development question being asked.

4. What is the probability that the operator will accomplish specific tasks under various task/environmental conditions? How is that probability affected by various equipment characteristics?

This two-part question is an extension of question (1) which dealt with operator capability to perform general functions. This question deals with detailed tasks and interrelates task performance with task conditions and equipment characteristics.

It has already been indicated that it is difficult to talk about a behavioral function to be performed without also considering the equipment which is the object of that performance. This suggests that the basic behavioral unit of prediction is the function or task (depending on the specificity of the question asked; which is in turn a function of system development stage) plus the equipment being operated/maintained. For greater predictive precision one should also include in the predictive unit the conditions that modify the performance of the unit, although this is not strictly necessary.

5. What is the effect of different amounts of manpower on task performance?

We have already referred to the fact that the system is usually not composed solely of individual operators working alone with their machines, but rather includes numbers of personnel working in coordination. Any HR technique which cannot formulate its predictions in terms of more than the single operator is in trouble, because its answers will be deficient.

6. <u>How does the performance of one task affect the performance of a second task which occurs either concurrently or sequentially, and how is this performance interrelationship affected by various types of tasks and task conditions</u>?

Since any system operation involves more than the single task, the HR technique must be able to interrelate its predictions for individual tasks, both vertically (in terms of sequence of tasks performed by the single operator) and laterally (tasks performed concurrently by multiple operators). This involves a consideration of the dependency relationships among tasks which is a problem of great severity for presently proposed HR techniques. The reason is that, in contrast to equipment performance dependencies which are of a binary type (the component is either dependent or independent of another component) there may be different degrees of independence-dependence among tasks. We shall address this problem in greater detail later.

7. <u>How does the operator's task performance vary as a function of repeated trials in (a) learning to perform the task and (b) performing a learned task (i.e., fatigue)?</u>

At some stage in system design the developer is faced with the problem of determining how much training should be provided the operator. We do not deal here with training content, which is primarily a function of the nature of the task. However, the determination of training duration is in part influenced by the highest degree of performance of a given task which one can expect as a function of repeated learning trials.

In many system applications the developer is concerned about the effect of fatigue on performance, particularly in terms of determining an optimal work-rest cycle. It is well known, for example, that detection probability degrades as a function of time spent monitoring sonar/radar scopes. To determine an appropriate work period for functions such as these the amount of performance degradation to be expected as a function of repeated trials must be known.

D. What Are The Criteria Of An Effective HR Technique?

It is now possible to specify the criteria a proposed HR technique must satisfy to be considered effective. These criteria are highly pragmatic; that is, they focus primarily on making the technique an acceptable one for system development and operational use. It is unlikely that any presently available technique satisfies these requirements (although presumably some at least will eventually), nor are all these criteria of eoual importance.

1. The technique should be usable by non-specialists, e.g., engineers and operational (i.e., military) personnel. This means that the technique will be relatively simple. For example, for a user to ask a question of HR should not require formulating that question in a special (e.g., mathematical or symbolic) language; the answers he receives from his interrogation of the HR system should not require translation.

2. The technique should not require excessively tedious calculations, as some presently available techniques do. In view of the well-known rapidity of the design decision-making process, tedious calculations will cause the answers received to be delayed until after firm design decisions are made. To

what extent this difficulty can be avoided by use of computerization is not known.

3. The technique should not require the application of performance data which are not readily available. Some techniques may require extremely molecular data, phrased in terms of the specifics of the equipment configuration for which task performance is to be predicted. To derive data of such specificity often means that experimentation must be performed, which not only cannot be performed because of time and cost but defeats the goal of prediction. Nor should the technique require the derivation of applicable data through techniques of expert judgment.

4. The technique must lead to usable design recommendations, whether these recommendations deal with equipment characteristics, training, manpower or procedural suggestions. It is not necessary that HR provide answers phrased directly in terms of such recommendations, but at least the answers supplied must be logically translatable into such recommendations.

5. The technique should be capable of being utilized at all stages of system development, including operational exercise of the system. It should be able to handle all system elements in both molar and molecular form. A technique which supplies only partial answers is unlikely to secure acceptance by system developers and users, in which case it will remain only a research tool.

6. Because fundamentally the user is interested only in determining whether system personnel can do their job, and how that job performance is influenced by various factors (including task conditions), the answers the HR technique provides

must be formulated in task performance terms, at least. What this means is that the metric employed by the technique must be understandable in terms of concrete system operations. Artificially derived coefficients which do not readily relate to individual tasks or procedures will be relatively useless.

7. The technique must be capable of being validated by the collection of performance data in the operational setting. A non-validatable technique is a scientific anomaly. The means that the technique must contain a measurement logic which is visible and can be translated into real-world data-gathering operations.

8. The predictive outputs of the technique must be compatible with (capable of being combined with) those of equipment performance predictive techniques (i.e., reliability). Since most HR predictions are made for a system which receives a reliability predictive index, it is not only non-parsimonious to employ two predictive techniques which cannot be combined, but it is likely that system developers and users will be suspicious of HR unless it can be combined with reliability. From this we derive another requirement: that since the reliability prediction is formulated in probabilistic terms, the HR technique must be formulated in comparable terms. Another reason for the use of a probabilistic metric is that the HR index is unlikely to be able to predict the performance of the single task event; hence it will be necessary to deal with the likelihood of events occurring over a series of performances.

9. The HR technique must be capable of assimilating data from various sources. It is unreasonable to think of a technique which, having once developed a data bank, assumes that that data

bank will remain static. Moreover, the technique should be able to operate upon the new data to apply it to the various system elements for which HR predicts. In other words, if new data are received which describe task performance under specifield equipment, task and environmental conditions, it should be possible to partition the performance effects of each condition and to treat them as individual data elements; it should be possible to take task data and categorize it as being a subset of functional data.

For example, assume that the following performance datum is received: operator reads meter at distance of 18 inches under 30 footcandles illumination. It should be possible to partition the datum into values corresponding to the following elements: reading task; meter; viewing distance; illumination. The reading function should be capable of being subsumed under, say, stimulus recognition function.

Admittedly the above criteria are quite stringent and it is not assumed that HR in its initial development will have to satisfy all the requirements. Ultimately, however, it must; and any technique which ab initio lacks the capability of being developed to those requirements should be automatically disqualified.

E. <u>What Are The Elements Of The HR Technique?</u>

Considered as a predictive <u>system</u> HR has certain elements. It is useful to examine what these are, because these may suggest certain problems that need solution:

1 Assumptions and goals underlying the HR structure.

2. Definitions and taxonomic categorizations of HR elements, e.g., definition of functions/tasks; specification of types of perceptual functions.

3. Specification of the HR data required to answer HR questions, e.g., data on the effect of environmental conditions, performance data relative to equipment types and characteristics;

4. A metric or way of expressing HR outputs;

5. Rules or operations for

   a. asking questions of HR;
   b. retrieving data from HR;
   c. outputting HR predictions;
   d. combining performance predictions;
   e. extrapolating, interpolating or generalizing new data from already available data;
   f. incorporating new performance data.

6. Categorization of behavioral and equipment parameters which are assumed to modify performance, e.g., presence or absence of feedback, accuracy requirements, organization of internal components.

Each of these with the possible exception of assumptions and goals presents some problem which requires solution if the HR technique is to be implemented. These will be discussed below. Before doing so, however, it is worth spending a short time on the assumptions which underly our concept of an HR methodology. This is because these assumptions are in part responsible for the problems we encounter.

It will be noted that these assumptions are not peculiar to the HR methodology but stem from and apply equally well to our concept of behavior in general. Not all possible assumptions are included in the following list, because to do so would require writing a book on the theoretical foundations of psychology.

1. One most important assumption is that a variety of parameters influence behavior. We all accept this assumption, and volumes of empirical data justify our faith. From the standpoint of an HR methodology this assumption means that our methodology must include at least the major parameters. The major parameter is equipment, of course, but there are others, such as accuracy and time requirements imposed on behavior (what I have in previous papers called the "pacing" factor), the presence or absence of feedback, etc. To the extent that behavior is modified by the occurrence of a parametric variable in the context of the operator's performance, it is necessary to include the parameter in the performance prediction. However, the problem arises of which parameters are most important (because one can hardly include them all) and how to determine when a significant parameter is affecting behavior. The assumption has consequences, obviously, for the development of the HR data bank, since the bank, to be valid, must include data on these parameters.

2. We assume also that molar units of behavior are composed of smaller elements; that consequently these molar behavioral units can be partitioned by analysis into molecular elements or can be built up by adding molecular elements; that molar behavior subsumes molecular. Typically we assume that functions are composed of tasks, tasks of task elements, etc.

The kinds of questions asked about behavior imply that the HR methodology must supply meaningful answers about each behavioral level. The implication of this assumption for HR is that the methodology must incorporate rules for combining or analyzing these elements.

3. Behaviors interact. A single task may consist of several task elements, as, for example, when the operator must read an indicator and concurrently press a switch. Each task element has an independent performance value; how are these to be combined? Moreover, the performance output of a sequence of behaviors cannot be fully understood without consideration of their behavioral interaction. To put it in HR terms, if our basic behavioral unit consists of the task, then at least some tasks are interdependent, and we must account for their interdependence in our predictions.

4. Equipment characteristics also interact with each other and with behavior. An equipment is described by at least several equipment characteristics, all of which interact. For example, if we take a simple control panel we see that it consists of at least a number of controls and displays arranged in some pattern across the face of the panel. The two characteristics - number and arrangement - interact so that we must include both in our HR predictions. Singly each characteristic has an effect; in interaction the effect on performance of one is presumably modified by the effect of performance of the other. For an accurate prediction of performance in relation to the control panel we must supply the interactive effect of the two characteristics. This increases the complexity of the predictive job.

In summary, the types of interactions we are dealing with therefore are:

a. multiple equipment characteristics within the same equipment;

b. multiple task elements within the same task;

c. the interaction of equipment and behavioral functions in a single task;

d. the interaction of multiple concurrent or sequential tasks;

e. the interaction of modifying task conditions and single and multiple tasks.

The HR system must account for all of these.

The problem has significant implications for the structure of our Data Bank. At least two different data bank structures are possible:

Data Bank Structure 1	Value	Data Bank Structure 2	Value
Behavioral function$_{1-n}$	A_{1-n}	Task$_1$ + equipment$_1$ + modifying factor$_1$	A
Equipment$_{1-n}$	B_{1-n}	Task$_1$ + equipment$_2$ + modifying factor$_2$	B
Characteristics$_{1-n}$	C_{1-n}		
Type$_{1-n}$	D_{1-n}	Task$_2$ + equipment$_2$ + modifying factor$_2$	C
		. . . .	.
		. . . .	.
		. . . .	.
Modifying task conditions$_{1-n}$	T_{1-n}	Task$_n$ + equipment$_n$ + modifying factor$_n$	N

The first structure presents predictive values for each molecular and molar element of behavior and equipment. To derive a performance value for any single task it is necessary to organize the elements in various ways corresponding to the type of behavioral function, equipment, etc. for which a prediction is to be made. For example, data bank structure 1 might have the following extract:

Behavioral function -	reading,	value	.9876
Equipment -	meter	value	.8754
Equipment characteristic	scale type	value	.9753
Modifying task condition	reading under specified illumination	value	.8777

(all values are purely hypothetical)

To determine the probability of performing the task, reading meter with given scale type under specified illumination, one would have to combine each of the above values in certain ways.

In the second structure it is unnecessary for the user to organize elements, but the data bank must have an extremely large number of task-equipment combinations from which one selects the one combination corresponding to the behavioral unit one is attempting to predict. For example, data bank structure 2 might have the following:

Reading meter with scale type X under 10 footcandles ----	.8543
Reading meter with scale type Y under 10 footcandles ----	.8655

Reading meter with scale type Z under 10 footcandles ---- .8666
Reading meter with scale type X under 30 footcandles ---- .8865
etc.

The permutations are very large. Since it is unlikely that the second type of data bank will have all possible combinations listed, for certain behavioral units no prediction can be made. Both types of data banks therefore pose problems.

Examining the assumptions underlying the HR technique should not be considered an academic exercise. Since the HR methodology we develop is a product of our underlying assumptions, it is necessary to make those assumptions explicit in order to determine whether a given methodology is scientifically sound.

F. Problems of Developing An Effective HR Technique.

Before discussing each problem in detail, let us list them to see what they consist of. Few of these problems will come as a complete surprise to the reader; they have been anticipated throughout the previous sections of this paper.

1. Definition of the behavioral units whose performance is to be predicted.

2. Definition of equipment units to be included in the HR methodology.

3. Definition of the number and type of parameters to be included in the HR system.

4. Determination of a suitable HR metric, and its empirical distribution.

5. Determination of what is to be included in the data bank, how it is to be organized; how it is to be developed, and how it is to be used.

6. Specification of the rules for combining predictive values for: combinations of equipment characteristics; sequences of behaviors for the same individual; tasks performed by two or more indiviuals working concurrently or consecutively on the same job.

7. Specification of the rules for combining HR predictive values with equipment performance predictive values.

8. Determination of methods for validating HR predictions.

9. Determination of methods for incorporating new data from a variety of data sources into the data bank.

10. The solution of all the preceding problems with special reference to the prediction of maintenance technician performance.

In the following discussion we propose to examine each problem and supply a method of attack on the problem without, however, providing a complete solution, since as noted previously the effort involved in solving these problems requires continuing research.

1. Definition of the behavioral unit.

Reduced to its simplest terms, this is the question of how molecular the behavior included in the unit should be. I have defined the behavioral unit as including a behavioral function, the equipment implementing that function and any significant modifying factors. Depending on the questions which the HR technique is supposed to answer, I might include various level of detail in that unit. In other words, my unit might consist of: activates typewriter key; or types message (of designated format and length); or transmits messages using typewriter. The level of detail included in the behavioral unit would depend on the specific questions asked by the user, since he could ask questions about every level.

However, if one includes various levels of detail in the behavioral unit, the problem of getting from one level to another immediately arises. One must have data for each level, or else provide a method of combining more molecular levels to derive data for more molar ones. Other researchers may well prefer to have only one or two levels. Pragmatically it makes the whole problem simpler if one restricts the number of levels.

The essential thing to remember (and this applies to most if not all of the problems discussed) is that the problem can be solved by definition. Whether one uses one or more levels or which one(s) are decided upon is purely a heuristic consideration. Once the HR developer has defined the limits within which his technique will work, he has solved his problem.

2. Definition of the equipment unit.

The same problem exactly applies to the specification of the equipment unit. I may wish to include in my predictive technique: the equipment characteristic (e.g., arrangement of controls or type of scale marking); the equipment component (e.g., joystick, meter, knob); type of equipment (e.g., sonar/ radar scope, aircraft throttle), etc. If I include the various levels I must indicate the rules under which I will include them in my predictions. The kinds of questions asked by system developers forces us, I suspect, to include all levels in the HR methodology; depending on the type of question I will select that equipment category which will answer the question. Thus, if the developer wishes to know how one arrangement of displays will affect operator performance relative to another arrangement, I will supply him with performance data about these equipment characteristic. If he asks the question whether he should use a legend light rather than a meter for a designated function, I will work at the component level.

3. Definition of number and type of parameters.

If my HR system is to provide meaningful answers, it must consider the parameters which influence operation performance. To the extent that I do not include significant parameters in my HR technique, my prediction will be lacking. Which parameters? Logic and already available experimental data will suggest the parameters to be applied to modify the prediction of the behavioral unit, but the choice is the HR developer's. The ones that would appear to be most important to this author are: the number of components from which the control to be activated or the display to be read must be selected; the organization of these; the presence or absence of feedback; the sequence of responses to be made; response accuracy and speed requirements; the type and number of stimuli presented. Not all of these are effective in any one task, so that rules for determining when they are effective must be developed and applied. Moreover, it is quite possible for more than one parameter to affect the individual behavioral unit, so that I am faced with the problem of integrating the predictive value associated with each parameter.

For every parameter I select for inclusiòn in my HR system, I must secure appropriate performance data from some source. This would tend to limit the number of parameters selected. At the same time, however, the variety of questions asked of the HR technique force me to be more rather than less comprehensive in my selection of parameters.

4. Determination of a suitable HR metric.

There are a variety of ways in which the HR prediction can be expressed. Generally, following equipment reliability practice, HR workers have utilized a probabilistic error or task measure in which the four figure value (e.g., .9999) represents the probability that a given behavioral unit will be satisfactorily accomplished. The author follows the same practice on the grounds that what the system developer and user want to know, reduced to its most simple terms, is, will the job or task be accomplished as required? Even when the question asked relates to an equipment characteristic such as the effect of one arrangement vs. another, the developer still wishes to know whether that particular characteristic will affect the operator's probability of task accomplishment.

It occurs to the author that task elements do not readily fit into the task accomplishment framework, since the task element is not goal-oriented. One may ask what the task is that the task element is seeking to accomplish? However, one might assume that accomplishing the individual task element serves as a goal or task in itself. Thus the probability value associated with striking an individual typewriter key is the probability of the correct key being struck with sufficient force to imprint on paper.

Task accomplishment probability can be expressed in a number of ways, e.g., that the task will be accomplished within a required time period, to a required level of accuracy, etc. The nature of task requirements will determine the specific meaning of the task accomplishment measure, but the probabilistic formulation can handle them all. Because of the variety of interpretations the probability value may have, the data bank must indicate what the correct interpretation of the probability value is: e.g., probability of error occurrence, probability of task completion, probability of task completion in a specified time period, etc. This has the disadvantage that multiple probability values may be required for each data category in the bank. However, it is unlikely that the data bank will supply all the requisite values, at least initially. The metric utilized should of course provide a range of values or confidence limits to indicate also the variance in the values supplied.

It is of course possible to utilize a completely artificial scalar measure such as a scale from 0 to 100, in which the values represent degree of accomplishment or "goodness of response"; but the difficulty here, besides the problem of defining the nature of the scale, is to relate the scalar measure to some concrete task performance. If a task receives a value of 50 on the scale, what does this mean for the likelihood that the task will be completed adequately? Another difficulty is that a non-probabilistic measure cannot be readily compared with or combined with probabilistic measures of equipment performance.

The author does not consider the selection of an appropriate metric to pose a great problem for the HR developer.

Assuming other problems can be solved, e.g., definitions, the availability of appropriate data, the probabilistic metric would seem to work well enough to supply desired answers.

5. The Data Bank, organization, development and use.

The data bank or store of applicable performance data is constrained by two factors: the kinds of questions which it must answer; and the availability of performance data to answer these questions. It is apparent from what has been said previously that system developers and users will ask a wide variety of questions; in consequence the data bank must be very comprehensive if it is to satisfy these demands. In particular it must include as minimal elements the following: data on

a. equipment characteristics

b. equipment component

c. equipment types

d. task elements

e. tasks

If appropriate transformations are available, procedures and functions need not be included; if transformations are not available, data on these must be included.

f. modifying parameters, including

(1) task conditions

(2) environmental conditions

(3) physical and physiological limitations on performance

(4) effect of repeated trials.

There may not of course be data to fill each of the above categories. Presently available data banks will not supply sufficient data. Unless provisions are made to collect appropriate data either in the laboratory or in operational system exercises, consideration should be given to the problem of transforming general behavioral (i.e., psychological) data into man-machine equivalents to fill the gaps.

6. Combinatorial Rules

Outside of the development of the data bank one of the most difficult problems the HR developer faces is the specification of rules for combining performance values for equipment characteristics, equipment components, task elements and tasks to provide a single figure of merit which represents the probability of personnel performance. As indicated earlier, the interdependency of system elements is so complex that the rules developed and used successfully for equipment reliability predictions do not appear to be adequate for HR - although they are acceptable as a first step. It has been suggested that if one plots out graphically (for example) all the contingent behaviors possible in a given operation, and assigns a probability to each behavior, then combination reduces to a judgment of whether the behavior is independent or dependent, in which case one applies the appropriate probability equations. The difficulty the author sees in this is that there may be degrees of dependence or independence between equipment characteristics or tasks.

If two tasks are completely independent, they may be dealt with multiplicatively. If they are completely dependent, such that failure to accomplish task 1 prevents task 2 from even being initiated, this too can be readily handled. But suppose that in addition to the probability of accomplishment for each task, if it were performed in isolation, there is a dependent probability such that if task 1 is not accomplished or accomplished with some inaccuracy, the probability of accomplishing task 2 with required accuracy is reduced by a specified amount. The probability equation for such a situation becomes very complex, one suspects. In other words, the combinatorial problem becomes fierce when we attempt to predict not just whether or not a task will be accomplished, but whether it will be accomplished in terms of some degree of degradation in required accuracy.

Since the behavioral literature offers few or no suggestions, the only solution is to define one's rules (on the basis of logic or any other evidence) and check against empirical performance to see if the prediction is reasonably close. The selection of appropriate rules can be helped by the computer. A number of alternative rule possibilities can be tried out on the computer and the most productive selected for checking against operational performance.

7. Rules for combining HR predictions with equipment reliability predictions.

The desirability - indeed the necessity - for combining or at least contrasting HR predictions with equipment reliability (ER) predictions, has been pointed out previously. This requirement imposes certain constraints on HR. Obviously, the metric utilized by ER - probability - must also be employed by HR. Of equal importance, the two measures must be coordinate. In other words, to combine an ER prediction for an equipment with the HR prediction for the same equipment, the HR prediction must include all of the tasks involved in operating (or maintaining) that equipment.

Given these premises, there appears to be no special difficulty in combining HR with ER predictions. If one assumes that the behavioral and equipment aspects they predict are co-equal and equally necessary for system operation, since the system cannot function if either its personnel or its equipment fail to accomplish their tasks, then the combination of values can be achieved by simple multiplication of the individual HR/ER values.

8. Methods of validating HR predictions.

The requirement for validation imposes certain constraints on the methodology. The structure of the HR methodology must be such that measurements can be taken in the operational environment which will produce data capable of being compared with the HR predictions and later included in the HR data bank. For example, the definition of the behavioral unit must be such that the behavioral unit can be reliably observed and measured. The data secured by measurement in the operational setting must be capable of being transformed into the HR metric; and the individual unit values must be capable of being meaningfully combined according to the rules specified by HR.

From this it would appear that one way (among others, of course) of testing the effectiveness (although not necessarily the validity) of the HR technique is to require its developers to specify the rules to be followed to validate that technique.

True validations (i.e., involving collection of comparison data with operational systems) have been few and far between. The mere use of an HR technique to make a prediction cannot be considered a validation, even though the results achieved may seem reasonable. Therefore, none of the HR techniques presently available can be considered validated. This is because the validation must be performed in an operational setting or one which reasonably reproduces the major characteristics of that setting. As a first step in that direction, the collection of performance data in high-fidelity simulators is acceptable, but must be followed by collection of the same data in the operational environment.

One cannot insist on rigid standards of correspondence between HR prediction and operational task performance as indicating validity or invalidity. A statistically significant difference between a prediction and actual performance does not necessarily mean that a technique is invalid, since it should be expected that any technique will in its inception be rather crude.

9. Methods of incorporating new data into the data bank.

Because of the criticality of the data bank to any HR technique it is necessary to ensure a continuing flow of data from external sources into the data bank. Three sources of such data exist: the operational setting, the general behavioral literature, and expert judgment. The last appears dubious at best, although certainly it presents fewer problems than the others. Attention is presently being paid to methods of incorporating general behavioral (largely laboratory) data into a man-machine data bank in a study being performed by the author and his colleagues for the Human Engineering Division of AMRL. The possibility of using such data does not, however, relieve the military of the responsibility for throwing open its resources for the collection of appropriate performance data. Indeed, once an acceptable HR technique is tentatively organized, the next immediate step should be to start collecting operational data.

The collection of data in an operational military setting presents many problems of control to the investigator. Consideration should be given to the development of self-reporting data collection techniques on a mass basis or a more conservative process involving observations of samples of various types of tasks to secure a representative subset of the total task population.

10. The prediction of maintenance technician performance.

We have left maintenance performance (i.e., trouble shooting) for the last in this discussion because it represents a quantum jump in difficulty over the development of an HR technique which predicts operator performance. Theoretically an effective HR technique should be able to handle both operator and maintenance behaviors, and ultimately it will do so; but for the present the problems mentioned previously are magnified a thousand-fold as they apply to maintenance. For one thing, the definition of the behavioral unit of prediction is much more difficult to specify, if only because many of the functions performed by technicians are highly cognitive and hence covert. What is difficult to observe is quite difficult to categorize meaningfully. The factors that modify maintenance performance are largely unknown. My colleagues and I have done some work to determine these factors and have come up with accessibility, diagnostic information, equipment structure and operator capability as the major factors determining that performance. However, performance data bearing on these parameters is practically non-existent.

For purely heuristic reasons it is possible that the effort to develop an HR predictive technique which is applicable to maintenance behaviors should be performed independently of other HR technique development efforts. This does not mean that the two efforts should be uncoordinated or that there should be no cross fertilization of ideas, but unless a significant amount of time is given to the maintenance aspects of the technique, the results will be inadequate. Ultimately of course the maintenance aspects of HR must dovetail with the operator aspects, but the problems involved in maintenance alone deserve more than a passing effort.

The general logic described in the preceding sections of this paper would seem to apply to maintenance behaviors just as it does to operator behaviors. In other words, an HR predictive system which is to be applied to maintenance actions must answer the same types of system development questions, meet the same criteria, start with definitions, etc. There appears to be nothing inherent in maintenance behaviors per se which would justify a markedly different approach.

From this consideration of the problems to be solved in developing an effective HR technique, it is apparent that these are of two types. The first is what we term "definitional" or analytic, which can be solved by establishing definitions and rules and does not require an empirical effort, except at a later time. The second type of problem requires experimental or empirical data collection efforts.

By far the largest number of problems examined are of the definitional or analytic type. With the exception of the development of an appropriate data bank - and then only because present data banks are so lacking - all the problems can be solved analytically. Since the HR technique is an artificial construction, it can be created as the developer wishes, subject only to the proviso that the technique can be validated.

Suggestions for an Approach to Problem Solution

It would be hard to say that one of the problems cited was any more important than any other. This is because all of them must be solved in the course of developing an effective HR technique.

From the Navy's standpoint, in view of the fact that several techniques are presently available, it would seem reasonable first to determine whether any of these satisfy the criteria specified in this paper and offer substantial promise of developing into a usable technique. There is no sense in re-inventing the wheel, if already available wheels will permit the cart to move. While no presently available technique is completely adequate according to the criteria specified in this paper, it seems unlikely that they must be discarded in toto. Therefore an attempt should be made to constrast the techniques and select the most effective elements of each to arrive at a starting point for further research. A preliminary attempt in this direction was made by Freitag (1966) for NELC, but not at any detailed level.

Without ignoring any of the problems discussed earlier, it is felt that major research efforts should be concentrated on the following areas:

1. Assuming that the superstructure of the HR system is decided upon, e.g., specification of system elements, their

level of detail, etc., major attention should be paid to the development of combinatorial rules.

2. The development of a data bank with particular attention to the establishment of methods of securing appropriate performance data from Navy operational systems. Sooner or later, even if efforts to transform general behavioral data into man-machine equivalents are successful, it will be necessary to supplement these data with operational performance data. The development of an operational data gathering system would also solve the problem of HR validation.

3. The development of a technique for predicting maintenance technician performance. Here it appears necessary to go back to first principles to build the structure referred to in (1) above. So little has been done in this area that there would be no great loss if the effort to develop an HR technique for maintenance behaviors were started afresh.

The development of an effective HR technique should be viewed as a multi-stage effort. First effort should be to develop a relatively simple system, devoid perhaps of refinements, which might be able to predict gross behaviors only. This, once validated, should be used as a foundation to build a more sophisticated system.

In all that has been said we have mentioned computerization only in passing. Obviously, the ultimate HR technique will require computerization, if only to handle the masses of data needed and to operate the combinatorial rules adopted. It would be a mistake, however, to begin the HR development effort by orienting it specifically to computerization. The HR structure needs to be nailed down; then the development of a prototype data bank (building on what we have already) can begin; after the prototype data bank is available, one can look realistically at the requirements for its computerization.

To those who have been working in this area for some time, it may appear as if what I have been talking about is curiously old fashioned, the problems raised having been around for some time. Yet because these problems have received no solution, it is all the more pressing to consider them once again.

REFERENCES

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

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

3. Swain, A.D., Quantification of Human Performance. Proceedings, 8th Reliability and Maintainability Conference, Denver, Colorado, July 7-9, 1969, 251-254.

4. Fayne, D. and Altman, J.W., An Index of Electronic Equipment Operability: Report of Development. AIR-C43-1/62-FR. American Institute for Research, Pittsburgh, Pennsylvania, 1962.

DISCUSSION OF MEISTER's PRESENTATION

Harris - I think we have to distinguish between the human reliability methods at the outstart. I'm not sure that the method itself need be or should be understandable by anybody. I think the concern of the engineer as the user is the data and some instruction as to how they may apply them in the design of the system. He doesn't know what he has to pass to apply them or even to applying the method itself.

Meister - Perhaps what I really meant to say was the methods developed should be capable of being utilized by the immediately involved human factors man, a somewhat different thing perhaps. In other words, it should not be the private preserve of a specialist in a special technique.

Regulinski - Why only a human factors person, why are you confining it to them?

Meister - Why am I confining what to the human factors people?

Regulinski - Why are you excluding engineers?

Swain - I'd like to answer that if I could, Dave. I work with reliability specialists in systems and equipment. I don't understand some of the techniques they use in going through circuit analysis. I don't expect them to understand the methods I use in human factors analysis. If they tried to do it I'd be scared stiff. I wouldn't want them to depend on their judgments in a skill for which they have no competence. Similarly I wouldn't deal in their field for which I have very limited competence.

Meister - It would certainly help a lot if the equipment reliability types that you have to work with understood the techniques. One hopes that if a system reliability man were interested in this kind of prediction method he would, with indoctrination, be able to make use of the technique.

Swain - What scares me. I don't want to see anybody who doesn't have a human technology background get in there and try to predict human behavior.

Tolcott - You put a very heavy burden on this technique, a burden I think that may be heavier than was implied by the ADO. I think it is reflected in the kinds of questions that the engineer might pose that your suggested technique must answer and I'd like to submit that maybe this technique does not have to answer all of those questions and certainly can't answer all of those questions all of the time. It seems perfectly defensible to answer some of these questions by first, "I don't know", and secondly, "but, I can find out. I can test the concept that you're asking me about, Mr. Engineer, and find the answer."

Meister - Test how?

Tolcott - Test by simulation. I think it's important to understand, it's important to constrain the goals of the system that we're trying to develop here to something that is indeed feasible.

Meister - Admittedly what I've been describing is really an idea concept. I would not expect that initially the human reliability system would be able to answer all questions to the desired degree of accuracy. On the other hand I must protest against the concept that the engineer will stand still for being told "I don't know the answer but if you give me a simulation apparatus I will be able to tell you," because unfortunately, and this is unfortunate, the design proceeds so rapidly that in many cases he doesn't have time to wait.

Tolcott - I'm not asking him to stand still, I don't care if he doesn't stand still, but I want to avoid giving him answers that are wrong or implying that I can give him answers that I can't give him.

Meister - All human reliability predictions are going to be just that, predictions, with a certain error of estimate. He knows that. All equipment reliability predictions are predictions with errors of estimates and he knows that too. Engineering is not an exact science and he knows that too better than any of us, so he is generally willing to accept reasonable approximations of answers, but I don't think that you can slough him off by saying "I don't know the answer." I suspect that, initially, in many cases that is going to be the only truthful answer you can give him, but if you are going to develop an effective, emphasis on effective, technique you are going to have to answer many questions which will pose severe difficulties.

Siegel - There is no doubt about that Dave, but within the constraints of a reliability prediction technique, we're not trying to give you the answer to cancer, nor to all problems in one pill. All we're trying to do is give you a limited set of answers and there are other tools available to answer other questions.

Meister - Let me say this, Art, with all due recognition to our own fallibility as scientists, we must keep in mind what this human reliability structure should be able to do. Because if we say, as you have just said, well, obviously it is impossible to do this right now, and I can't give you the answer to cancer, and all that sort of thing, what generally happens is that the necessity for doing research to answer these questions gets thrown into the background. What we have, then as a consequence, are some of these piddling little techniques which are unfortunately only too characteristic of the human factors area; which satisfy us as specialists temporarily and don't satisfy any real system development use.

Blanchard - Are we talking here too about when the question is asked, like presumably this ADO is going to provide a mechanism by which this question can be asked earlier. Maybe we won't have, hopefully, the kind of urgency that we are talking about, this need to stand there at the drawing board and make snap decisions. What we're hoping is that if we can get into design early enough in the conceptual stages, that these questions will become obvious much earlier and would allow us perhaps to apply more rigorous tests, to take more time to run simulation studies than perhaps we have had the opportunity to do in the past. That would be my hope anyway.

Meister - Let me say this, Bob, even if the human engineer were right in on the initial design concept stage, as we all would like to be, the more molecular kinds of questions, such as those dealing with human performance equivalents of component attributes, manpower, and so forth, will eventually appear. You get down to a more molecular stage of system development and more molecular questions will be asked. It's not a function then of whether you were present right at the inception of the system development. These questions do arise and I suspect that even if you had a very direct hand in the initial conceptualization of the system that you're still going to get those questions asked of you.

Regulinski - I think they'll come from the engineer; I have a great uneasy feeling of a faulty conception of our reliability engineer, and I'd like very much to describe one to you, if I may. There are at least three universities now that I'm aware of, which give graduate degrees in reliability engineering. I mention obviously first my own institution, the Air Force Institute of Technology, then Arizona which followed a few years ago and Stanford which has picked up the ball just recently. If you examine their curricula you'll find that these engineers are not total ignoramuses in your field. Why? Because the students are required to take at least two graduate courses in your area of competence, namely numan engineering and human factors. At the Air Force Institute of Technology we've a behavioral scientist, and not an engineering type, teaching this too. So I don't think that you should be frightened if this reliability engineer should predict human performance reliability. I might even suggest to you that he might be horrified at your attempts to do human reliability modeling, thinking "This man has absolutely no knowledge of mathematical modeling. For him to predict human performance and reliability is nonsense."

Jenkins - Well, Ted, as a user of this concept, I think that the point is not whether the person wears a title of human factors or R&M engineer, it's his capability to do what is required, that is, what is necessary when the government or the company imposes this kind of requirement. I don't think we care what his title is as long as he is qualified. I don't think, really, Dave cares that much as long as he is qualified.

Meister - I don't care at all.

Regulinski - I detected a line...

Swain - You're putting words in my mouth, so I'll clarify my thoughts. I've been working with a group of reliability engineers since 1961. I know them very well, and they're personal friends of mine. I know what their capabilities are, what their limitations are. The point is this, Ted, I taught some of them in course work at the graduate level a little bit of what you can teach them in 1 or 2 courses in human factors. This does not make them an expert in the human behavior technology area, any more than two courses in some engineering field would make me an expert in engineering. The big thing with the course that you people teach, I think, is to get the reliability specialists to recognize when they need expertise in the human factors area; not to make a human factors expert out of them. I feel that in the human reliability field that unless somebody has a real good background, including the technology and the expertise in behavior technology, he has no business doing human reliability prediction work.

Meister - Isn't this somewhat irrelevant, who does what? I think you're passing off on to a side issue, if I may say so. The question is not who uses what, but what that what consists of, and how that what is to be developed, and whether in fact the criteria that I called out are criteria that should be employed and used to evaluate in the presently available techniques.

Swain - I think that the concensus here is that your first criteria is not appropriate if you define it as saying that people not trained in human technology can do it.

Meister - O.K., let's pass on from Number 1. Any comment on Number 2?

Coburn - I'd like to comment on some of these areas if I might. I don't think you commented on Number 3 when you were speaking, Dave, but in the last part of your paper you mentioned that the technique should not require derivation of data through utilizing judgment by experts. I would not like to accept this at the outset as a criterion. I think before this session is over maybe we'll have other viewpoints. Let's consider it as a possibility.

Blanchard - Can I ask about the 3rd one so we can take them in order. The initial comment here, Dave, suggested that the technique should not require the use of performance data which are not readily available. I'll agree to that if we clarify it. We don't want this technique to become specific to a body of data that just might be available at that point in time. In other words, you might be building a poor model to reflect poor data. I think the model should be amenable to current data and we should consider in its development what we have available by way of data, but we should keep in mind that this model has got to be flexible to some future point in time at which we have good data. I interpreted you to mean use of current data was the primary criterion in model development, I would consider it to be very secondary.

Meister - Yes, really what I should have done and I didn't do, is to rank the criteria in terms of order and importance. They're not all equally important.

Coburn - I think we'll find Number 4 is the most important one.

Meister - Let me say something about this availability of data. I think that a technique which requires one to sit down and perform various types of experimentation in order to utilize the technique to apply it to a specific system development project is one which fails, in a certain sense, because the whole idea of prediction is that you don't have to do the specific experimentation in order to get the kind of answers you want. That's all I had in mind.

Blanchard - Later on I'll have a couple of more comments on this question.

Jenkins - We have to run along a little bit, I just had one question which I think is important from the management point of view. You make four recommendations and the first one is to combine the various techniques which we have now, or to analyze them, and to select the most effective elements for these to arrive at the starting point for further research. Now, I don't know how many techniques there might be around right now.

Meister - I don't think there are that many.

Jenkins - I don't think so either, but if we look at the total design process from the time that the system concept is first thought of down to the time that it is in the fleet or in the service, it appears to me that there are certain more critical stages than others that human reliability should address ourselves to. For example, as much as we would like we're not really going to be getting into the system concept phase very much. When they come out with a new piece of equipment the design requirements do come from the operational commands. The ideas which they have are fairly rigid. They are given to the builders of the system, Naval Material Command, to achieve. Now human factors just doesn't get in there to any appreciable extent. The first time we're affected by a concept is when the responding command says here's a system now, let's, as best as we can, match the people portion of this system with the equipment portion. It is that point that human factors can enter. When it comes to validation of whether we're right, sometimes it's easy and sometimes it's not. Generally, the point they feel in validation is, "Does the equipment meet some very broad requirements?" not, "Do you get operational accuracy to the nth degree?", because there's too many intervening variables which cannot be controlled at the test site. I think that if our approach can be directed toward that portion of system development which can be most easily influenced by human reliability methods and data banks, that's where we should put our money.

Blanchard - Let me ask a question, "It seems to me you're giving us the official viewpoint in terms of how concepts are evolved. It has been my experience in several instances that PTA's are not really a requirement document written by CNO. Most of these PTA documents, requirement documents, are written by labs and systems commands. They are the ones who have the ideas. These documents do get funneled up and, officially, CNO says, "This is what I want", but a lot of the conceptualization and the study that should go into testing concepts are done by the laboratories and by the technical command, in my opinion.

Jenkins - I can't deny that, but the point is that among the people in the laboratories and the technical commands who do that, the human factors participation is very small.

Tolcott - Jim, isn't that exactly what this ADO is aimed at, reliability prediction methodology. I think of that as a tool to help us overcome this deficiency of inadequate participation.

Jenkins - Right, but because the ADO in this program has to be sensitive to things outside of its particular sphere of influence. For example, the capability of the laboratories and the systems commands to have human factors people at the right time in the right place. We have to live in a real world. They're not there. It doesn't look as if they're going to be there.

Meister - Are you objecting to the criterion that the methodology should be capable of being applied at any stage in the system development?

Jenkins - I'm saying that I think it has to be refined to a point of where it would get the optimum payoff.

Tolcott - Jim, I would suggest that, if this effort is indeed successful, it will open up the gates to more human factors application at the point where it is needed most. I think one of the reasons for the lack of application and people right now is that there is nothing like this that can be applied. We're working toward that. We should not accept that particular constraint as a constraint on the objectives.

Momiyama - I think we have a little bit of a paradoxical situation here where, as I said, this is an Advanced Development Project because that is where you get more money to do things. We're talking about a whole spectrum of exploratory development or even going back to the research and since in the human factors area, we don't have a generic human factors project in the research area or Exploratory Development, we have to sort of expand our scope here in this Project a little bit to include some of the things we're talking about. At the same time advanced development project does call for a specific result.

Blanchard - An ADO, as I understand it, is addressed to an available capability with potential application. The ADO states or is fixed to a concept or can there be alternative concepts under consideration?

Momiyama - An ADO is normally for advanced development and engineer development. We should have definite approaches of what we are doing. We don't have too many alternatives. Exploratory development is where we are evaluating the various alternatives.

Meister - All I'm trying to say is, that any technique, if it is sound, if it's effective or useful at one stage in development, whether or not it's an advanced stage of development, should also be effective at an earlier stage. If the Navy sometime ever gets around to the point of developing a system which does require a heavy human factors input right at the TDP, whatever the stage is you want to talk about, it should be useful at that stage too. It should be capable of being utilized at all stages. I see continuity of kinds of questions throughout system development, so that if a technique is useful and will answer certain questions, say midway through the design, it should also be able to answer questions very early, because there is a logical continuity.

Mills - Dave is not referring to a technique. Let's put it this way he is referring to a technique I believe, as the total technique involved which under it is subsumed possibly several other techniques.

Siegel - I didn't read that in the paper.

Mills - Maybe I read more into the paper than the rest of you did. I felt that it established an operating structure for the development of the technique under which all other more specific techniques that are involved are subsumed. The whole concept may take 20 years to satisfy and all this is, is an operating structure for how you're going to direct your effort. What are you going to direct your effort to first, for example, is what I'm saying?

Mackie - I think there is possibly a very real problem here, at least it seems to me so. I'm not encumbered by knowledge of this field. I was impressed with your description, Dave, of the kinds of data that we're interested in and it seemed to me that you implied that every kind of performance data that psychology, or any other discipline, has developed is applicable in some way to this problem area. Then we have a variety of criteria that relate to problems of validation, although, as I read in this field, it seems to me that virtually none of the types of data you described have been validated. I wonder if we wouldn't end up with quite a different approach if we started, let's say, by looking first at the operational world and deciding at the outset what it is that we're going to be able to use as validation data. I understand why you emphasize, as strongly as you do, the answering of questions for these Army engineers. But it seems to me that this is what leads you to a requirement for data. For example, on what the probability is that a particular toggle switch will be thrown. It would seem that if we start to think about what is going to be possible in the way of obtaining operational data for validation, we might end up considering quite different classes of behavior. I think you made the statement that eventually we want to go to the Navy and specify the kinds of data we need for validation of the model. I'm not sure that we can do it that way. I'm concerned that maybe we should start the other way around and look at the operating context of the system of interest and decide what is ever going to be obtainable, realistically, in the way of operational data. It may be that when we do that, we'll find that we won't be able to validate toggle switch throwing.

Meister - We may not be able to validate toggle switch throwing in the operational environment but we may be able to validate it in a laboratory setting. Since the data bank aspect is an essential part of the human reliability methodology, since the data bank assumes that we are going to be able to validate prediction in the operational environment, in the initial development of the human reliability structure you do have to take into account what it is you will be able to do in the way of gathering data operationally. However, I would not like to make the present availability of opportunities for gathering data operationally to be the primary consideration in developing this human reliability structure, because our opportunities right now are quite limited, as you suggest, because of the recalcitrance of the military. If we go that route, then the kinds of answers that we're going to get, the kind of human reliability structure we're going to build, will be extremely inadequate. In my paper I said something about developing a technique which is actually useful, and not just a research tool. We can develop all sorts of research tools that can be based solely on simulation studies and things of that sort. These will not answer the questions that are asked during development. It seems to me that we have to look at all of the elements that would enter into a human reliability structure in attempting to develop that structure; to concentrate just on 1 or 2 or 3 will be to give us inadequate answers. Now, I sort of have a feeling of being the devil's advocate in this discussion. I do not suggest that we are not faced with very real problems of data collection, I do not suggest that the criteria that I've advanced have to be met fully, this year or next year, but I do say, and I think I can support this point of view, that on a long term basis a human reliability structure requires a long term program even though the military customarily likes to think they get answers in one year and that's it. Over a long term the structure that we develop, if it is to be really effective, must answer these questions, must have these elements. I don't necessarily look for us to have such a structure next year, or two years from now, it may be five years from now, it may be longer, but we have to address ourselves to this kind of question. Otherwise what we will get out of our efforts will be piddling and, speaking as somebody who has seen an awful lot of piddling research and an awful lot of piddling human factors techniques, I think of the human reliability program not just in terms of what we're involved with for the Navy but in terms of importance to the entire human factors discipline. That's too damned important to be left to the military customer. It is essential to our entire discipline, it may in fact be the heart's blood of our entire discipline. Therefore it has to take into consideration all of these things. Maybe the military will only give us an opportunity to get half way.

Mackie - I agree with your argument, Dave, except I want to interrupt you to reinforce this point. I don't think it's the recalcitrance of the military; I think we lack the corresponding development of data gathering and observational techniques in the operating environment that have been so well developed for the laboratory where most of our data come from. My only suggestion is I think emphasis on the operational world will lead us to a different level of definition. We can't ask one approach to be applicable for all needs. I think that we haven't paid sufficient attention to the kind of data that are obtainable if we sharpen up our observational tools for the operational environment. I think this is why there is so little validation; we just haven't got the right kinds of data and there are some kinds we're never going to get unless the military allows us to wire the guy up and get his GSRs and everything else, while he's operating. We know we're not going to be able to do that. The question is, what can we do, then, in the operating environment?

Meister - I want to fall back to my prepared position. This is my main defense. Will your technique, anyone's technique, supply answers to these system development questions. They must, otherwise the technique has no significance.

Siegel - I think that the answer has to be, to paraphrase Shakespeare, "It's not that I love your system less, but I love utility more", and he's arguing that you just are not going to get there by collecting such data (as you've written) down to four decimal points. It's just a will of the wisp. It's almost a practically impossible thing to do on the basis of operational situations. I think you're going to find there is a myriad of operational situations you're talking about for this four decimal point number and it may vary by + or - .001 as the operational situation varies. You have pointed out it's not the task but it's the interactions of these various things that are important. It seems that we always have to reinvent all of psychology to get to this data bank. It seems to me that you might ask yourself the question, what is the purpose of this reliability technique? As I read CNO, it is to help the decision maker and maybe the fundamental question is not what is the predicted validity but maybe the fundamental question is, to what extent does these numerics or whatever technique you folks come up with help the decision maker? Rest assured that he's not going to make his decision only on the basis of these inputs and the criterion in this case would be did I help him in deriving the decision? I would also ask the question, if this is the case, is it the technique that you're working on which needs validation or is it the decision makers decision which was based on my data and other data which we should be validating? Then it becomes a question of: do we predict the validity or do we base our arguments on construct validity, content validity, concurrent validity and so on? There is really no sacred cow about predictive validity. There is no reason why we should dance to the tune of the test constructor's music. Maybe we hear a different tune.

Meister - Let me ask you this question. You've been talking about all the things that you can't do. Let me put it to you bluntly. What kind of questions can your technique answer? What's the numeric you can give me as a system development user, and we'll go on from there? What's the minimum that you can give me? What kind of questions can you answer? If you can't answer any questions, then I'm a little bit dubious.

Mills - This is the point, these questions direct the development of these techniques to a large extent. They also direct the operational validity test. What kind of data do you want to go out and collect. You have at least this effort going on along with the development of this kind of a data system, and that is the development of the operational principles. The only way you're going to know how to direct these questions of data collection, data development, and data system development is if you have some notion of what kind of questions you're going to be asked, that to me is the problem here. What he is trying to do in this paper, I felt is to try to set up some sort of an operating structure for the development of this kind of a data system. Now which questions should first be answered. That's a point of discussion and debate.

Mackie - I think it sure is. The thing I'm perplexed by is what seems to me to be an approach which says let's start with all conceivable kinds of behavior that we know of because our engineer may ask us about anything. Therefore, we have to have enormous quantities of data from some place and all psychology hasn't produced enough systematic data to tell us even how people solve simple troubleshooting problems. You can't start that way. If you look at it from the standpoint of CNO, one way you could start is to say, "Okay, where did my mission fail?" "What broke down here?" It may be that we don't have to worry about 99% of the behavior that went on but only rather infrequent and limited types of behaviors.

Meister - You will recall that I said that one of the elements of the human reliability methodology was determination of what equipment and behavioral parameters we're going to include. It was suggested that not every parameter is to be included or should be included. Whatever approach you take, you have to have in mind some kind of an output in terms

of an answer to a question that your approach will answer. Whether that approach will answer any questions or only a few, this is a matter for determination by the people who fund you and so forth. What you must be able to tell me is what the output of your approach is going to be. If you can't tell me that, then I don't know how you can even begin to start.
It seems to me that not to have in mind the specific questions which your technique, your methodology will eventually answer is to assume that all behavior is your province, because you have not narrowed anything down. I don't care whether you say your technique can handle function analysis or manpower loading or system layout or what have you. But I do think it's incumbent upon people who work in this area to say, "Well this is what my approach will handle, hopefully, when I complete it," whatever it might be. Otherwise you are asking the military and the other people who are in our field to buy a pig in a poke. I don't really think it's unreasonable to ask this question.

Regulinski - I'd like to suggest that there are some questions that your system engineer is definitely going to ask you. Our current approach in systems analysis may be demonstrated by the following example: Take a machine operating in a system but requiring interconnection by some human operation. The systems engineer before he goes through prognosis, whichever technique he may choose to use (not the medeival techniques), he's likely to ask the simple question "Is this operation that you would like to affect subject to classification as to whether it belongs to some generic family of activity distribution." If you say yes to him, he is likely to ask you: "What is the underlying density function governing that distribution?" If you can answer that, you've got an answer to what data you're supposed to be collecting. The systems engineer is not interested in reliability of the human operator at the switch. He could care less about that. To him, and to the systems modeler the governing density is of utmost importance.

Jenkins - Could you state that in perhaps another way, namely, what the systems man is interested in is the effects of human error with respect to total mission success not the reasons for human error.

Regulinski - I would not say it that way. The systems engineer is basically looking for a mathematical model to describe the functionality of the man-machine system in toto. What he looks for is basically this: Assume that the system I've sketched on the blackboard is some man-machine control system in which the following is happening: This is the human being. He is judiciously chosen. He is a pillar of stability in a certain sense. He's not likely to be ruffled. He's got to make a judgement and a decision before, say, this critical interconnection is made which may in turn cause, for example, the lid to open up and the 152 system to rise majestically, some time later KIEV to be no more, and our state department to pound its chest ... mea culpa, mea culpa (and damn it that's an expensive culpa)! To predict the man-machine reliability, the system engineer must know the density function governing the behavior of subsystems interconnected, and if the behavioral scientist can give him the density function of the human performance, the two functions jointly answer the question: "What is the density function governing this particular man-machine operation. If you have it, then the system engineer has the total mathematical model, but if you give him something less, like a point probability, he is likely to give it back to you as useless.

Blanchard - I think there is a point here and that is the level at which we are operating with this systems engineer in providing an overall reaction to his requirements may not be compatible with the level we would have to operate in order to come up with performance predictions. We must be careful not to allow this to be a guideline in terms of the level at which we might have to proceed in human reliability. We might have to be at a much lower level than perhaps his modeling effort in the left and right hand boxes shown. I'm sure his models are far more detailed in the center box, as ours must probably be.

Meister - There are questions which the technique and methodology must answer. These questions do not necessarily have to dictate the way in which you get the answers to the questions. If it were possible, for example, to get all the answers using mathematical modeling simulation techniques or physical simulation studies or what have you, it would not make any difference, but whatever the approach utilized it must have an output which will answer Dr. Regulinski's questions as well as the design engineer's questions, or some questions. You must be able to specify that there's an answer somewhere along the line. If we can't do that, then we're in trouble, aren't we?

SOME RECENT EFFORTS TOWARD THE DEVELOPMENT OF A HUMAN PERFORMANCE RELIABILITY DATA SYSTEM AND SUPPORTIVE HUMAN PERFORMANCE RELIABILITY PRINCIPLES

Robert G. Mills

Air Force Systems Command
Wright-Patterson Air Force Base, Ohio

SOME RECENT EFFORTS TOWARD THE DEVELOPMENT OF A HUMAN PERFORMANCE RELIABILITY DATA SYSTEM AND SUPPORTIVE HUMAN PERFORMANCE RELIABILITY PRINCIPLES[1]

Robert G. Mills[2]
Aerospace Medical Research Laboratory
Aerospace Medical Division
Air Force Systems Command
Wright-Patterson Air Force Base, Ohio 45433

Introduction

The intent of this paper is to briefly acquaint the participants of the Human Reliability Workshop with recent efforts in the area of human performance reliability (HPR) being performed by the Systems Effectiveness Branch (MRHS) of the Aerospace Medical Research Laboratory. I have chosen this approach first because our program is newly established and may not as yet be familiar to most Workshop participants. Secondly, I feel that a discussion of our present activities will serve quite well to point out where at least we feel some of the deficiencies in HPR exist.

At the present time, we have two major HPR projects underway at MRHS. The long range objectives of these are to eventually develop a computerized HPR data system for the Air Force and conduct in-house research studies designed to empirically develop some of the HPR behavioral principles which are so badly needed in this area. A third major project is a joint research effort which is being funded by Rome Air Development Center (RADC). The long range objective of this effort is to determine the impact of human performance variables on Air Force equipment reliability and to develop a reliability predictive technique incorporating these variables. Accordingly, and to be consistent, I prefer to think that this aspect of HPR research involves primarily the determination of operational (man-machine) principles of HPR.

Unfortunately, this Workshop comes at a time when our program is just getting under way; and with the exception of the latter project above, I have very few tangible results to report. Aside from this limitation, however, I am confident that knowledge of these HPR projects can contribute to the development of the Navy's program in this area.

Development of a Human Performance Reliability Data System[3]

The development of a human performance reliability data system (HPR-DS) has long been a recognized but relatively unsatisfied need in the HPR area. In fact, the lack of an HPR-DS has undoubtedly severly constrained the capability of human factors personnel to advance HPR applications in system design. We embarked on a goal of establishing a limited computerized HPR-DS within the late 1972 time period. Although the problems which will have to be dealt with in this project are monumental, I view this time frame as being realistic and limited only by future funding environments and the perseverance of the individuals involved in its development.

Among the problems which I feel must be resolved, the following are considered to be of greatest importance.

1. The determination of HPR data requirements in terms of users' applications needs at varying levels of system design.

2. The determination and establishement of a valid, applicable, and substantial HPR data base.

3. The development of the necessary transformation and classification techniques to get from (2) to (1) above.

4. The development of the necessary models for applying HPR to the reliability analysis of systems.

5. The development of the necessary software for the implementation of an integrative HPR-DS which is interactive with users and data as well as modifiable in terms of new HPR methods development.

6. The formation of a formal organizational structure for the management of an HPR-DS.

In presenting this list, I do not intend to imply that it is exhaustive. In fact, it is a gross but necessary oversimplification. Furthermore, it is not meant to imply that a state-of-the-art does not already exist for each of these problem areas. For example, it is my opinion that the area of model development for application of HPR data (item 4) has progressed quite rapidly.

In a contractual effort recently initiated by MRHS, we hope to get a start toward resolution of the first three items above. Under this effort, we will examine all levels (e.g., conceptual, basic hardware design, and task allocation) of the man-machine development process to determine the engineering requirements for HPR data at each level. These requirements will be in the form of critical parameters, metrics, assumptions, etc. We will also examine analytical methods for applying HPR data at each design level and for assuring compatibility with hardware reliability. Based on this examination, a set of HPR data specifications will be formulated.

A second aspect of this effort will be to examine existing data sources to determine their capability for meeting HPR data requirements. It is not expected that data from these sources will be available in a form readily suitable to defined HPR requirements, thus necessitating the development of data extraction and transformation methods. In fact, the general lack of HPR data has received pointed discussion throughout the literature but with few results. Studies such as the RADC effort discussed below and that reported by Askren and Regulingski (1969) appear to be practically nonexistent.

However, what may be an even more serious deficiency, not generally recognized, is the lack of a methodology for HPR data extraction from either general sources of experimental data or even more specialized sources such as those used by Payne and Altman (1962). To presume that specific HPR research will provide necessary data within several decades is, in my opinion, unrealistic. This is because of the sheer effort and cost involved in conducting research which would eventually cover a sufficient number of gaps in HPR data. It would, therefore, be more practical, realistic, and well worth the risk--especially in terms of the demonstrated competence of human factors personnel in model development--to attempt to develop the necessary transformation methods for extracting HPR data from experimental literature. This does not negate the need for research; rather, it puts this need into better perspective. Thus, given that massive HPR data can be made available, research is needed to derive the necessary behavioral and operational principles and assumptions to be included in HPR applications models.

It should be emphasized that this first effort is exploratory and developmental in nature. It is intended to lay the groundwork for future development of the HPR-DS by establishing the standards for data collection both in terms of the kinds of data to be collected to assure applicability to the system design process and the sources of these data (i.e., in terms of user requirements and capabilities for data extraction). Furthermore, we are not contending, absolutely, that valid data can be obtained from the experimental literature. Instead, we feel that methods for extracting data from this vast source should be explored more extensively than has been the case in the past.

The development of methods for deriving a HPR data base is, of course, the portion of this effort which is most uncertain in outcome. Some of the uncertainty will be alleviated once the HPR data specifications are written because these will serve as a formative structure. To date, however, all that we can be sure of is that the general procedure will require a series of coordinated classification schemes relating, for example, machine, behavioral and system development functions, and also environmental conditions. I understand that Dr. Meister will have more to say about this area in his paper for the Workshop.

Concerning item 5 above, I do not anticipate a great deal of difficulty with the development of HPR-DS storage and retrieval software which are easily within the state-of-the-art. The interactive capability of the HPR-DS may be considerably more difficult to develop, however, especially with regard to HPR data input requiring computerization of transformation methods which may rely a great deal upon subjectivity.

For HPR data output, software will have to be developed which will provide users with a wide query capability both latitudinally (across HPR data) and longitudinally (across system development functions) within the data system.

Establishing the organizational structure for managing an HPR-DS is a formidable problem indeed. Most of the problems in this area have been noted before by Swain (1964) who also recognized that such a system should be managed through a governmental agency. Since it is doubtful that a large number of competitive industrial organizations could freely participate in an industrially managed system, this would probably be the most satisfactory solution. Another concern here is whether or not all the services should sponsor a single central HPR-DS instead of individually sponsoring several ones. While a central data system might be more efficient for development and cost, it might also be unwieldy in operation because of the diversity of operational systems. Although we recognize the significance of these questions, the stage of development of our program in this area is embryonic and still somewhat insecure. Thus, we have not as yet directly addressed this problem.

Research to Determine Behavioral HPR Principles[4]

As noted above, given a capability for obtaining specific HPR data, much of the required HPR research should be directed toward establishing general behavioral HPR principles. The most pressing problem in this area involves the assumption of independent task elements required by the use of the product rule in HPR models. Its solution requires an attempt to ascertain and quantify the relationship between low-order, behavioral task elements variously defined as task elements, behavioral units or discrete stimulus-response units and high-order tasks--also variously defined as subtasks, specific behaviors, or task functions. We have recently initiated an in-house, basic research program designed to study this problem and others, such as the determination of the characteristics of HPR and time distributions and determining the effects of behavioral redundancy upon HPR.

Our first study has just been initiated and data are not yet available. However, a general description of this study should serve to exemplify our approach to this research area. I hope some preliminary data can be made available at the time the Workshop meets.

The study is similar in purpose to that of Buckner and McGrath (as reported in Swain, 1964) who found that the empirical probability of detecting a stimulus with combined auditory and visual stimuli was .91. This was compared with a predicted probability of .97 obtained under the assumption of independent, auditory, and visual detection elements. The study is also similar in purpose to that of Askren and Regulinski (1969) who derived and tested a general mathematical model of HPR for time continuous tasks.

The procedure we are using is a bit complex and is best described by the event sequences shown in figures 1 and 2 In general, HPR and time distributions are being obtained on a set of six discrete, what I will call, reference tasks. The reference tasks being used vary in behavioral complexity and may be summarized as follows.

FIGURE 1-SEQUENCING OF EVENTS FOR HPR INDEPENDENT TASKS, SUBJECTS COMPLETE EACH TASK IN BLOCKS OF 50 TRIALS.

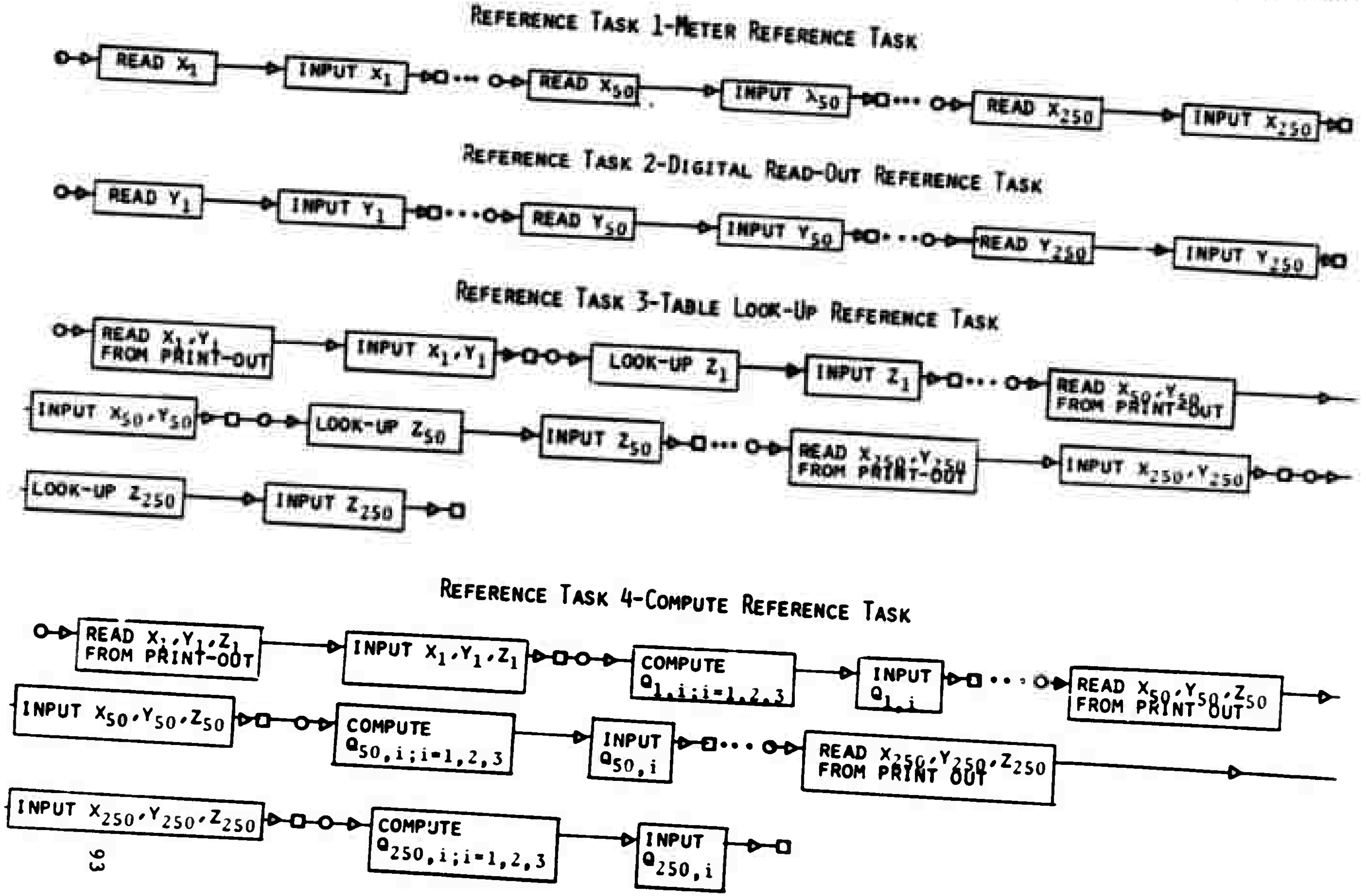
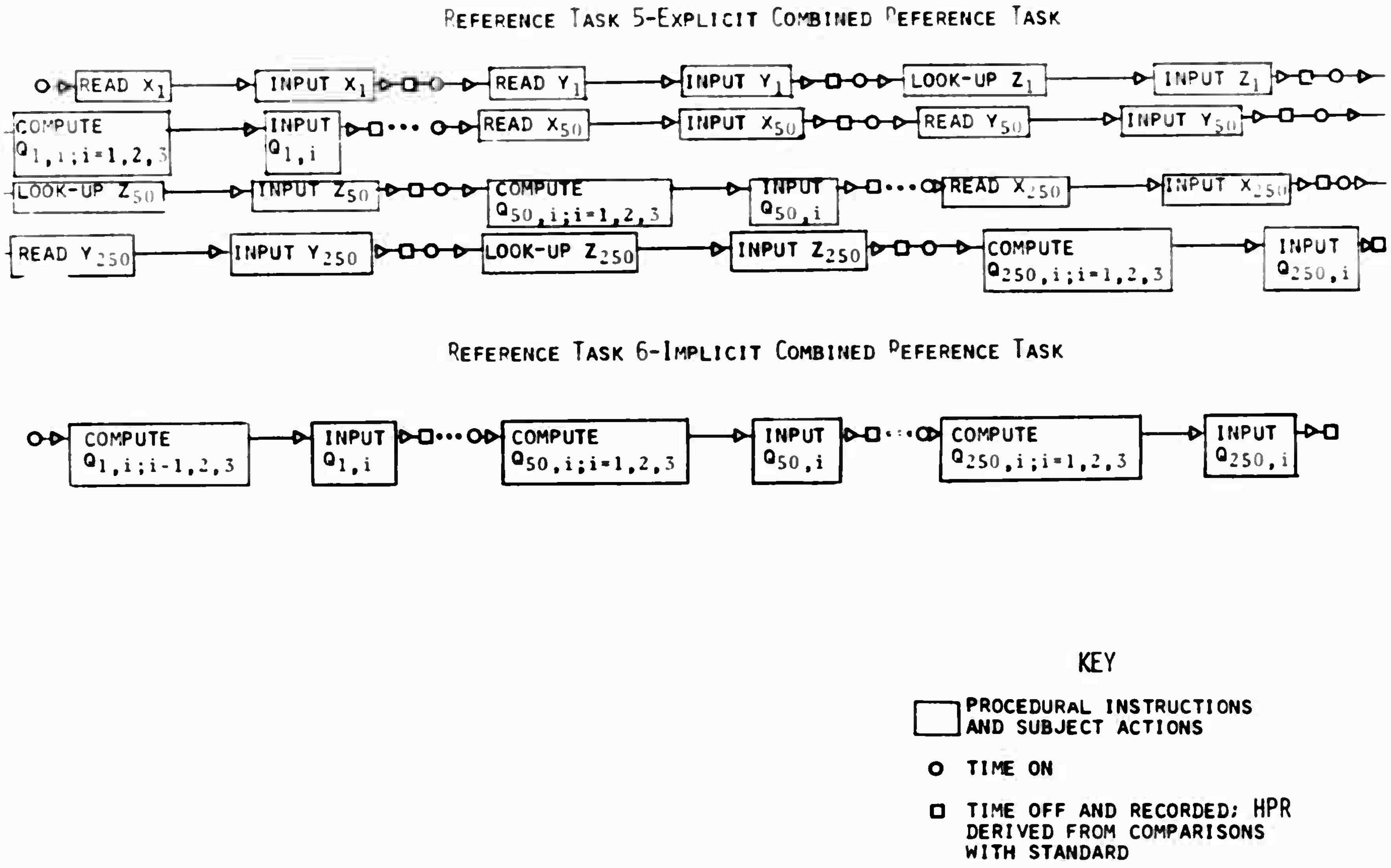


Figure 2-Sequencing of Events for HPR Combined Tasks. Subjects Complete Each Task in 3locks of 50 Trials.



Reference Task 5-Explicit Combined Peference Task

Reference Task 6-Implicit Combined Peference Task

1. For reference task 1: the subject reads a value (X) displayed on a meter and inputs the value on a numeric keyboard. The time to read X (read and input) is recorded along with the encoded X value.

2. For reference task 2: the subject reads a value (Y) displayed on a digital read-out display and inputs the value. The time to read Y and the encoded Y value are also recorded.

3. For reference task 3: the subject reads X and Y from a computer print-out (which displays these values simultaneously). These values are input and recorded with the time. The subject then looks up a table value (Z) by accessing a set of tables using X and Y as coordinates. The time to perform this operation is also recorded along with the Z value.

4. For reference task 4: the subject reads X, Y, and Z from a computer print-out and inputs the three values. These are recorded along with the time. The subject then computes and inputs a value (Q) calculated using one of three formulas selected randomly, e.g., Q = X·Y/Z. The time to perform the computation and input, along with the Q value, is recorded.

5. For reference task 5: the subject obtains X, Y, and Z as in tasks 1, 2, and 3 and computes Q as in task 4. Time and values are recorded after each operation.

6. For reference task 6: the subject performs the same operations as in task 5 with the exception that he is not given procedural instructions to read X, read Y, and look up Z. These operations are implicitly assumed to have been performed in order to compute Q. The time to perform all implicit operations and compute Q is recorded along with the Q value.

Reference tasks 1-4 are considered to represent independent task elements. In the case of reference tasks 3 and 4, however, independence is difficult to assure because these tasks involve operations performed 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 and compute aspects of reference tasks 3 and 4.

Measures are being obtained over a large number of trials (a minimum of 250 in blocks of 50 trials) on each task in order to develop sufficient HPR and time distributions. The apparatus includes a device comprised of 5 meters, one of which is used to display X, a digital numeric display Y, and a numeric keyboard which subjects use to input required values. The remaining 4 meters are used to display extraneous information that is not used in any of the tasks. Also used is a device for displaying single lines of a computer print-out on a trial-by-trial basis and two books of Z tables. All recording is automatic and subjects complete tasks in random order.

HPR distributions will be derived for each reference task on the basis of comparisons with standards. Thus, for example, HPR on reference task 1 will be distributed in the form of deviations from standard X values. HPR on reference tasks 2 and 3, however, will be binary in form, i.e., correct or incorrect. At the time 250 trials have been accumulated on each task, obtained HPR and time distributions will be evaluated. If these are considered inadequate, the entire task order will be replicated. This will provide 500 observations on each of the reference tasks.

Using this procedure, it is hoped to obtain relationships between HPR and time measures determined for the independent task elements (see figure 1) and the same measures determined on the combined reference tasks (see figure 2). A typical comparison to be made, for example, will be to determine the relationship between subject performance on reference task 1 and performance on the same task when it is an explicit task element of reference task 5. Other comparisons will involve determining the relationship between the performance of subjects on reference tasks 1-4 combined and performance on reference task 6. The latter analysis will specifically examine the problem of predicting HPR and time functions for higher-order tasks given a knowledge of these functions for lower-order independent task elements.

It should also be noted that the procedure exemplified by this study can be expanded to include other high-order tasks, task elements for serial redundancy checks, tests of the effects of time stress, etc. Furthermore, while the study itself is admittedly basic in nature, it is felt that it represents the systematic approach required to i. estigate behavioral principles of HPR.

Research to Determine Operational HPR Principles (Impact of Operator Performance Variables on Equipment Reliability)[5]

The importance of determining the impact of operator performance variables on equipment reliability lies in the fact that although a number of studies, such as that of Shapero, et al (1960), have demonstrated that human error exerts a substantial influence on systems effectiveness, little is known of the specific nature of these errors. The principle reason for this is that earlier studies have not been conducted at the required level of failure examination to discern equipment failures which are the result of human action (HIF, human initiated failure). This in turn prevents determination of the factors causing HIF and furthermore the quantification of the contribution of HIF to equipment and unreliability. As a result, the quantitative relationships between human error and system design and effectiveness parameters have not been determined. It was with these observations in mind that RADC and MRHS initiated a program in this area.

Thus far, one rather extensive field study dealing with the impact of operator performance variables on airborne equipment reliability has been completed. A final technical report (Meister, et al, 1970) will be available shortly. The general objectives of this study were as follows.

1. To develop and test a methodology for on-site determination and investigation of HIF and other human related failures as they affect the operational reliability of a select group of airborne equipment. Other human related failures refer to failures reported falsely by an operator (FR, false report) or failures reported by operators which could not be duplicated (CND, could not duplicate).

2. To relate operator performance and equipment design variables to the incidence of HIF, FR, and CND.

3. To determine the types, causes, and consequences of HIF, e.g., an HIF which is the result of a maintenance action or one which results in equipment damage.

4. To develop a means of combining operator and equipment reliability indices.

Data collection for this study was conducted over a period of 5 months by Bunker-Ramo personnel stationed at the SAC airbases of March AFB, California, and Wright-Patterson AFB, Ohio. The equipments selected for the investigation were the AN/ASG-15 or MD-9 Fire Control System, the ASB4A or ASB9A Bomb/Nav System, the MD-1 Auto Astro Compass and the APN-59 Search Radar. Each of these equipment items were selected according to a set of criteria involving principally the requirements of sufficient operator interaction, operational recency (i.e., representativeness of avionics state-of-the-art) and a sufficient failure rate.

The procedure for data collection involved interviewing operators reporting failures during the debriefing period following their flights, observation of the maintenance performed on the failed equipment, and interviews with maintenance technicians. In all, a total of six different data forms were filed by the investigators for each failure report. This number of forms was required in order to record results at each stage of the maintenance reporting system under observation. These forms were designed to obtain data from the operator concerning, for example, at what point in a mission a reported failure was first noted and what symptoms were observed. From maintenance observation and interviews, data were obtained concerning whether or not a failure could be duplicated, failure cause in the maintenance technician's opinion, etc.

The results and conclusions of this study are far too extensive to detail here. However, table I provides a categorical summary of the failure data collected. As indicated in this table, the failure categories of HIF, FR, and CND represent 36% of all failures reported. (The HIF percentage can be considered conservative because a failure was classified as HIF only when it was nearly certain no other classification would be adequate.) It is obvious from these data that the incidence of failures in these categories can be expected to contribute substantially to equipment unreliability. This, in fact, was verified through further analysis.

Table I. Summary of Frequency and Percentage of Failure Categories as Determined by Meister et al (1970)

Total Number of Failures Investigated	Total HIF		Total FR		Total CND		HIF+FR+CND		Total EF*	
	N	%	N	%	N	%	N	%	N	%
552	69	13	27	5	101	18	197	36	355	64

*EF represents the equipment failure category.

A second major conclusion of the study was that there appears to be a relationship between the occurrence of HIF, FR, and particularly CND, and certain classes of components involved in reported failures. For example, it was shown that the incidence of HIF and FR, as might be expected, was reflected best by display components which involve operator interaction. The incidence of maintenance technician caused failures, also as might be expected, was reflected best by internal components and tended to vary directly as a function of the number of connections between components. The CND category which, as shown in table I, comprised the largest human related failure category was reflected primarily by the components requiring fine perceptual and psychomotor activities (e.g., activities involving the use of cross-hairs).

The importance and contribution of this study to the area of HPR is unquestioned. While it has demonstrated that HIF has an effect on equipment unreliability, as previous studies have shown, it has in addition attempted to quantify this effect. It has also verified rather precisely the existence of several categories of human related failures and, most improtantly, demonstrated relationships between these categories and types of equipment components. Because of the significance of these results, this type of study should be continued and extended into areas of ground equipment and reliability prediction.

However, the methodology employed in the RADC study is not without rather severe limitations. For example, it can be estimated that the delineation of a single HIF cost on the average approximately 39 manhours of data collection effort alone (i.e., three full-time collectors on site for five months obtained 69 HIFs as shown in table I). This is simply the result of rigorous detective and clerical work required to ascertain the nature of these failures. A second limitation not unrelated to the first is the lack of failure rates on these equipment items sufficiently great enough to provide a large data sample during a short period of investigation. This has special implications for attempts to determine the relationships between equipment design parameters and human related failures. Thus, in this study there were many classes of components for which there were zero or small numbers of HIF, FR, and CND. This result could have obviously occurred either because these components were unrelated to these failure categories or because the sample size was not large enough to develop approximations to true failure distributions indicating existing relationships.

There appears to be no easy way around these limitations. Although the initial methodology represented by the RADC study can surely be improved to provide more efficient data collection, it cannot be comprised and still result in required data. It will, therefore, continue to be a costly procedure.

Conclusions

In this report, I have tried to summarize the major activities of MRHS in the field of HPR. In doing so, I have indicated the principle reasons for developing each program and the particular HPR deficiencies to which it is being directed. I view the need for an HPR-DS as an overriding consideration in the area and one that is fundamental to the progress of human factors in man-machine systems design. The lack of such a data system, however, also points to the greatest deficiency in HPR--that of a valid and applicable data base. I would like to see a much stronger, more coordinated effort in this area. Although each of the efforts described above has been initially directed toward this problem, these programs are relatively minor compared to the magnitude of the tasks ahead in this area.

APPENDIX

Research to Determine Behavioral HPP Principles: Preliminary Data

The preliminary data presented in this Appendix were manually transcribed for our first subject's second and third blocks of 50 trials each. (An occasional trial was lost for this subject; thus, not all sample sizes have N = 100.) These data are merely intended to exemplify the study's methodology and what it may be able to yield. The data are certainly not intended to be conclusive.

Figures A-1 and A-2 are histogram distributions of task times for each reference task. Mean times are shown next to each distribution. Means again are used for exemplary purooses only. It is realized that many of these distributions are non-nromal. Eventually, we will obtain the probability density functions for the reference task time distributions; as well as other descriptive measures such as confidence estimates.

There are two principal observations to be made about Figures A-1 and A-2. First, it will be noted that the distributions for the independent tasks (i.e., top half of Figure A-1) differ somewhat from those obtained for the same tasks when combined in the explicit combined reference task (i.e., bottom half of Fiqure A-1). These differences occur primarily in Z and Q. Thus, for example, times to look up Z are markedly skewed in the combined task as compared to the independent task. Because the times for computing Q are spread widely, any developing skewness in the combined task is presently not apparent. However, the times have definitely shifted left becoming, generally, shorter.

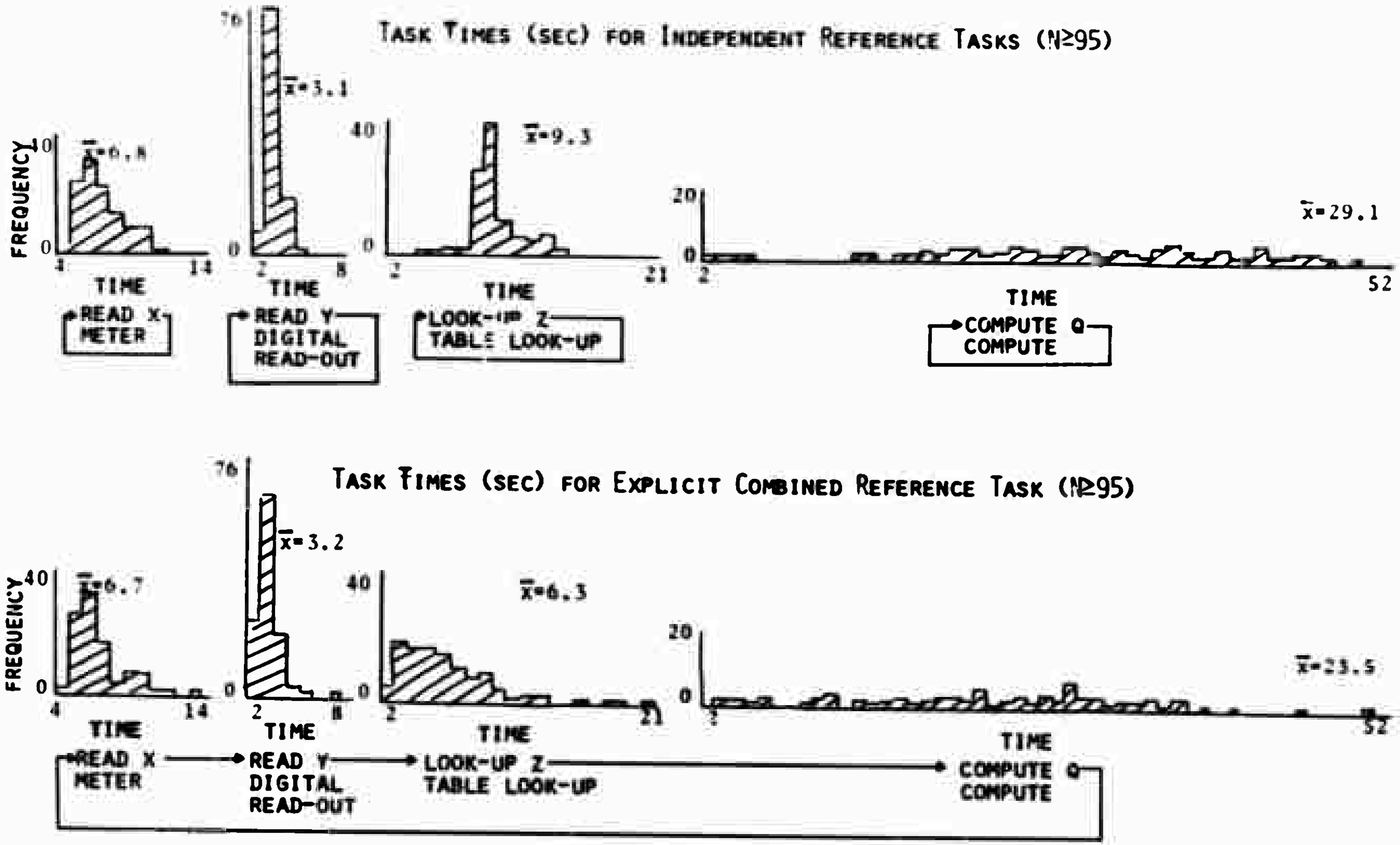
The second observation to be made from FiguresA-1 and A-2 is that the assumption of additivity of task times does not yield an accurate prediction of the task times for the implicit combined reference task (i.e., bottom half of Figure A-2). Thus, summation of the mean task times for the independent tasks yields a predicted implicit combined task mean time of 48.3 seconds as compared to an obtained mean of 55.6 seconds. The mean obtained for the explicit combined reference task (treating each task as independent and summating yields 39.9 seconds) also underestimates the implicit combined task mean.

Figure A-3 provides some examples of probability HPR formulations for the reference task excluding those involving Q. On the basis of our empirical data, we are able to obtain the HPR of the independent reference tasks, of the same tasks derived independently from the explicit combined reference task, and of the *actual* joint probability of tasks when combined in the explicit task (e.g., Pr $({}_{EX}X_{\pm 1} \cap {}_{EX}Y_{O})$ is read the probability of *both* X *and* Y being correct within a given trial of the explicit combined reference task).

It can be recognized immediately that there is a reduction in the HPR of each independent task when it is a part of the explicit combined reference task. For example, the HPR of X has been reduced from 0.98 to 0.85. This then suggests that one may not be able to accurately predict the HPR of combined (subtask) task performance given independently derived HPR for task elements.

This is further supported by the results of the combinations in Figure A-3. Thus, implication 1 states that the HPR of combined task performance cannot be accurately predicted on the basis of HPR obtained for independent tasks <u>or</u> our product-rule assumption for combining task HPR is incorrect. This statement is based on the fact that neither X and Y or X and Y and Z HPR is predicted accurately from X, Y, and Z independent HPR (i.e., 0.98 vs. 0.83 and 0.98 vs. 0.76).

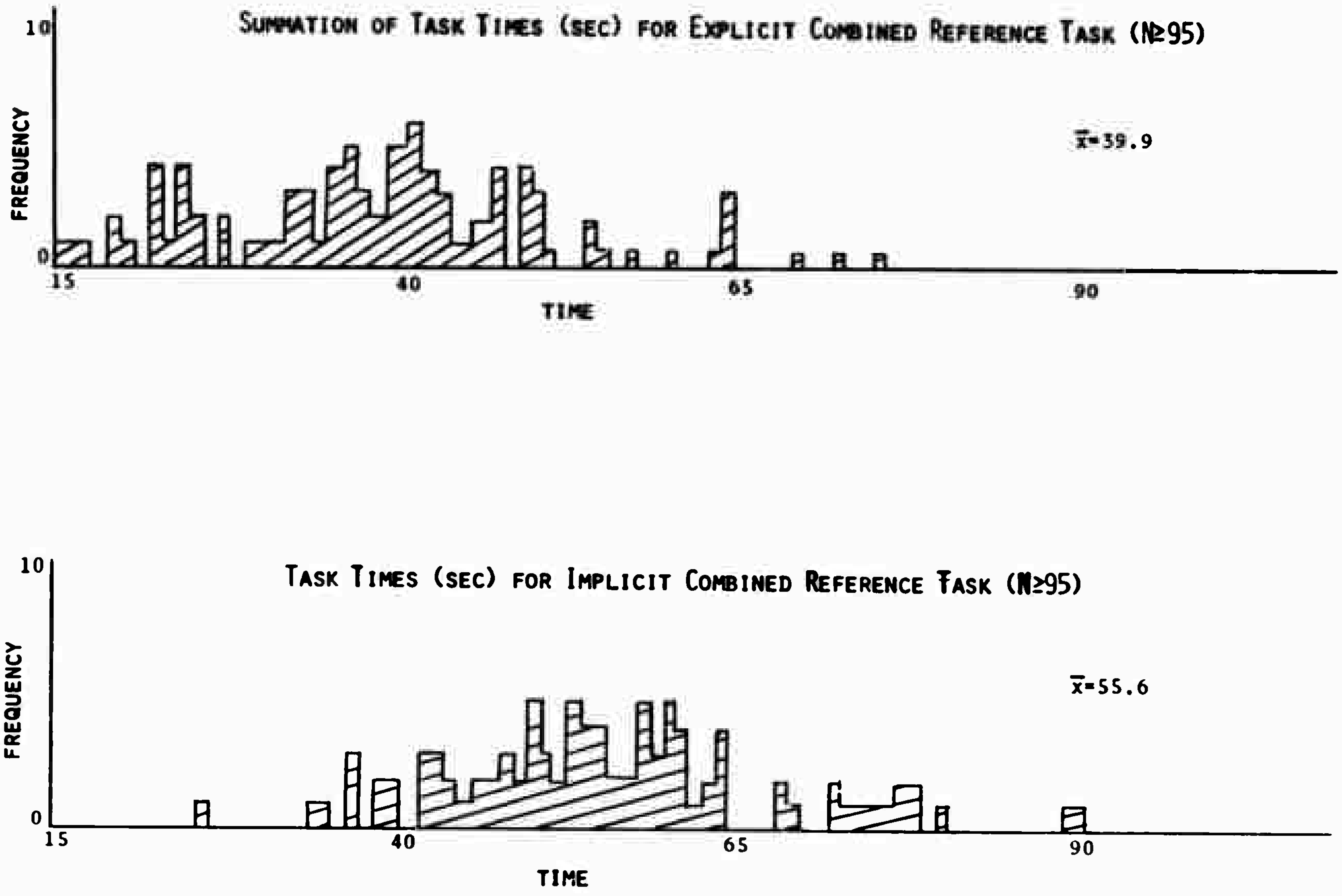
Implication 2, however, states that there may be nothino wrong with our product-rule assumption but rather, some behavioral aspect of the combinatorial process is responsible for the poor prediction. Thus, 0.833 or the obtained HPR for X and Y in the explicit combined reference task compares quite favorably with predicted HPR or 0.83 obtained from combining X and Y on the explicit task under the assumption of task independence. The same is true for the table task or 0.783 vs. 0.76.



Task Times (sec) for Independent Reference Tasks (N≥95)

Task Times (sec) for Explicit Combined Reference Task (N≥95)

Figure A-1



Summation of Task Times (sec) for Explicit Combined Reference Task (N≥95)

Task Times (sec) for Implicit Combined Reference Task (N≥95)

Figure A-2

Examples of Probability Formulations Carried Out to the Table Look-Up Task. Number of Trials= 100 for A Single Subject. $\mathrm{P}\dot{\mathrm{R}}$= Predicted Probability of Correct Performance. Pr= Obtained Probability of Correct Performance. Lefthand Subscripts in= Independent Reference Task and ex= Explicit Combined Reference Task. Righthand Subscript= tolerance (e.g., $\mathrm{Pr}({}_{EX}X_{\pm 1})$= Predicted Probability in the Explicit Combined Reference Task of Reading X Correctly to ±1 Unit Relative to the Standard X Value).

Empirically Determined Values

$\mathrm{Pr}({}_{IN}X_{\pm 1})=0.98$	$\mathrm{Pr}({}_{IN}Y_0)=1.00$	$\mathrm{Pr}({}_{IN}Z_0)=1.00$
$\mathrm{Pr}({}_{EX}X_{\pm 1})=0.85$	$\mathrm{Pr}({}_{EX}Y_0)=0.98$	$\mathrm{Pr}({}_{EX}Z_0)=0.93$

} Includes only errors in Z which can NOT be attributed to errors in X or Y

$\mathrm{Pr}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)=0.83$

$\mathrm{Pr}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0\cap{}_{EX}Z_0)=\mathrm{Pr}({}_{EX}T_0)=0.76$

Some Combinations

$\mathrm{P}\dot{\mathrm{R}}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)\longleftarrow\mathrm{Pr}({}_{IN}X_{\pm 1})\bullet\mathrm{Pr}({}_{IN}Y_0)=0.98\bullet 1.00=0.98$

$0.98\not\cong 0.83$

$\therefore\mathrm{P}\dot{\mathrm{R}}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)\not\cong\mathrm{Pr}({}_{IN}X_{\pm 1}\cap{}_{IN}Y_0)$ Implication 1

$\mathrm{P}\dot{\mathrm{R}}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)\longleftarrow\mathrm{Pr}({}_{EX}X_{\pm 1})\bullet\mathrm{Pr}({}_{EX}Y_0)=0.85\bullet 0.98=0.833$

$0.833\cong 0.83$

$\therefore\mathrm{P}\dot{\mathrm{R}}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)\cong\mathrm{Pr}({}_{EX}X_{\pm 1}\cap{}_{EX}Y_0)$ Implication 2

$\mathrm{P}\dot{\mathrm{R}}({}_{EX}T_0)\longleftarrow\mathrm{Pr}({}_{IN}X_{\pm 1})\bullet\mathrm{Pr}({}_{IN}Y_0)\bullet\mathrm{Pr}({}_{IN}Z_0)=0.98\bullet 1.00\bullet 1.00=0.98$

$0.98\not\cong 0.76$

$\therefore\mathrm{P}\dot{\mathrm{R}}({}_{EX}T_0)\not\cong\mathrm{Pr}({}_{IN}T_0)$ Implication 1

$\mathrm{P}\dot{\mathrm{R}}({}_{EX}T_0)\longleftarrow\mathrm{Pr}({}_{EX}X_{\pm 1})\bullet\mathrm{Pr}({}_{EX}Y_0)\bullet\mathrm{Pr}({}_{EX}Z_0)=0.85\bullet 0.98\bullet 0.93=0.783$

$0.783\cong 0.76$

$\therefore\mathrm{P}\dot{\mathrm{R}}({}_{EX}T_0)\cong\mathrm{Pr}({}_{EX}T_0)$ Implication 2

Figure A-3

REFERENCES

Askren, W. B. and T. L. Regulinski. *Mathematical Modeling of Human Performance Errors for Reliability Analysis of Systems*. Aerospace Medical Research Laboratory Technical Report 68-93, Wright-Patterson Air Force Base, Ohio, 1969.

Meister, D., Finley, Dorothy L., et al. *The Effect of Operator Performance Variables on Airborne Equipment Reliability*. Rome Air Development Center, Air Force Systems Command, Griffis Air Force Base, New York (in press), 1970.

Payne, D. and J. W. Altman. *An Index of Electronic Equipment Operability: Report of Development*. AIR Report AIR-C-43-1/62-FR, American Institute for Research, Pittsburg, Pennsylvania, 1962.

Shapero, A., Cooper, J. L., et al. *Human Engineering Testing and Malfunction Data Collection in Weapon System Test Programs*. WADD Technical Report 60-36, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio, 1960.

Swain, A. D. "Some Problems in the Measurement of Human Performance in Man-Machine Systems." *Human Factors, 6*, 687-700, 1964.

FOOTNOTES

[1] The research reported in this paper, except where otherwise noted, was conducted by personnel of the Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio. This paper has been identified by Aerospace Medical Research Laboratory as AMRL-TR-70-70. Further reproduction is authorized to satisfy needs of the U.S. Government.

[2] A number of individuals have been largely responsible for the efforts described in this paper. I would like to particularly acknowledge the participation of Dr. David Meister, who is the principal investigator for the two contractual efforts described; his colleagues at Bunker-Ramo Corporation; and Lt Shirley A. Hatfield, who is assisting me in conducting 6570 AMRL/HES's in-house human performance reliability research program.

[3] Human Performance Reliability for Man-Machine Systems Design, Contract No. F33615-70-C-1518.

[4] Primarily because of time limitations, the workshop presentation dealt with this section only. The preliminary data of the study described below which were presented at the workshop are described briefly in the Appendix. The Appendix was added following the workshop.

[5] The Effect of Operator Performance Variables on Airborne Equipment Reliability, Contract No. F30602-69-C-0140. Mr. Lester Gubbins (EMNRC) is RADC's monitor for this effort. Maj Donald A. Zink was 6570 AMRL/HES's initial monitor. He and Mr. Gubbins formulated the original statement of work for this effort.

DISCUSSION OF MILL's PRESENTATION

Tolcott - One of the problems of presenting the measures that you're using is that you've applied some arbitrarily defined all-or-none criteria of task accuracy. Plus or minus 1.

Mills - This occurs in the real world.

Tolcott - Yes, but what happens is that once you've applied your criterion, your probabilities are now based on that particular criterion.

Mills - I have a great tool that I work with - the computer. I can obtain these probability measures across a wide range of tolerances. If tolerance proves to be an important factor we'll have to account for it in our models. These data are all on punched tape and need the computer to get them off. Now, some data, as I mentioned, was manually transcribed. We derived these data from computer listings of the second and third sessions for this particular subject so that I could demonstrate what we're doing with this methodology. Again, I must emphasize that I do not swear by these. It's the methodology that is important right now. These are the kinds of studies that need to be done to try to develop modeling rules or at least to give us some indication of where we are with regard to the assumptions that we've been making.

DEVELOPMENT OF A HUMAN ERROR RATE DATA BANK

A. D. Swain

Human Factors and Quality Control Division
Sandia Laboratories

SC-R-70-4286

DEVELOPMENT OF A HUMAN ERROR RATE DATA BANK*

A. D. Swain
Human Factors and Quality Control Division
Sandia Laboratories
Albuquerque, New Mexico

July 1970

ABSTRACT

A program to develop a human error rate data bank for use by system planners and designers and by human reliability analysts and human engineering specialists is described. The point of view is taken that sufficient work has been done in reliability technology (including the mathematical modeling) and in behavior classification (task taxonomy) so that an interim human performance data bank can be initiated almost immediately. A procedure for establishing a noncomputerized, manual entry, interim human performance data bank is outlined. The basic criterion variable suggested is the human error rate. A method of identifying and scaling the independent variables (called performance shaping factors) is presented.

An approach for refining and validating the data bank is described. Validation includes using multivariate analysis techniques to find out what performance shaping factors have maximum postdictive power for various types of equipment/task combinations for which human error rate data was collected. It is questioned whether the gains realized from automatic, high-speed data retrieval would be worth the added complexity (and related unreliability and down time) of a computerized data bank and the time, effort, and costs of developing and maintaining it.

*This work was supported by the United Sates Atomic Energy Commission and in part by the Department of the Navy.

FORWORD

Since the early 1960's work in the area of quantifying the effects of human errors on system reliability, maintainability, and safety has been expanding. The major cry then, and now, is that there is no listing of human error rate data comparable to the tables of defect rates that specialists in equipment reliability analysis use. Several of us in the human factors field have been complaining long and loud that a Department of Defense agency should sponsor the development of a human performance data bank to provide the human error rate data and other data needed to enhance the quality of human reliability analysis work. Now is our chance. We have been told to "put our time and effort where our mouth is." We are grateful to the Naval Ship Systems Command, the Naval Air Development Center, and the Office of Naval Research for sponsoring a human reliability workshop to address just such problems as the data bank.

This paper was prepared in response to the direction to "assume that you had been given the responsibility to tell the Navy how to develop a human performance data bank." Although the opinions expressed in the paper are my own, the paper has been influenced greatly by other members of the human factors staff at Sandia Laboratories. Special thanks are due Lynn V. Rigby for his suggestions on the approach and related techniques to follow and to Messrs. Robert G. Webster and Henry E. Guttmann for their careful reviews and critiques.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEVELOPMENT OF A HUMAN ERROR RATE DATA BANK

In the early 1950's efforts were made to incorporate into system effectiveness estimates the quantitative amount of degradation that would be introduced by humans in the system. By 1964 enough persons were working in this area to warrant "The Symposium and Workshop on Quantification of Human Performance" held in Albuquerque (Nicklass et al, 1964). In the 1960's a number of human reliability predictive techniques were offered (cf. reviews by Meister, 1964; Swain 1964b; and Swain, 1969b). The primary difficulty with these techniques lay not with the mathematical modeling, but with the dearth of human performance data to be used by the techniques. In 1969 it could be said with regard to human reliability work that "... our primary need is still a central bank of human performance data. Several years ago we had reached the point where our mathematics and reliability technology had far outstripped the available human performance data. Today, if anything, the gap has widened." (Swain, 1969b).

The purpose of the present paper is to propose a program for developing, testing, and validating a human performance data bank for human reliability work. The purpose of the data bank is to enable human factors specialists to provide system planners and designers with quantitative estimates of the influence of human performance on various measures of system effectiveness. These measures can include reliability, safety, maintainability, or any other success criterion.

A Point of View

To achieve the above purpose of the human performance data bank it is necessary that the data and associated quantification of the influence of human performance on systems be compatible with conventional reliability statistics and technology. This aim has implications for the dependent and independent measures to be used and how they are treated mathematically (or modeled).

The Mathematical Model

As early as 1952 an application of conventional reliability technology was used at Sandia Laboratories to estimate during the planning stage of a nuclear system the potential effects of human errors on the effectiveness of the system when it became operational (described in Swain, 1964b, p. 688). Since 1962 this basic approach has been expanded and applied to such areas as design for producibility (Rook, 1962, and Rigby and Swain, 1968), evaluation of proposed weapon concepts (Swain, 1963a and 1963b), flight simulators (Swain, 1967a), evaluation of emergencies (Rigby and Edelman, 1968), assessment of human reliability in nuclear reactor plants (Swain, 1969a), and safety (Swain, 1969c). One explanation for the acceptance of this method by designers and reliability analysts was expressed by a Sandia manager who said, "There's nothing new to this method; it's simply an application of our reliability technology to human behavior." It can be judged, then, that a method which is an extension of conventional reliability technology is most likely to find acceptance <u>and</u> <u>use</u> in system planning and design. And that is the name of the game!

Thus it appears that it would be unwise to devote much time to mathematical modeling for quantification of human performance effects in systems. Conventional reliability technology exists and it can handle our modeling problems, including the knotty problem of nonindependence of behavioral events (Swain, 1967b) and the relationship of human behavior with other system events such as equipment malfunctions and enemy actions (Swain, 1963a and 1963b). We might well forgo any further modeling until much later in human reliability work; we should devote 99 percent of our efforts now to the data bank problem.

The Criterion Measure

There is one important aspect of human performance that is compatible with common reliability statistics and technology. That aspect is the human error rate. Human error rates correspond to the defect rates used in equipment reliability studies. If human error is defined to be any variant of human performance that reduces the probability of system or mission success (however defined), then failures due to human errors can be treated in a manner very similar to component failures; that is, human errors can be predicted as a probabilistic function of the variables determining or influencing human performance related to the system.

As Rigby (1967) has stated, human error rates "... is a straightforward, unequivocal, and generally acceptable concept; it describes exactly the kind of information we can use most effectively; and the acronym, HER, is guaranteed to get attention. More euphemistic terms such as 'human reliability,' 'zero defects,' or 'human success probability' mean different things to other specialists, such as flight surgeons, quality inspectors, and personnel people. Most people seem to be ready to accept the fact of human error, and this fact can be dealt with more effectively if dealt with openly. Too, if it is called 'human error,' it is more likely to be dealt with by behavioral scientists, as it should be. It is both useful and important, however, to distinguish, as Rook (1965) does, between situation-caused errors (SCE) and human-caused errors (HCE). Emphasis on SCE, especially when setting up error collection programs, helps remove the unfortunate and inappropriate onus attached to the words 'human error.'"

Therefore, in this paper is it recommended that human error rates (HER) be the basic human performance criterion measure to be used and that the data bank be called the Human Error Rate Data Bank.

HER must be defined broadly as indicated above. And we generally accept the fact that in systems work, a human behavior is an error only when it has the potential effect of reducing (1) system reliability or (2) the likelihood that some other system success criterion will be met. Operating within the above restriction, the following definition of human errors is suggested (Swain, 1963a):

Human errors occur:

(1) When a man fails to perform a task or part of a task (e.g., step),

(2) When he performs the task or step incorrectly,

(3) When he introduces some task or step which should not have been performed,

(4) When he performs some task or step out of sequence, or

(5) When he fails to perform the task or step within the allotted time.

The above definition is quite independent of any consequences of human behavior. In a data bank, the consequences of human error should be considered to be system specific, and failure to perform a task may or may not lead to a specific consequence. Thus, the data bank should concern itself only with the probability of error. Of course, an operator's knowledge of the consequences of his actions is one of the independent variables that influences the error rate and should be recorded.

In order not to dilute the data collection effort necessary for a data bank, the criterion measure should be restricted to HER alone. There are other possible criterion variables, but they have not proved to be highly useful in applied human reliability work.

Task Taxonomy

In my first review of human reliability (Swain, 1964b) the following was stated: "Even before the relatively recent emphasis on quantification of human performance, it was apparent to many psychologists that a task taxonomy was needed. The need is even greater now. A whole system of task nomenclature is needed, nomenclature sufficiently general to apply to all man-machine system tasks and yet sufficiently specific (in its subclasses of tasks) to enable the user of the data bank to obtain information on human performance in any context of use for any man-machine system. This is a very large order."

It now appears the cart was placed before the horse. When a human reliability analyst looks at a list (always it's a draft list) of behaviors in a task taxonomy it is easy to get discouraged. If he were restricted to any existing list (e.g., Chambers, 1969), he would not usually be able to describe all of the independent variables which were really important for the HER in question.

Therefore, one can despair at any attempt to develop a detailed task taxonomy in advance of years of data collection. It now appears that the only workable task taxonomy will come out of (that is, be induced from) the use of the reams of data that can be put into a data bank. Over a period of time those of us in practical human reliability work will learn what types of data are useful--what gives us the best predictions. This experience must be used to massage the data bank, that is, to attach weights to the independent variables in the bank on the basis of what combination of data and weights best describes the results obtained in many field situations.

A task taxonomy derived from the human error rate data bank will have to be useful because the task descriptions in it will have been related to the criterion measure--to measure error rates in systems. These comments are not meant to be disparaging of the task taxonomy studies that have been done and the important research now being done in this field by the American Institutes for Research (Fleishman et al, 1968, 1970). It is merely suggested that the job vis-a-vis applicability to human error rate classification and prediction is far too complex for the primarily deductive work that has been done and that the emphasis now should be placed on a more inductive approach as suggested in this paper.

Computerization of the Data Bank

Concern with the computerization of the data bank is appropriate, but efforts in this area should be secondary to developing a data bank which is useful. The computer must not become the tail that wags the dog, as has happened in other endeavors. Let's first develop a workable and useful data collection and description scheme without considering any computer limitations. Premature attempts to computerize human performance data tend to oversimplify the varieties of human performance, to introduce clerical errors into the data, and to misconstrue the intent of various researchers by forcing their data into a restricted format.

Too much cannot be said about the efficiency of computers in handling routine operations. But there is nothing routine, at this stage, about human performance data. Only a dedicated specialist can hope to make sense out of what is available, and the interpretations that he must make at every step in data categorization and reduction will permanently color the character of both the bank and its contents. Until we are much better prepared, I feel that any human error data bank should be handled manually and only by human factors specialists who have long experience with the human error literature and the basic psychological methods underlying that literature.

Major Problems

The following six major problems are discussed:

(1) what data to collect;
(2) how to collect the data;
(3) how to store the data for use by planners and designers;
(4) how to use the data early in the developmental program and later on when validation work has been completed;
(5) how to optimize predictability of human performance influences on systems when using the data bank; and
(6) who should solve the above problems.

What Data to Collect

Types of Tasks

For the data bank to have maximum representativeness and therefore usefulness in applied work, emphasis should be on obtaining data from real-life situations whenever possible. The real-life variables, especially the people variables of training, experience, motivation, and stress, are not well simulated in the laboratory situation. The performance of the college sophomore, the typical laboratory subject, and the usual artificiality and restrictiveness of the laboratory tasks employed make for real problems of generalizability to a military field situation, particularly when dealing with error rates as small as 0.001 or even 0.00001.

Although emphasis should be on real-life tasks, it should also be worthwhile to collect data on tasks with less representativeness to real-life jobs. Even for laboratory tasks, there is always the possibility of calibrating the data (Swain, 1967a) so that it can be of use in predicting real-life behavior. Moreover, laboratory tasks may offer the only hope for detailed investigation of some types of independent variables, especially those involving stress. (There are notable exceptions, e.g., Berkun, 1964, but the preceding statement is generally valid.)

The following types of tasks offer possibilities for collecting worthwhile data for the data bank:

(1) real-life, routine operations, e.g., preventive and corrective maintenance, production, field retrofit, field use;
(2) real-life special operations, e.g., field tests, unscheduled exercises (e.g., SAC Standboard), fleet maneuvers;
(3) simulator trials, e.g., exercises in operational flight trainers; and
(4) laboratory studies.

Dependent Variables

The human error rate (HER) should be the primary criterion or dependent measure to be collected. The denominator required to produce a rate causes problems in some data collection schemes, especially those managed by quality assurance personnel. Too often quality assurance measures deal with the number of defects per unit of time or per assembly produced. Rigby (1967) has pointed out the need for data more amenable to analysis and prediction work as follows: "In any assembly task, for instance, it is not sufficient merely to record the number of soldering errors per number of units produced. In order to be fully meaningful, the data must show the number of soldering points per unit, at least. It is also helpful to show any differences among the soldering points that might make a difference in either frequency or type of error. For instance, were all wires inserted through holes and soldered, or were some looped, wrapped, or pigtailed?"

Thus, the definition of what constitutes an error is very important for the data store. The definition should include time, quality, and quantity considerations where they apply. For example, if a human action is not done in time in a paced task, the time requirements should be stated as part of the definition of the error. If a cold solder joint is considered a human error, what criteria did the inspectors use to identify a cold solder joint? If unsatisfactory performance is defined as not being able to track X number of aircraft simultaneously on a search scope, this information should be part of the definition of error.

When time, quality, or quantity define an error, these facts should be recorded. However, it does not appear to be desirable to saddle the data collection system with routinely collecting task completion time, task quality, or task quantity. With regard to maintenance time, for example, Bond (1970) has recently stated, "...efficiency, in the sense of minimum number of checks or minimum number of minutes, is not an extremely important criterion (to the technician). What the technician apparently wants to do is to find the trouble within a reasonable time." For many other types of tasks the time required to complete a task is not generally important or at least there is a wide latitude. Therefore, it is suggested that task completion time be recorded and entered into the data bank only when the time variable is likely to be considered an inherent part of the task as a performance shaping factor.

The same type of argument can be made against routinely recording quality and quantity output data.

An important advantage of the above restrictions (aside from simplifying the data collection and management) is that if an error is defined as the probability of not achieving some goal (whether time, quality, or quantity dependent or not), the probability equations are far simpler than if these three output factors are considered as variables. And, based on our eight years' experience in human reliability prediction studies, it appears that little value would be lost in following this approach.

In recording HER, the single most important statistic is the mean error rate. At this stage of human reliability prediction work, a single-point estimate of the error rate is all that is generally used, and this single point is nearly always the mean. However, for data collection for the data bank, the following additional statistics should be recorded whenever possible: median, upper and lower range, standard deviation, and shape of the curve. If feasible, the entire set of raw data should be recorded. Finally, the number of opportunities and how this number was defined should be entered into the data bank.

Independent Variables

Table 1 (from Swain, 1967a) shows general categories of factors that shape performance in a man-machine system. The problem for a data bank is to quantify all the individual performance shaping factors (PSF) that appeared to materially influence the HER recorded. This problem is not a small one. Consider just the people variables. Rigby (1970)

Table 1. Representative Categories of Performance Shaping Factors

Extra-Individual		Intra-Individual
(1) Instructions . Procedures required . Verbal or written communcations (2) Task-Equipment Characteristics . Perceptual discrimination required . Motor discrimination required . Complexity (information load) . Repetitiveness . Continuity (discrete vs. continuous) . Feedback (knowledge of results) . Human engineering factors Re: Man-machine interface design (prime eq., tools, fixtures, job aids) (3) Situational Characteristics . Equipment reliability status . Weather factors visibility/turbulence . Enemy actions . Friendly actions Commander & Team Partner Ac.ivities . System organization . Sys.em social structure	(4) Psychological Stresses . Task speed . Task load . High jeppardy risk . Threat of failure . Monotonous work . Long, uneventful vigilance periods . Reinforcement absent or negative . Sensory deprivation . Noise/distracting signals (5) Physiological Stresses . Fatigue . Pain or discomfort . Hunger or thirst . Temperature extremes . G force extremes Weightlessness High G . Atmospheric pressure extremes . Oxygen Insufficiency . Vibration . Movement constriction . Lack of physical exercise	(6) Individual (Organismic) Factors . Previous training/ experience . State of current practice or skill . Personality and intelligence variables . Motivation . Knowledge of required performance standards . Physical condition

states, in discussing the nature of human error:

> "People differ in more ways than any other class of objects in the known universe. The following estimates of the total number of measurable human dimensions are representative, and probably conservative:

100	Mental Abilities (e.g., IQ, inductive logic)
200	Personality Traits (e.g., masculinity, need for achievement)
500	Perception Indexes (e.g., visual acuity, reaction times)
1000	Anthropometic Dimensions, (e.g., height, weight, arm length)
5000	Physiological Dimensions (e.g., blood types, heart rates)
10,000	Sociological Dimensions (e.g., age, language, education)
200,000	Occupational Titles (areas of experience and competence)
1,000,000	Proficiency Indexes (specific knowledges and skills)

> The above measurable dimensions overlap and are inter-related in a very complex fashion, but they are relatively independent. People in any one occupation usually vary randomly across nearly the full range of all dimensions except those few on which they were selected; and on those few they usually cover more than half the full range."

To refer back to Table 1, our problem is that we normally cannot know in advance what PSF have material influences on HER. Therefore, in collecting data for the data bank we must initially err on the conservative side and depend upon after-the-fact analyses (postdiction studies) to see what PSF and weightings account for most of the variability in HER.

As the first step in developing the human error rate data bank, it is suggested that the responsible agency develop a master list of categories of PSF. This master list would serve as a general guide to the data collection effort. The listing in Table 1 can serve as a starting point. Note that the suggestion is for a list of <u>categories</u> of PSF rather than a detailed listing of individual PSF such as that provided by Chambers (1969). It is judged that a list which is much more detailed than that shown in Table 1 would not likely be useful to the data collection effort--and, in fact, might even impede it by the implication that the listing was complete. To aid in the development of a master list of PSF categories, it is suggested that the responsible agency review the work currently being done by the American Institutes for Research on development of a taxonomy of human performance (Fleishman et al, 1968, 1970).

How to Collect and Report the Data

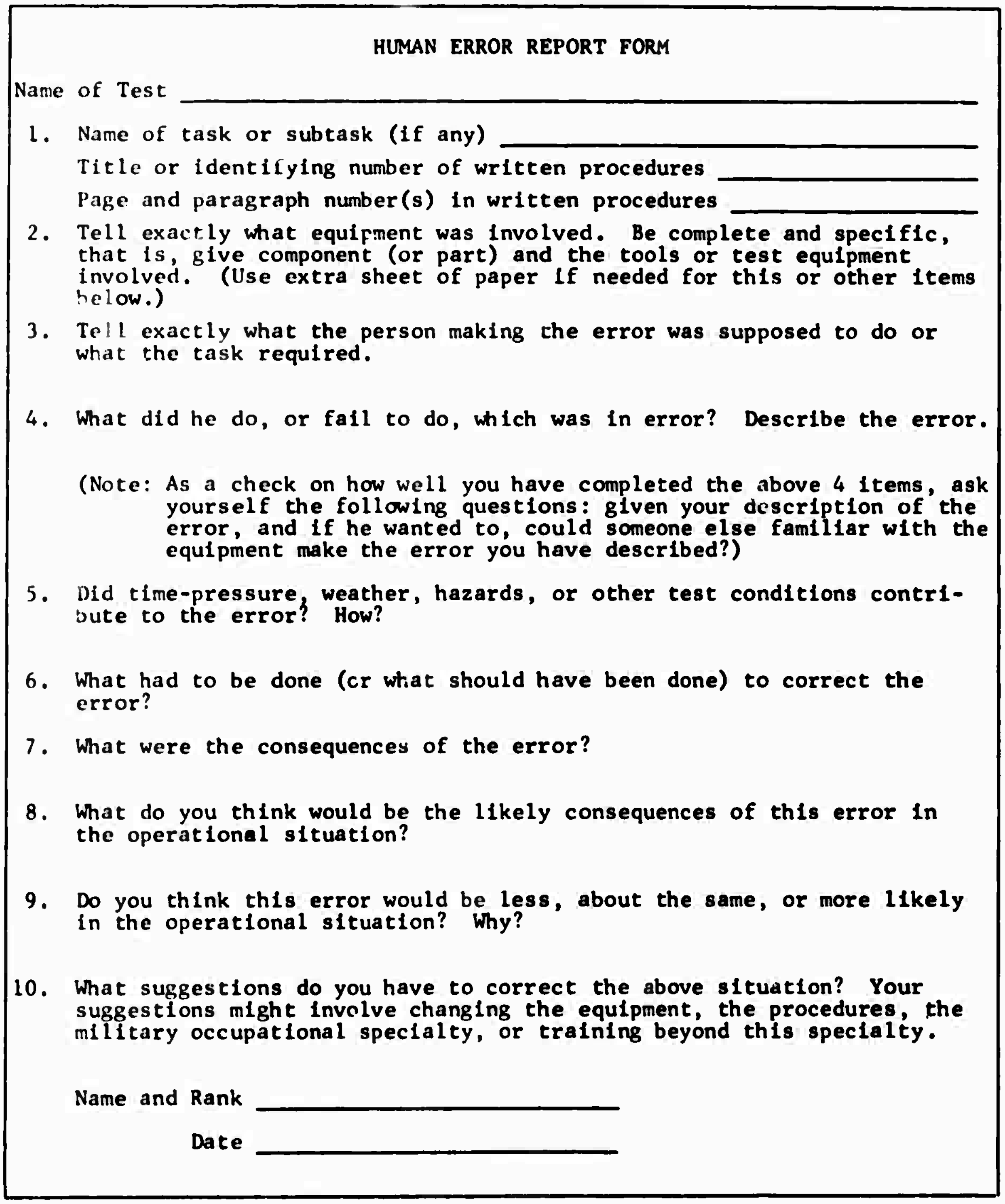
The Initial Data Collection and Reporting

Using the above master list as a guide, the responsible agency should develop an initial procedure for data collection and reporting for the human error rate data bank. Two forms would be required--one to describe the error and related details and one to fill in for entry into the interim data bank.

The initial data collecting and reporting should include all HER data (and related factors) collected during the data bank developmental program, previously published, or in individual company files. Each company participating in the program would fill in the data bank forms with data from its own human error rate studies. The responsible agency (or its designee) would check other publications for human error rate data to be entered into the interim data bank. Some surveys have already been made and the results should be obtained for analysis. For example, Sandia Laboratories made a survey of over 5,000 human performance articles in the hopes of adding to its HER data bank. Of these articles only a handful contained useful data for human error rate prediction. The results of this survey can be made available to the responsible agency.

Data Recording Form

The form used to describe the error and associated factors should be one that can be used by persons untrained in human factors. Probably most initial error reporting is done by such persons, e.g., inspectors, Inspector General personnel, and so on. This form should be relatively unstructured. An example of a human error recording form is found in Figure 1. This form was prepared for Army troop field evaluation of



HUMAN ERROR REPORT FORM

Name of Test ______________________________

1. Name of task or subtask (if any) ____________________

   Title or identifying number of written procedures ____________

   Page and paragraph number(s) in written procedures ____________

2. Tell exactly what equipment was involved. Be complete and specific, that is, give component (or part) and the tools or test equipment involved. (Use extra sheet of paper if needed for this or other items below.)

3. Tell exactly what the person making the error was supposed to do or what the task required.

4. What did he do, or fail to do, which was in error? Describe the error.

   (Note: As a check on how well you have completed the above 4 items, ask yourself the following questions: given your description of the error, and if he wanted to, could someone else familiar with the equipment make the error you have described?)

5. Did time-pressure, weather, hazards, or other test conditions contribute to the error? How?

6. What had to be done (cr what should have been done) to correct the error?

7. What were the consequences of the error?

8. What do you think would be the likely consequences of this error in the operational situation?

9. Do you think this error would be less, about the same, or more likely in the operational situation? Why?

10. What suggestions do you have to correct the above situation? Your suggestions might involve changing the equipment, the procedures, the military occupational specialty, or training beyond this specialty.

Name and Rank ____________________

Date ____________________

Figure 1. Human Error Report Form

electronic equipment (Swain, 1964c) and was intended for use by enlisted personnel with no training in human factors. The form is still in use. (In practice two sheets of paper are used to allow room for writing.)

The most important thing is that the data collection instructions and data form for use by persons untrained in human factors should not be restricted to a selected number of categories. Otherwise, little useful information may be obtained. For example, in an unpublished study by Rigby on defect reporting at one industrial plant, the form previously used by inspectors had 41 coded categories. The inspectors were supposed to check the categories that applied to any defect they reported. But the inspectors, being human, did not commit the 41 definitions to memory nor did they carry the papers with them that defined each category. They did not use 9 of the categories even once, 7 other categories were used very infrequently, and another 7 categories accounted for 77 percent of the defects. The basic problem was that the defect categories were structured for the convenience of the data analyzers without much thought as to whether the 41 categories would be used by the inspectors. It was simply assumed they would.

Our experience is that the only kind of defect description scheme that will work (that is, that will be useful in identifying causes and patterns of defects and in enabling error rates to be determined) is one that allows the inspectors to describe a human initiated defect in their own language such that someone else familiar with the production process *could repeat the defect if he were told to do so*. This approach means that an extra analysis step is needed before data of this kind can be summarized and encoded. And this is as it should be. It allows the inspector to do what he can do reasonably well--namely, describe what happened. And it allows the data analyzer to do what he should be capable of doing--namely, summarize and encode data. It is false economy to think that the data recorder can perform both functions.

It is not proposed that the data recording form shown in Figure 1 be used for the human error rate data bank. But it does provide an example of the kind of relatively unstructured form we feel is appropriate to the initial data gathering process.

In addition to this unstructured form (or as part of it) recording space is needed for certain physical measures which pertain to the

HER in question. Environmental factors such as temperatu e, humidity, windchill, noise, etc., can be measured and such values should be entered into the data bank form. Any other physical measures pertaining to PSF levels should be recorded.

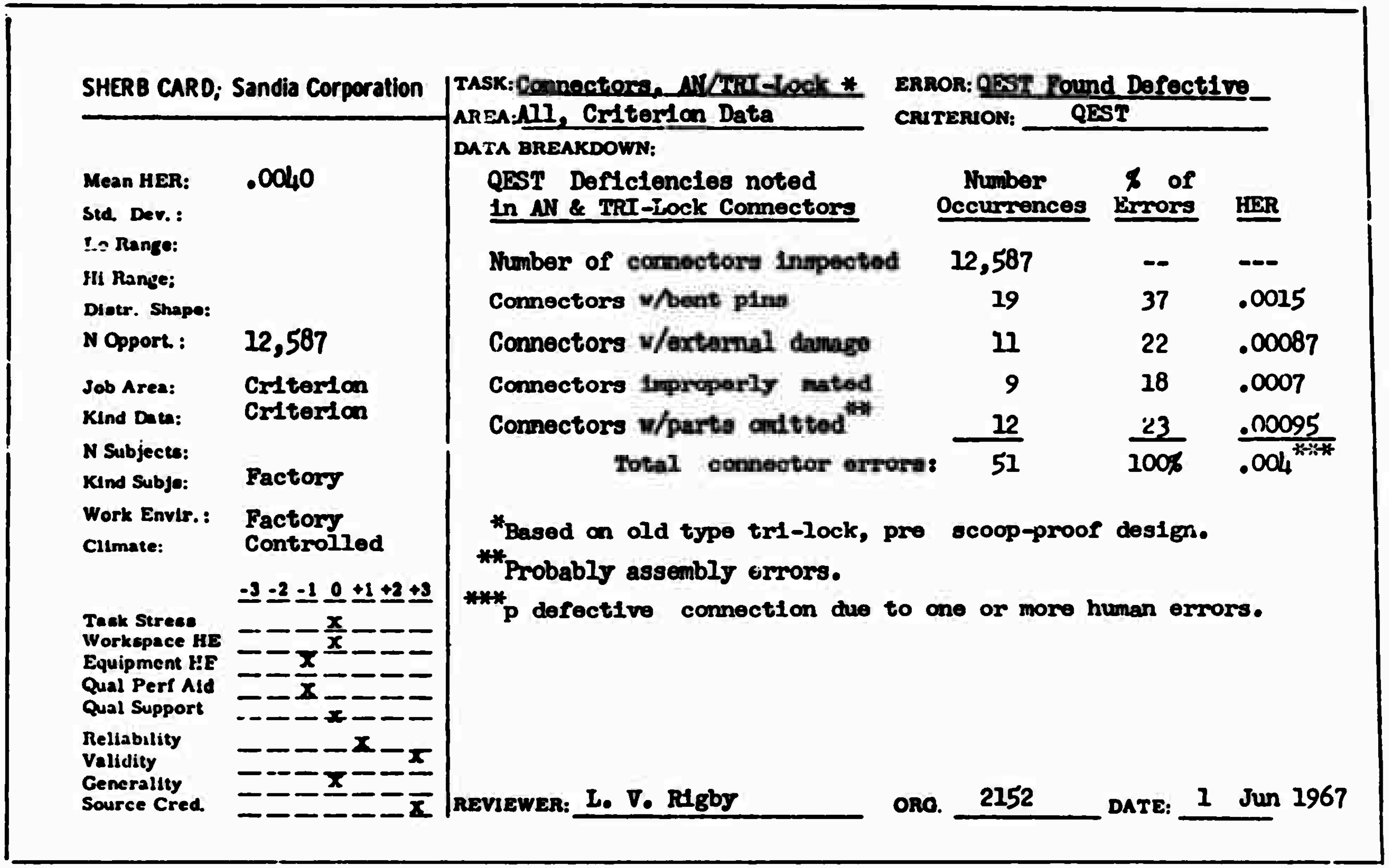
Data Bank Form

The above type of data recording form would not provide sufficient information for the human error rate data bank. A special form, the data bank form, is needed for the analyzed and encoded raw information from the data recording form. The data bank form would have on it the equipment and task involved plus the dependent and independent variables as described earlier. Detailing of this information requires judgments based on skill and experience in human behavior technology and cannot be done by clerical personnel. The data analyzer must also be familiar with the task in question.

At the 1962 meeting of the Human Factors Society, Sandia Laboratories offered to collect and disseminate HER data (Swain, et al, 1963). There were no takers, but we developed the Sandia Human Error Rate Bank--SHERB (Rigby, 1967). Although we have made a start, personnel with the necessary skill to manage such a file have been too busy doing human reliability and human engineering work of higher priority. However, the SHERB card (Figures 2 and 3) contains the type of information we felt was important in being able to use the HER data to generalize to other tasks for which human error rate prediction was desired.

Rigby's 1967 report describes how to use the SHERB card, so that description will not be repeated here. But even with our limited experience with the SHERB cards we now recognize some limitations and need for changes which have some bearing on a data bank form for the human error rate data bank.

First, there obviously is not sufficient room to write down all the relevant information connoted by the various categories on the card. Thus, we tend to go back to the original data record (or some other report) to see if the data on a card should really be applied to the prediction task at hand or to judge what correction factor, if any, should be applied. This difficulty plus the inconvenience of trying to reproduce extra copies of the 5 x 8 inch card leads us to recommend the use of standard size (8-1/2 x 11 inch) sheets of paper.



SHERB CARD; Sandia Corporation

Mean HER:	.0040
Std. Dev.:	
Lo Range:	
Hi Range;	
Distr. Shape:	
N Opport.:	12,587
Job Area:	Criterion
Kind Data:	Criterion
N Subjects:	
Kind Subjs:	Factory
Work Envir.:	Factory
Climate:	Controlled

	-3	-2	-1	0	+1	+2	+3
Task Stress				X			
Workspace HE				X			
Equipment HF			X				
Qual Perf Aid			X				
Qual Support				X			
Reliability					X		
Validity							X
Generality				X			
Source Cred.							X

TASK: Connectors, AN/TRI-Lock * ERROR: QEST Found Defective

AREA: All, Criterion Data CRITERION: QEST

DATA BREAKDOWN:

QEST Deficiencies noted in AN & TRI-Lock Connectors	Number Occurrences	% of Errors	HER
Number of connectors inspected	12,587	--	---
Connectors w/bent pins	19	37	.0015
Connectors w/external damage	11	22	.00087
Connectors improperly mated	9	18	.0007
Connectors w/parts omitted**	12	23	.00095
Total connector errors:	51	100%	.004***

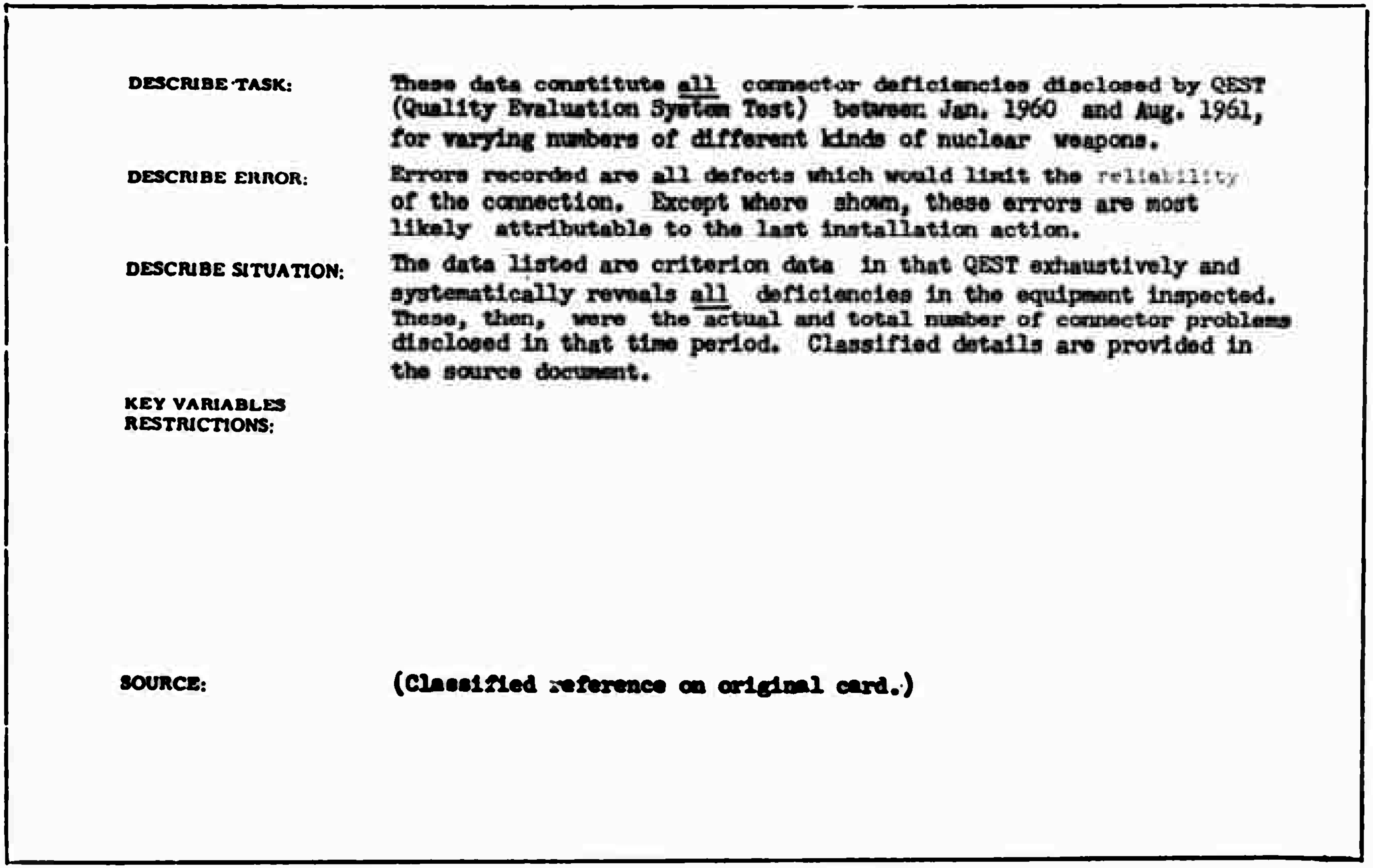
*Based on old type tri-lock, pre scoop-proof design.

**Probably assembly errors.

***p defective connection due to one or more human errors.

REVIEWER: L. V. Rigby ORG. 2152 DATE: 1 Jun 1967

Figure 2. The Front Side of a Typical SHERB Card



DESCRIBE TASK:	These data constitute all connector deficiencies disclosed by QEST (Quality Evaluation System Test) between Jan. 1960 and Aug. 1961, for varying numbers of different kinds of nuclear weapons.
DESCRIBE ERROR:	Errors recorded are all defects which would limit the reliability of the connection. Except where shown, these errors are most likely attributable to the last installation action.
DESCRIBE SITUATION:	The data listed are criterion data in that QEST exhaustively and systematically reveals all deficiencies in the equipment inspected. These, then, were the actual and total number of connector problems disclosed in that time period. Classified details are provided in the source document.
KEY VARIABLES RESTRICTIONS:	
SOURCE:	(Classified reference on original card.)

Figure 3. The Back Side of a Typical SHERB Card

Moreover, one shouldn't feel obligated to restrict oneself to one sheet of paper when reporting a HER and the associated PSF. There must be sufficient room to report sufficient detail. One guiding principle we use is to describe each PSF in the same level of detail that one would use in describing the experimental design and related procedures for a Ph.D. dissertation.

There also needs to be room for the data analyzer to report his own subjective judgments as to what, in his opinion, may have been an extremely important contribution to the error rate. To cite one obvious example, an unacceptably high inspector error rate was discovered in one sample of assembled items. Records were able to point to one inspector who, it was later determined, had had unusually severe emotional stress related to a lingering illness and subsequent death of his wife. In our opinion, this one PSF was the overriding one, and that inspection HER data bank form would so indicate so that we wouldn't later ascribe the HER to other factors.

Our experience, then, is that the data bank form should be relatively unstructured, rather like the SHERB card. Thus, it does not appear desirable to use a detailed PSF listing which the data analyzer uses to check off the relevant categories. Rather the PSF categories listing should be available as a guide, and the data encoder instructed to describe on the form those PSF under each general category which appear to have material influence on the HER reported. By material influence is meant a PSF level other than zero (as described later).

Another difficulty, a minor oversight, was that we have a place to describe the task, but we did not specifically call out the type of equipment involved. It turns out that in practical human reliability or human engineering work, the most usual search term is type of equipment--whether prime equipment (end item), tool, work fixture, handling equipment, or other type of equipment. The level of description of the equipment features should be as detailed as that found in the AIR Data Store (Munger et al, 1962).

In addition to calling out the task and task equipment, the data bank form should have entries for "preceding tasks" and "following tasks." The data analyzer should enter task information in these entries which is directed at identifying task-task interaction effects. (Or it may be appropriate to identify this interaction at the step level.) The important thing is that the data analyzer set down pre- or post-task activities which he believes contributed materially to the error rate observed. "Materially" can be defined as a nonzero PSF level as described below.

Another difficulty we experienced in using the SHERB card was the attempt to use a 7-point sigma scale to describe levels of a PSF. We felt it was important to indicate in some fashion how much of each relevant PSF was involved in the HER reported. The values in the scale indicated the followings ranges:

-3 = equal to or worse than -3σ (i.e., 99.9%, or 999/1,000, are better than -3)
-2 = -2σ up to -3σ (i.e., 98%, or 49/50, are better than -2)
-1 = -1σ up to -2σ (i.e., 86%, or 6/7, are better than -1)
0 = $\pm 1\sigma$ (i.e., the average or middle 2/3's)
+1 = $+1\sigma$ up to $+2\sigma$ (i.e., 86% or 6/7, are worse than +1)
+2 = $+2\sigma$ up to $+3\sigma$ (i.e., 98% or 49/50, are worse than +2)
+3 = equal to or better than $+3\sigma$ (i.e., 99.9%, or 999/1000, are worse than +3)

For example, in describing task stress, the 7-point scale can run from -3 through 0 to +3, each number representing approximately one standard deviation value. We still think this approach is a good one, but it is clearly obvious now that for each PSF to which this scale is to be applied sample definitions of the sigma values should be included to guide the data compiler. Thus, it is suggested that in the initial data collection effort, the data analyzer who uses a 7-point scale to describe a level of a particular PSF should also include his definitions of each of the seven points. This type of information would help the revisions of scaling of PSF levels to be done in subsequent phases of the program.

The use of the 7-point sigma scale (or some other scaling) is recommended for the initial data collection for the human error rate data bank because it is a convenient and workable method of differentiating at least grossly among levels of a PSF and can be used for an interim data bank. The method is convenient because it involves the use of psychological scaling, a method familiar to most human factors personnel, or at least readily learned by them. The method is workable because it provides a ready method of quantifying the subjective judgments made as to PSF level and one which should enable various human factors personnel to use each other's human error rate data. As noted by Rigby (1967), "The use of this kind of scale is not intended to imply greater accuracy in rating; rather, it simply forces us to think in terms of a normal distribution of events. The great majority of events are 'more or less average,' and they receive the middle, or zero, rating. This kind of rating scale seems to be more useful and more appropriate to probability analysis than a linear scale."

Operationally, we tend to think of the 7-point sigma values as follows:

-3 = Difficult to think of a worse level of this PSF in a practical situation.

-2 = The value of this PSF is highly detrimental for human reliability.

-1 = The PSF level is sufficiently below average that the HER would likely be materially increased.

0 = Probably as long as the PSF is within this range, it should not materially affect the HER; other factors would be much more important.

+1 = The PSF level is sufficiently above average that the HER would likely be materially reduced.

+2 = The value of this PSF is highly beneficial for human reliability.

+3 = Difficult to think of a better level of this PSF in a practical situation.

We feel that the same type of 7-point sigma scale should be used to describe the data analyzer's impression of task difficulty. This type of subjective judgment could have real predictive value. We must remember that although we are interested in predicting absolute error rates we may well find that quantified subjective judgments of the PSF will turn out to have statistically significant predictive weightings.

By way of illustrating the use of the 7-point sigma scale, the following paragraphs describe two PSFs which we have found to be so important in determining the HER for a task that we feel they should be added as separate categories on a data bank form. These categories are personnel redundancy and manner of use of performance aids. While our most recent experience is primarily in the production area, military reports suggest that these categories are also important in operations areas.

By personnel redundancy we mean the extent to which one or more checks are made on the accuracy of a task and the amount of relative independence of these checks. A 7-point sigma scale for personnel redundancy is as follows:

-3 = One person performs the task and does not check his work.

-2 = One person performs the task and unsystematically rechecks some of his work by looking over it

-1 = One person performs the task and systematically checks his work but without the independence of redoing it (i.e., if he made cable connections, he looks for tightness; if he did arithmetic calculations, he goes over his work visually).

0 = One person performs the task and checks his work by a method that provides more independence than rank -1 above (i.e., if he made cable connections, he feels for proper tightness; if he did arithmetic calculations, he repeats the calculations on a different sheet of paper and then checks the second answers with the first).

+1 = One person does the work and another watches him (the "over-the-shoulder" method of monitoring or inspection)

+2 = One person does the work and then steps away from the task and another comes up and inspects the output of the first but without repeating the work (i.e., analogous to rank -1 but using a second person).

+3 = One person does the work, steps away from the task, and another performs a check with the independence of rank 0 above.

It is clear that in reporting the human error rates, it must be recorded whether the HER is for one man, or for a man and one or more monitors. This information is not always clear in defect reporting, but it must be clear in reporting HER. In some cases it is not possible to partial out human errors between operator and inspector. In such cases, the HER must be ascribed to operator plus inspector.

The manner of use of performance aids has been demonstrated to have an important influence on HER. The SHERB card has a 7-point sigma scale for the quality of performance aids, but no category for manner of use. For written instructions, the most common form of performance aids, the following 7-point sigma scale could be used:

- -3 = Written instructions not used
- -2 = Written instructions used only occasionally, e.g., when supervision is around or when an audit is being made.
- -1 = Written instructions used frequently but not every step is read by the operator (i.e., he picks and chooses).
- 0 = Operator reads written instructions to himself but without doing each step right after he reads it.
- +1 = Operator reads written instructions aloud (especially if working with someone) and does each step after he reads it.
- +2 = An oral reader reads off each step and observes when other operators complete step (or is advised by telephone).
- +3 = Same as rank +2 except that the oral reader checks off each step as it is completed.

The above two examples of 7-point sigma scales illustrate the kind of aided judgments we feel can be valuable to human reliability work if made by persons skilled in human behavior technology and with practical experience in human factors.

In recording the HER, the data analyzers should record on the data bank form whatever significant digits are provided by the data and leave any rounding off to those who will use the data once it becomes available to data bank users. HERs should be listed as decimals, for example, as .0021 rather than 21×10^{-4} or to some standard base such as 10^{-6}. Decimals are more easily grasped and more commonly understood at least up to five or six decimal places. (It is not necessary to write 0.0021 since in dealing with probabilities, the numbers are always 1.0 or less.)

There is one final classification scheme which should be considered for the human error rate data bank form. With some minor additions, this classification scheme is the one developed by Rook (1962) in attempt to classify HER into categories that would suggest a means for correction of error-likely design situations. Such a classification scheme should have real utility for those who use the human error rate data bank.

The classification scheme employs a cross-cutting system of error classification using four types of error versus the input, mediating, and output components described in the AIR Data Store (Payne and Altman, 1962, and Munger et al, 1962). The types of error are:

Type A - Intentional Without Malevolence (i.e., the operator thought he was doing the right thing but was following an incorrect procedure whether written or not. As Rook says, "the operator intends to perform the act correctly, but erroneously performs it out of limits")

Type B - Unintentional (i.e., "there is no element of intent in the performance of the act; it just happens." (Rook, 1962). Example: operator accidently burns a hole in wire insulation when soldering)

Type C - Omission Without Malevolence (i.e., "While such acts are intentional, and thus logically form a subclass of Type B, they are influenced by factors different from those relevant to Type B errors" (Rook, 1962))

Type D - Malevolent (i.e., error done with intent to inflict injury, damage, or other undesirable consequences)

Using the above definitions plus the three components from the AIR Data Store, we arrive at the system of error categories presented in Table 2. Rook (1962) discusses how this classification scheme can be used. He notes that the system of classification "is not proposed as a means of explaining the cause of human error, but merely as a means of subdividing the phenomenon of human error into categories which are both manageable and suggestive of corrective action to be taken."

Table 2. System of Human Error Categories

Errors due to acts which are	Errors due to behavior components of		
	Input (I)	Mediation (M)	Output (O)
A - Intentionally performed without malevolence	AI	AM	AO
B - Unintentionally performed	BI	BM	BO
C - Omitted without malevolence	CI	CM	CO
D - Performed with malevolence intended	DI	DM	DO

Interim and Final Data Collection and Recording

On the basis of experience with the initial data collecting and recording, and with the help of those participating in the developmental program, the responsible agency should be able to revise the data collection and recording scheme and make appropriate revisions to the interim data bank.

Inputs should be obtained from those who use the interim data bank for human reliability and human engineering work and from those who supply filled in data bank forms to the data bank. (Hopefully, there would be a great deal of overlap between these persons.) Experience with the 7-point sigma scale (or equivalent) should suggest desired improvements in scaling the PSF and defining (with examples) the various PSF levels.

For various PSF, rather than depend on the relatively subjective 7-point sigma scaling, it may be desirable to develop scale values using more objective psychological scaling described in Swain (1967a) and Rigby and Edelman (1968). For any one PSF, various levels (as obtained from the data bank) can be pair compared by experts (i.e., persons familiar with the program and doing applied human factors work). (Shortcuts of the lengthy paired comparison methods are available--see Guilford, 1954, p. 168ff., and Torgerson, 1958, p. 191ff. Blanchard et al (1966) contains some recent application of paired comparison techniques to the human reliability field.) The resultant ranking can be converted to an interval scale using techniques described in the referenced texts above and in Edwards, 1957, p. 31ff. The scaling of emergencies described in Rigby and Edelman (1968) is a case in point.

The final data collection and recording methods and forms would be designed after the validation study described later and would include empirically derived weights for PSF levels.

How to Store and Use the Data Bank

The human error rate data bank must have easily accessible data in order to be useful to designers and planners as well as the human factors specialists doing human reliability work. The suggested interim data bank is different from the final data bank, so these banks are discussed separately.

Interim Data Bank

The interim data bank will include the completed data bank forms and possibly also the completed data recording forms. It may well be useful to include both completed forms since one represents a summary and analysis of the other. But for discussion purposes, just the data bank form will be referred to.

At least for the interim data bank, it is suggested that the most economical method (and quickest way) of setting up a useful data bank would be to set up a publishing vehicle for the completed data bank forms. Thus the responsible agency would publish the received data bank forms on some periodic basis and distribute them to those organizations participating in the program. The name "Journal of Human Error Rate Data" might be appropriate for this continuing series.

Each using agency would be responsible for its own filing scheme using whatever search terms they feel are appropriate. It is suggested, however, that the responsible agency develop a list of suggested search terms for each completed data form they publish in the Journal. On the basis of our experience, we would suggest equipment type (e.g., the name of the particular prime equipment, tool, work fixture, handling equipment, etc.), task (e.g., connecting), and PSFs which receive minus ratings. Thus, if the user is interested in errors in using screwdrivers, he might look up the data forms filed under the equipment type category "screwdriver." (It might be determined that such a specific category would be too fine a cut, and some more general term such as "fastening tool" might be more useful, at least for the interim data bank.) If the user is interested in screwdrivers used in making cable connections, he might also look up the data forms filed under the task "connection." If he is interested in effects of cold temperature on screwdriver use, he might look up the data forms filed under the PSFs having negative ratings for "temperature." For the interim data bank, however, it would appear to be presumptious to attempt a cross filing scheme such that a man interested in screwdrivers used to fasten cables together under cold temperature extremes would have only to look under one category in the file. Such refinements could be considered for the final data bank, but the limited usefulness of such a cross filing scheme might not be worth the added cost and complexity.

For the user, the 7-point sigma scaling of PSF levels would be helpful since he would know that someone trained in human factors technology and familiar with the task in question had assigned the sigma ratings for the PSF. Thus, if he saw a plus or minus rating for a certain PSF when he is looking at a completed data form he would be alerted to be careful about generalizing that particular HER to the task for which he wants to predict a human error rate unless the same PSF rating applied to the new situation.

As the responsible agency receives more and more completed data bank forms and comments from the data analyzers responsible for the forms, appropriate revisions would be made to the human error rate data bank, including addition or deletion of search terms. With appropriate feedback supplied from the users of the data bank, a most useful data bank could result.

Final Data Bank

After the validation part of the program described later, the final version of the data bank could either be computerized or noncomputerized. It should not be just assumed that computerization of the data bank would be a worthwhile undertaking. It might well be that the extra benefits of computerization (assuming these could be demonstrated and not just taken for granted) would not be worth the extra cost and complexity. We feel that the kind of manual system described above could handle our needs.

In any event, the final data bank would be one that had benefitted from the experience in using the interim bank. The search terms would be ones that had proven to be useful. The PSF and levels would be empirically derived--from postdiction exercises as well as exercises in prediction (as discussed in the next section). It would now be possible to use the data collected to induce a task taxonomy that would have practical value for human error rate prediction.

Validation and Management of the Data Bank

After the human error rate data bank has been in use for some time and there are many entries, it should be worthwhile to see which PSF and levels of PSF are related to the HERs for various equipment/ task combinations. Levels of PSF would include any physical measurements plus the weights determined by use of the 7-point sigma scale or by more refined psychological scaling described earlier. The point of this effort would be to see which combinations of PSF and PSF levels correlate highest with the HERs for various equipment/task combinations.

It is not a foregone conclusion that it will be possible to derive equations which have great predictive power. It may simply be impossible to develop a sufficiently small set of predictive equations to be manageable. Nevertheless, the effort will not have been wasted; the bank of filled in data forms, revised to include empirically derived weights, would still be a major step forward in human reliability work. And it would be possible to develop/induce a final version of a task taxonomy which would have maximum usefulness in human factors work.

There will be a need for continuing management of the human error rate data bank. Based on continued experience with the bank, further revisions and refinements of the data bank forms and related collection and recording procedures and techniques should be made.

Who Should Solve the Above Problems

Who should do the work involved in developing a human error rate data bank? The responsible agency, whether in-service or some research group on a contractual basis to the military sponsor, cannot develop the data bank without the active support and help of a large number of companies, the potential users of the data bank. But how can these companies be motivated to supply the information necessary for the data bank?

There are problems to be faced. One major problem is the reluctance of some companies to admit that their personnel make errors or at least to advertise the fact by disseminating quantitative data related to human errors. In our work, we have seen cooperative lower management at one industrial plant agree to allow us to analyze their defect data, only to be overruled by their upper management for fear of the harm that such information might do them if it got into the hands of their competitors. This fear is not unrealistic considering competitive bidding for military contracts.

A second major problem is that persons with inadequate skills are frequently used in data collection and analysis areas. But we cannot expect good (usable) data for the human error rate data bank unless highly knowledgeable and experienced persons are assigned to the data analysis work described earlier. Our predictive work will be no better than the quality of the data available. As noted earlier, since the data analyzer will have to make judgments requiring skills in human behavior technology, no one should be assigned to this task unless he has these skills.

A third major problem is that human performance data is often reported in such a way as to make it impossible to determine human error rates. The results of university laboratory studies, the kind performed by personnel in psychology departments, and which show up in publications such as the Journal of Experimental Psychology, often cannot be used for human reliability work because error rates are not reported. Defect reports collected by quality assurance organizations in industry often are not useful for identifying human errors and usually not conducive to determining rates of human errors (Rigby and Swain, 1968).

What can be done about these problems? The responsible agency can prepare documentation which will show the potential participators in the proposed program what benefits they can derive from active cooperation in establishing a human error rate data bank. It seems to be an accepted fact (if one can judge from presentations at symposia of the IEEE, ASQC, ASME, SAE, and other professional organizations) that the human element is responsible for a very large percentage of system and component failures. Certainly any data that would assist system planners and designers to reduce this adverse influence of the human element in systems and thus to include this factor in tradeoff considerations should be welcomed by all. These and other arguments can easily be found.

The above is the carrot. But a stick is needed. We believe that all military contracts for system development and production should have.

a requirement for HER collection. How to implement this requirement should be spelled out clearly and should reflect what is needed for the human error rate data bank. Money, time, and effort should be allocated for this purpose <u>and for no other</u>. Otherwise, the data bank effort would probably be dropped as overruns occur in a contract. The contract should also spell out the qualifications of the persons to do the data analyzing as described earlier. Otherwise, unqualified persons may be assigned to this important task. Finally, some safeguards must be established so that a company's reported HER data cannot be used to that company's disadvantage by its competitors. Perhaps it would be sufficient for the data bank form to omit the company's name.

Concluding Comments

It has been suggested that work begin very soon on an interim human error rate data bank which would be planned for manual retrieval of human reliability data. Mathematical techniques exist now which are compatible with conventional reliability technology. Thus, it seems inadvisable to spend early time and effort on further mathematical modeling. It has further been suggested that a hard look be taken on the worthwhileness of computerization of the data bank. Automatic high-speed data retrieval may not provide sufficient gains over a simpler manual data bank to warrant the extra complexity (and associated unreliability), time, money, and effort.

A workable data collection, analysis, and recording scheme has been suggested which does not depend on the development now of a behavior task taxonomy. The latter effort has been going on at least since the early 1950's and we seem to be very little nearer to realizing a task taxonomy which is useful for applied human reliability and human engineering work. However, the results of the proposed work should facilitate the later development of a useful task taxonomy for this purpose.

It has been asserted that the mere reporting of human error rate data with its associated performance shaping factors would be a major aid to human reliability and human engineering work and, further, that should this reporting be the only result of the proposed development of a human error rate data bank, the effort would have been worthwhile.

The specific steps proposed to develop the human error rate bank and suggested phasing of the work are summarized below:

Phase I - Development of Initial Procedures (4 months).

(1) Develop list of categories of performance shaping factors (PSF).

(2) Develop initial procedure for data collection which includes initial list of PSF categories and a subjective 7-point sigma scale for estimating PSF weights for each HER collected.

Phase II - Collection of Initial Data and Establishment of Interim Human Error Rate Data Bank (8 months).

(1) Using initial procedure, collect and analyze HER, operating times, and PSF weighted on 7-point sigma scale. The data should include unpublished and published data previously collected as well as that collected during Phase II. Have data collectors report their problems and successes.

(2) Set up interim data bank (including procedures for use) based on above results.

(3) Publish above procedures and, periodically, the HER collected in a Journal of Human Error Rate Data.

Phase III - Revision of 7-point sigma Scaling or Interval Scaling of Selected PSF and Revised Data Collection Procedure (6 months).

(1) Make revisions of 7-point sigma scale for PSF on basis of experience above.

(2) Where possible, have experts rank selected PSF (using paired comparison or related psychological scaling technique) and therefrom derive an interval scale of the PSF.

(3) On basis of the revisions and scaling, derive new PSF weights and enter them into interim data bank. Modify procedures for use accordingly.

(4) Develop revised procedure for data collecting, using new weights.

Phase IV - Data Collection and Revision of Data Bank (12 months).

(1) Using revised procedure, collect and analyze HER, operating times, and weighted PSF.

(2) Revise and (if worthwhile) computerize data bank to incorporate new data.

Phase V - Derive New Weights by Multivariate Analysis and Develop Final Data Collection and Data Bank Use Procedures (6 months)

(1) Conduct postdiction exercises to see what PSF and weights have maximum postdictive power for various types of equipment/task combinations. Change data bank accordingly.

(2) Develop final procedure for data collection and use of data bank.

(3) Construct task taxonomy based on data collected to date.

Phase VI - Continuing Management (Open ended)

(1) Conduct continuing management of data bank.

(2) Based on use of data bank, make revisions to data collection procedures and data bank and use procedures as appropriate.

References

Berkun, M. M., "Performance Decrement Under Psychological stress," Human Factors, 1964, 6, 21-30.

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

Bond, N. A., Jr., "Some Persistent Myths about Military Electronics Maintenance," Human Factors, 1970, 12, 241-252.

Chambers, A. N., Development of a Taxonomy of Human Performance: A Heuristic Model for the Development of Classification Systems, AIR-726-3/69-TR-4A, American Institutes for Research, Washington D. C., August 1969.

Edwards, A. L., Techniques of Attitude Scale Constructions, Appleton-Century-Crofts, Inc., New York, 1957.

Fleishman, E. A., Kinkade, R. G., and Chambers, A. N., Development of a Taxonomy of Human Performance: A Review of the First Year's Progress, AIR-726-11/68-TPR1, American Institutes for Research, Washington D. C., November 1968.

Fleishman, E. A., Teichner, W. H., and Stephenson, R. W., Development of a Taxonomy of Human Performance: A Review of the Second Year's Progress, AIR-726-1/70-TPR2, American Institutes for Research, Washington D. C., January 1970.

Guilford, J. P., Psychometric Methods, McGraw-Hill Book Co., Inc., New York, 1954.

Munger, S. J., Smith, R. W., and Payne, D., An Index of Electronic Equipment Operability: Data Store, AIR-C43-1/62-RP(1), American Institute for Research, Pittsburgh, Penn., January 1962.

Nicklass, D. R., Meister, D., Obermayer, R. W., and Leuba, H. R., "Preface," Human Factors, 1964, 6, 553-554.

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

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, Los Angeles, Calif., June 1967, 5-1 to 5-13. (Also SC-R-67-1150)*.

Rigby, L. V., "The Nature of Human Error," Annual Technical Conference Transactions of the ASQC, American Society for Quality Control, Milwaukee, Wisc., May 1970, 45 -466. (Also Preprint SC-DC-69-2062).

Rigby, L. V., and Edelman, D. A., "A Predictive Scale of Aircraft Emergencies," Human Factors, 1968, 10, 475-482, (Also SC-R-69-1208).

*Sandia Laboratories publication numbers in parenthesis.

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

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

Rook, L. W., *Motivation and Human Error*, SC-TM-65-135, Sandia Laboratories, Albuquerque, N. M., September 1965.

Swain, A. D., *A Method for Performing a Human Factors Reliability Analysis*, SCR-685, Sandia Laboratories, Albuquerque, N. M., August 1963.

Swain, A. D., *Human Factors Associated with Prescribed Action Links* (U), SCR-634, Sandia Laboratories, Albuquerque, N. M., May 1963, Conf.

Swain, A. D., *THERP*, SC-R-64-1338, Sandia Laboratories, Albuquerque, N. M., August 1964.

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).

Swain, A. D., *Volume 3: Human Factors Test and Evaluation*, Test Plans and Evaluation Dept., U. S. Army Electronic Proving Ground, Fort Huachuca, Arizona, November 1964.

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

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 Laboratories, Wright-Patterson AFB, Ohio, May 1967, 17-31. (Also SC-R-68-1697).

Swain, A. D., *Human Reliability Assessment in Nuclear Reactor Plants*, SC-R-69-1236, Sandia Laboratories, Albuquerque, N. M., April 1969.

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

Swain, A. D., "A Work Situation Approach to Improving Job Safety." *Proceedings, 1969 Professional Conference of the ASSE*, American Society of Safety Engineers, Chicago, Illinois, August 1969, 233-257 (Also SC-R-69-1320).

Swain, A. D., Altman, J. W., and Rook, L. W., *Human Error Quantification, A Symposium*, SCR-610, Sandia Laboratories, Albuquerque, N. M., Apr. 1963.

Torgerson, W. S., *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York, 1958.

DISCUSSION OF SWAIN'S PRESENTATION

Meister - The particular examples of the types of factors that you consider to be critical are some things which I think are terribly difficult, not only to measure, but also for a user to be able to make any sort of estimation.

Swain - But this is what I do. For example, I've had to predict how simultaneous could 10 operations take place amongst 10 different teams. I had to go to reaction time literature, I had to go to all kinds of different sources to try to get an idea, and I was continually trying to evaluate the applicability of all these disparate bits of data to my particular situation so that I could come up with an estimate of what this system was attempting to do. My goal, in using these scaled factors, is simply to make these kinds of judgements easier to make. They're still judgement. All I am asking for is a system that will help the applied man to do his work a little better. What I recommend later on then in this paper is that we actually have what I call the Journal of Human Error Rate Data. All the people who are doing work on developmental systems, weapons systems, radar systems and what not, often have an opportunity to collect performance data. I'm saying let's have a vehicle so where they can send in this data, and, furthermore, let's have the governmental agencies require that it be done.

Coburn - You would have so many different frames of reference! How would you organize all this material so it would be useful?

Swain - I've got some suggestions on that. I think that if we had all of the data put in some kind of a journal published periodically, it would be up to everybody to devise their own filing scheme. This is only for the interim situation. I recommend that the responsible agency do some necessary headscratching and so on and come up with a suggested search term. Equipment type, of course, is one that we use all the time in our work.

Mills - Isn't it true, that you may have to develop a rather large taxonomy. The question is whether or not such a large taxonomy would not be useful for developing a data base which would be computer stored.

Swain - We don't have a taxonomy that's very useful. We've been going since 1952 and this is 1970 we still do not have a useful taxonomy.

Mills - I'm not saying necessarily to develop a new taxonomy per se, but use some sort of categorizations. Somehow to get the base amount of data standardized in some form and stored someplace. This is what you're suggesting when you suggest a journal, it seems to me.

Swain - I'm just suggesting reporting the studies that are done.

Mills - This is going to be highly diversified, it won't be standardized.

Swain - I'm suggesting the thing be recorded using a data bank form which the responsible agency contracted would have to develop.

Mills - Wouldn't this be just as valuable if it were put into a computer?

Meister - Let's start with first things first. For example, what behavioral units, what equipment units are you going to include in this data base?

Swain - I really don't know.

Meister - I get the impression you don't really care.

Swain - No, I said, for example, that we have a task using this screwdriver to do something or other, you might want to use this as a search term.

Meister - Would your data base be dealing solely with task elements like simple stimulus-response connections or would it include functions like takeoff or landing or navigation, would it include various types of operational tasks, or is it in the production area, or what?

Swain - As I mentioned, all kinds of tasks, even including the laboratory tasks.

Meister - What I'm getting at, Alan, is you just can't propose a data base, you have to specify what it contains.

Swain - If you're saying that I haven't come up with all the answers, I'll have to admit that.

Jenkins - You're point is important because the program has a specific objective to establish a data bank and certainly we don't expect one to be perfect in the beginning. There are many problems on how you classify behavior but do you believe that given the method which you described here, that we can come up, say in just one area, command and control, where with the results of human error impact performance? Someone mentioned that you have to pinpoint the system function where the results of human error are so critical that by collecting data in this fashion we would then be able to go back and remedy the causes for types of errors which are not tolerable. There are some errors that are going to be tolerable but for intolerable errois do you think this kind of data is a useful tool?

Swain - Well, as I say I think this kind of a data bank can be started fairly soon and to the extent that there are tasks going on in the real world for which human error rates and time and so on can be collected. All I am suggesting is a systematic way of collecting this data.

Meister - For example, I am a user. I want to predict the error that will be associated with the SQS-26. That's a total console, is it not? It contains, dials, meters, controls and so forth. How do I enter your system, do I ask what is the error rate associated with sonar type consoles? Do I enter the data system asking what is the error associated with tracking using a joy stick, or what? How many items, what parameters do I enter your system with? You have to tell me that or I can't really use your system.

Swain - You're asking me to come up with the answer right now, I can't no. I know what we do. All I can say is what we do at Sandia. If we were going to participate in an error bank I can tell you what we would do. You're using some tasks I'm not familiar with.

Meister - We can't use your tasks, you won't tell uc about them.

Swain - What would I do? Let's say we have a situation where we have people using a certain item in a field exercise and let's say they have to open a combination lock. We have data that says out of so many opportunities, the error rate is so and so in opening this lock under these kinds of conditions. We would try to fill in the data bank form by getting as many different types of human error rates we have for that particular task. By types I'll use an example. Maybe an exercise of one night we had to do this in a heavy downpour. This would be a performance shaping factor which either did or did not materially effect the error rate based on some more benign operating conditions. We would be entering that type of data and we would be classifying it as combination locks. This would be one search term because if I'm interested in human errors related to combination locks I'm sure going to use that as a search term. Right off the top of my head I can't think of other search terms I would use, but I as a potential user and having other users needs in mind would try to devise as many search terms as I thought somebody might be interested in.

Meister - I take it you're recording down every time. For every datum you're putting down everything you can possibly think of that might have possible impact on that error datum.

Swain - I'm not suggesting that for any non-zero performance shaping factors, no. I've even restricted them further than that in the report. I've said that any minus performance shaping factors be used as search terms.

Tolcott - I think that what this discussion is pointing up, is that when somebody asks a question like they did this morning, "What is the density function for performance?", a general answer has to be that human performance is affected by certain factors but the specific factors that will affect it, and the amount of effect will differ from type of system to type of system. This is the first time today that I've heard command and control mentioned as a specific system which might be the focus of effort of this particular ADO, and if this is true I think one might usefully look at the characteristics of that kind of a system and try to identify the performance shaping factors as part of the data bank.

Jenkins - This is what I was getting at earlier. I said that in the structure of the program the key naval system should be looked at first. This is most important for human reliability. It is in command and control that we must put greatest effort in data bank generation.

Siegel - Why is it that although we have been yelling data bank since 1952, the data bank bankers have not come to grips with the problem and at least given us a taxonomy nonetheless the data bank. They can't agree amongst themselves on the taxonomy, nonetheless on what data to collect, after they get the taxonomy, to put in the bank. I'm glad I'm not your banker, frankly.

Meister - That's an argument ad hominum, It's not relevant. What you're saying, in effect, is we are deficient in our research.

Siegel - No, I'm not saying that at all. I'm saying is the quest for this taxonomy and data bank a will of the wisp?

Swain - I think it's after the Holy Grail, really. My feeling is that after several years if one had thousands of these data bank forms, you would have an awful lot of human error rates collected for a great many numbers of tasks. I think then one would be able to induce a task taxonomy that would have some usefulness to the Navy.

Meister - The term task taxonomy (shades of Bob Miller) is beginning to assume a very negative connotation. Forget about that term. Strictly from a user's standpoint, I have to know what categories of data you've got or I can't use your system. I don't care whether you call it task taxonomy or what have you. I have to have something to be able to enter the system with, and that is a judgement which the originator of the system must make at some point in the development of the system and desirably ahead of time rather than after the fact. Although you can obviously modify your classification system as you get more data in.

Coburn - Alan, can I ask about the form of the error rate data? You mentioned in your paper the difficulty of putting time down to estimate error rate, yet the examples that you give are time related.

Swain - It's not as easy you'd think.

Coburn - Certainly it's very difficult to collect error rates in any meaningful form. The problem is with the denominator.

Swain - To me time data is not very useful in my work.

Coburn - What kind of a probability do you come up with in a situation like that?

Swain - I wish I knew more about some of the operational tasks you deal with. Let's say that onboard a ship you certainly have a record of the number of times that you have F4 takeoffs in some period of time and the time is just irrelevant. You should be able to know from the task analysis what is done in the launcher operations for each F4 takeoff and if you are recording human errors you may find then that, let's say, out of one thousand times on F4 takeoffs the crews twice use the wrong holdback bolt. That's an error rate that's useful. That's what I'm saying, let's collect human error rate data. But these performance shaping factors are facts too. These are simply judgements; judgements are facts. They're not as nice scientific as physical units, but they are facts. I.et's take this further. Let's supposing that in both cases those holdback bolt misplacements were done at night so you'd want to know how many times these operations were done at night to get an error rate for night operations. Supposing the only times you had holdback bolt mistakes were at night. This is a very valuable piece of data. Night operation might be listed as a large negative PSF rating.

Mackie - Doesn't the operation imply a regularity, a homogeneity of events?

Swain - It depends on how you are going to define uniqueness. There is some level of commonality, obviously, or you'd never get anything done in command and control. And of course in an applied situation you know you're going to have some error.

Siegel - Most of these taxonomies or task data banks alluded to mocor tasks and that's what we've been speaking about mostly here. My current impression is that we have the motor aspect pretty well under control, frankly, from a design point of view. We have more problems with the intellective aspects of the task. Just assuming me to be right for the fun of it, if that's the case, we're making a mistake by emphasizing the motor aspect and secondly the intellective aspect. How are we going to get that into the taxonomy?

Meister - That's a very oifficult thing to do, obviously.

Siegel - It would be that aspect which I personally would be interested in. I would continue to argue that in terms of percentage variance accounted for in terms of human unreliability, in command and control situations which they are interested in, the pay-off is going to be the cognitive event. These are going to account for a high proportion of the predictive variance-whatever the percent may be or as small as it may be, rather than the motor.

Meister - What is important in developing this data bank is to be able to specify how you are going to accomplish the collection of your data.

Siegel - Without overemphasis on the easy part because you can see the motor aspect.

Meister - There's one thing that has to be pointed out in Alan's behalf, and that's even for motor activities the amount of performance data, performance prediction data, is abysmally low.

Regulinski - Turn to page 16 won't you? Would you fudge through with us please how you calculated reliability, human performance reliability, given an error rate of .0015 as an example.

Swain - There are 12,587 connectors that we lcoked at during stockpile sampling of items and in those 12,587 we found the errors that we actually listed below there.

Regulinski - I followed you quite well. I want you to tell me explicitly what it is that you want the reliability engineer to do with the figure .0015.

Swain - All right, .0015 is the probability of having bent pins with this type of connection in the type of task that is described on the other side of the card and this particular error rate was judged to be much too high. In this particular case we were able to show that four times out of a thousand you would have one or more of these kinds of human errors and, of course, if you've got connectors with parts omitted or connectors improperly mated and so on, most of the consequences in the particular application were intolerable. That is to say, they would have a severe result on the system reliability.

Regulinski - Alan, let me re-phrase my question again. I'd like very much to use the figure .0015 to compute reliability. Would you run through that for me, please?

Swain - I guess I don't understand the question, then. That is the estimated reliability, that is the estimated error rate for bending pins in these kinds of connectors.

Regulinski - Do I understand you correctly, that the error rate is in fact the reliability.

Swain - I normally use the term error rate. It and reliability are just the converse. In this particular case, the probability of the error is .0015. .9985 proportion of the time there will not be a bent pin.

Regulinski - Let me ask that question explicitly. You are saying that the reliability is 1 minus the error rate, is that correct?

Swain - In this particular case, yes.

Coburn - The term error rate is synonomous here, I think with the probability of error.

Meister - Are you implying, Ted, that there is something wrong with that?

Regulinski - Grossly!

Meister - Well, would you mind telling us?

Regulinski - Yes, certainly. In computing reliability there is one and only one condition which leads to reliability being equal to 1 minus the error rate. Only one: if and only if the error rate happens to be exponentially distributed. Let me demonstrate this on the blackboard. If we expand $e^{-\lambda t}$ by any power series theorems we obtain: $e^{-\lambda t} = 1 - \lambda t + \frac{(\lambda t)^2}{2!} - \ldots \ldots \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^k}{k!}$. If we assume that the quantity λt is very much smaller than unity, the quantity $e^{-\lambda t}$ is approximately equal to $1 - \lambda t$. Further, assuming unity time, one can equate reliability to approximately 1 minus the error rate.

Meister - Would it be acceptable to say that you could use one minus the error rate as a gross approximation of the true reliability?

Regulinski - How gross is gross? Again let me demonstrate on the blackboard how gross, gross can be. Let us say that $\lambda t = 0.4\text{x}10^{-6}$. This is a typical transistor figure. Another transistor has $\lambda t = 0.2\text{x}10^{-6}$. Assuming exponential failure rate of the transistors, we know for a fact that the mean can be obtained from the reciprocal of the failure rate. Assuming unity time and performing the indicated division, we see that the first transistor would have a mean of 2.5×10^6 hours, and the second transistor a mean of 5×10^6 hours! Would you like to hang by your posterior for the difference between 5 million and 2.5 million hours? Gross here is so gross that your reliability engineer could commit professional suicide using such approximations.

Meister - Your reliability engineer also deals with very gross estimates too.

Siegel - I'd just as soon hang for 5 times 10^6 as 2 times 10^6!

Meister - Are you saying that it is necessary to know the empirical distribution? In that case I don't disagree with you. Presumably if you collected enough data you would be able to arrive at the empirical distribution. We tried this on our RADC studies.

Regulinski - This has all the ingredients to the questions asked earlier: what sort of data we shall collect. If you hang yourself on this approximation then clearly you want weaker type of data then what I'm going to suggest from the point of view of the systems engineer, and his modeling needs. It is simply for the guidance of the Navy that we suggest that these approximations as gross as they are must be realized if they're going to guide the Navy to collect data. Please be aware of the trap you're in.

Meister - In order words, if I may recast it in my own terms, you're saying let us be aware of the assumptions under which we are working.

Swain - I just don't think it (Regulinski's argument about the grossness of the error rate) is as important as you make it out to be.

QUANTIFICATION OF HUMAN PERFORMANCE RELIABILITY PESEARCH METHOD RATIONALE

T. L. Regulinski
Air Force Institute of Technology

QUANTIFICATION OF HUMAN PERFORMANCE RELIABILITY

RESEARCH METHOD RATIONALE

T. L. REGULINSKI

AIR FORCE INSTITUTE OF TECHNOLOGY

FOR PRESENTATION AT THE NAVY HUMAN RELIABILITY WORKSHOP
SPONSORED BY
NAVAL SHIPS SYSTEM COMMAND, NAVAL AIR DEVELOPMENT CENTER
AND THE OFFICE OF NAVAL RESEARCH
WASHINGTON, D.C. 22-23 JULY 1970

INTRODUCTION:

To be able to study quantitatively the performance of a system, human or hardware, it is necessary to formulate for such system a mathematical analog whether it be in the analytic or numerical deterministic realm or in the analytic or numerical stochastic realm. This formulation constitutes the essence cf modeling. When the study is directed at a system governed by independent variables, the values of which may be chosen arbitrarily, and dependant variables, the values of which are determ.ned from the former, the formulation constitutes system deterministic modeling. When the study is directed at a system governed by functions of some random variables (variates), the formulation constitutes system stochastic modeling. Under consideration in this paper is time-space continuous stochastic modeling of the human system performance parameter RELIABILITY. In time continuous modeling, human reliability may be defined as:

$$R_h(t) = P\text{ (Task performance without relevant errors under constraint of time and environment)} \quad (1)$$

In time discrete modeling human reliability has been variously defined by point probabilities (Ref 1, 2 and 3).

MODELING PROCESS:

The modeling process encompasses the activities of four domains: the domain of random data generation and processing, the domain of mathematical model formulation, the domain of prognosis, and the domain of decision making. In the first domain, the data generated may come from a real world source as examplified by the observed times to some human task performance. The data so generated is subject to the usual statistical processing which would also include tests of randomness, stationarity, ergodicity, etc. In the absence of a real world source, the data may be generated by the Monte Carlo method. This would require a random number generator which could then be used to synthesize varied store of distributions (Ref 4).

The process moves from the data generation to mathematical model formulation domains via calculus of deductive reasoning. From the data statistics, a function is sought for the probability distribution of the random variable generated. Such a function is a homomorphic or an isomorphic mathematical model which analytically describes the behavior of the random variable. In the continuous case, the function is the probability density (pdf), or for the discrete case, the point probability. Both functions facilitate direct computation of the various characteristics of the random variable such as, for example, the expectation or the variance. Once the governing function is isolated, methods for which are documented (Ref 5 and 6), the process can move to the domain of prognosis via calculus of inductive reasoning. Here, using the governing function, we are involved in predicting the behavior of the random variable, and in establishing suitable criterion against which predictions can be evaluated. Correlation and regression modeling may be undertaken here also since, in essence, both constitute methods of prediction. The domain of decision making is entered via decision theory. It encompasses among others hypothesis testing, estimation, multiple decision and sequential testing, modeling of likelihood, utility and cost functions, and formulating rules for decisions (strategies). In this domain, as in the two proceeding ones, the governing function is fundamental to any model formulation which may be undertaken.

MODEL FORMULATION:

The human performance tasks that are most analogous to hardware system performance in time continuous domain, and thus are most amenable to classical reliability modeling, are continuous operation tasks such as vigilance, monitoring, controlling, and tracking (Ref 7 and 8). The human performance reliability of such tasks can be modeled by denoting the probability of human performance error during the time interval Δt, given errorless performance up to some time t, by $e(t)\Delta t$, where $e(t)$ is the human error rate analogous to hazard rate in reliability theory. The probability of errorless performance of at least time t can then be modeled as follows. Define the events:

A = errorless performance of time t duration

B = error will occur in interval $(t, t+\Delta t)$

Thus, P(A) is the probability of errorless performance, or reliability of human performance as a function of time, and consistent with (1) is denoted by $R_h(t)$. Further, from the definition of events:

$$P(B/A) = e(t)\Delta t \tag{2}$$

The probability of errorless performance from time zero to time t and from t to $t+\Delta t$ is the joint probability $P(A \cap \bar{B})$, where $\bar{B}$ denotes the event error will not occur in interval $(t, t+\Delta t)$. The joint probability, in turn, can be expressed as

$$P(\bar{B}/A)P(A) = [1-P(B/A)]P(A) \qquad (3)$$

$$= R_h(t+\Delta t)$$

or $$[1-P(B/A)]\ R_h(t) = R_h(t+\Delta t) \qquad (4)$$

Substituting (2) → (4)

$$[1-e(t)\Delta t]R_h(t) = R_h(t+\Delta t)$$

which leads to:

$$-e(t)R_h(t) = \frac{R_h(t+\Delta t)-R_h(t)}{\Delta t} \qquad (5)$$

But from the definition of a derivative

$$\frac{dR_h(t)}{dt} = -e(t)R_h(t)$$

or $$\frac{dR_h(t)}{R_h(t)} = -e(t)dt$$

It follows :hen that

$$\int_1^{R_h(t)} \frac{dR_h(t)}{R_h(t)} = -\int_0^t e(t)dt$$

and

$$R_h(t) = \exp\left\{-\int_0^t e(t)dt\right\} \qquad (6)$$

The relation between $R_h(t)$ and $e(t)$ given by (6) is a completely general model of human performance reliability in that it holds whether the error rate is a decreasing or an increasing function of time, or it is time invariant.

Other forms of (5) follow from the unique relationship which exists between the error rate e(t), the probability density function f(t), and the reliability function $R_h(t)$. The relationship between the three functions can be shown to be (Ref 9):

$$e(t) = \frac{f(t)}{R_h(t)} \tag{7}$$

It follows therefore, that the reliability function can be expressed in terms of the probability density function as

$$R_h(t) = \int_t^\infty f(t)dt \tag{8}$$

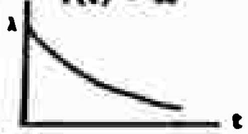
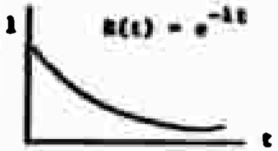
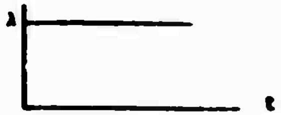
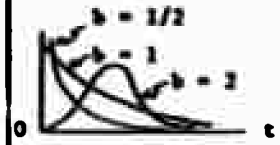
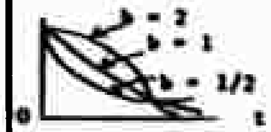
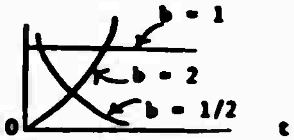
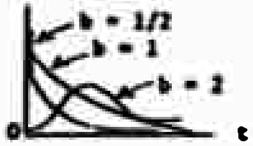
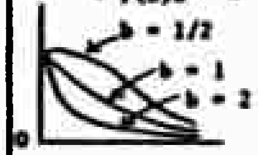
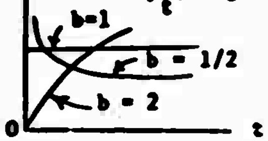
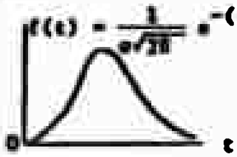
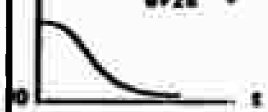
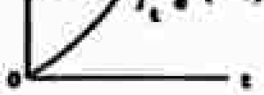
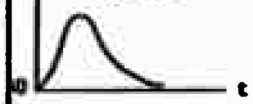
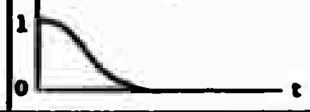
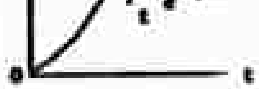
For each underlying probability density of time to error, the reliability function and the error rate function can be obtained from (7) and (8). The tabulation of these functions for the Exponential, Weibull, Gamma, Normal, and the Lognormal probability density is given in Table I.

Assuming that the probability density function can be successfully isolated, the first order moment can be obtained from

$$m = \int_0^\infty tf(t)dt \qquad .9)$$

which is the mathematical expectation of T, the random variable time to error. It is also referred to as the mean value of the variate T. In human reliability lexicon and analogous to classical reliability, m could denote mean time to first human error or mean time between human errors depending on corrigibility of the errors, and how the time to error is referenced.

TABLE 1

TYPE OF DISTRIBUTION	PROBABILITY DENSITY FUNCTION f(t)	RELIABILITY FUNCTION R(t)	ERROR RATE = e(t) = $\frac{f(t)}{R(t)}$
EXPONENTIAL	$f(t) = \lambda e^{-\lambda t}$  	$R(t) = e^{-\lambda t}$  	$e(t) = \lambda$ 
WEIBULL	$f(t) = \frac{b}{a} t^{b-1} e^{-t^b/a}$  	$R(t) = e^{-t^b/a}$  	$e(t) = \frac{b}{a} t^{b-1}$ 
GAMMA	$f(t) = \frac{1}{\Gamma(b)a} \left(\frac{t}{a}\right)^{b-1} e^{-t/a}$  	$R(t) = \frac{1}{\Gamma(b)a^b} \int_t^\infty \tau^{b-1} e^{-\tau/a} d\tau$  	$e(t) = \frac{t^{b-1} e^{-t/a}}{\int_t^\infty \tau^{b-1} e^{-\tau/a} d\tau}$ 
NORMAL	$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\theta)^2/2\sigma^2}$  	$R(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty e^{-(\tau-\theta)^2/2\sigma^2} d\tau$  	$e(t) = \frac{e^{-(t-\theta)^2/2\sigma^2}}{\int_t^\infty e^{-(\tau-\theta)^2/2\sigma^2} d\tau}$ 
LOG-NORMAL	$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-(\ln t-\mu)^2/2\sigma^2}$  	$R(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{\tau} e^{-(\ln\tau-\mu)^2/2\sigma^2} d\tau$  	$e(t) = \frac{\frac{1}{t} e^{-(\ln t-\mu)^2/2\sigma^2}}{\int_t^\infty e^{-(\ln\tau-\mu)^2/2\sigma^2} d\tau}$ 

An Adaptation to Human Performance Reliability of Distributions
Commonly Used for Reliability Analyses of Systems

Isolation of the Probability Density

The crux of the analytical methodology lies in isolating the underlying probability density which was identified earlier as the governing function. This, fortunately, is not an insurmountable problem in light of the work of Gumbel, Chernoff and Lieberman (Ref 10 and 11) who have investigated the use of graphical procedures for use in related type inquiries. The extension of their work leads to computerized isolation procedure developed for the problem on hand (Ref 12). The procedure involves plotting the individual time to error observations t_1, t_2, t_3,....t_n vs plotting positions on special probability paper. The attribute of such plotting paper is that if the observed values of t conform to the assumed distribution F(t), the relation between t and F(t) will be of the form

$$\mathrm{Ln}\{1/R(t)\} = A \,\mathrm{Ln}\, t + B \qquad (10)$$

which is a linear equation with independent variable Ln t, slope A, and intercept B. Once the observations are plotted, a straight line may be fitted through the plotted points provided the scatter of observations is sufficiently small. An alternate method would be to employ a numerical technique such as the Legendre method of regression (Ref 13).

If the fit is in doubt, then the acceptance or rejection of the assumed probability density function must necessarily be also in doubt in which case a goodness-of-fit-test could be used; e.g., the distribution free Kolmogorov-Smirnov test of hypothesis otherwise known as the d test (Ref 14). The test is set up by letting

$$D = \max |F(t_i) - S(t_i)| \tag{11}$$

where $F(t_i)$ denotes cumulative population distribution having a mean of m and standard deviation of σ, $S(t_i)$ denotes the cumulative distribution of the observations, and D the maximum difference over the n data points. The null hypothesis H_o that $\hat{m}$ and $\hat{\sigma}$ are estimators of m and σ is tested against the alternate hypothesis H_1 that $\hat{m}$ and $\hat{\sigma}$ are not estimators of m and σ. The null hypothesis is accepted at the α_d level of significance if $D < d$, where $\alpha_d = f(N,d)$. A list of α_d for all possible values of d and N is given in Reference 14.

Experimental Procedure

In general, the experimental design may call for subjects to perform, in a simulated or actual environment, a select task for a period of time sufficient to provide relevant performance errors. The observed times to first such error would then constitute the random data generated from which the mathematical model formulation could be attempted. In order to minimize errors due to learning and errors induced by fatigue, the experimental design should call for time limited task yielding reasonable error output; one which would require little or no learning, or one which would utilize fully trained subjects. If, for example, a simple vigilance task was to be selected, the experimental procedure may require the subjects to observe, say, a clock-type arrangement of light bulbs flashing sequentially on-and-off, and to respond to a failed light event by some such means as pressing a hand-held switch whose output could be coupled either to a graphical recorder or directly to a computer. The failed-light events would, of course, be programmed randomly to preclude pattern recognition. Congruous with vigilance type tasks the data acquisition would then consist of two different kinds of time-to-first-error; a miss error denoting that the subject did not detect the failed light event, and false alarm error denoting an error by anticipation.

PROGNOSIS:

For purposes of illustration let it be assumed that for the vigilance task discussed in the experimental procedure the miss error and the false alarm data were generated. The analytical methodology using density function isolation and the Kolmogorov-Smirnov test would then lead to the acceptance of the density function which governed the behavior of the miss and false alarm errors, and to the rejection of others subjected to the same test. Suppose the underlying probability density function was found to be Weibull, then

$$f(t) = \beta(\frac{t}{\alpha})^{\beta-1} \exp\left\{-(\frac{t}{\alpha})^{\beta}\right\} \qquad t \geq 0 \qquad (12)$$

Here α and β are positive constants. The former is the scale parameter and determines the dispersion of the probability density function about the mean. The latter is the shape parameter and determines whether the error rate is time variant or time invariant. The reliability function follows from (8) yielding

$$R_h(t) = \exp\left\{-(\frac{t}{\alpha})^{\beta}\right\} \qquad t \geq 0 \qquad (13)$$

Depending on the man-machine system under study, reliable task performance may be defined so as to preclude only the miss errors, or it may be defined to preclude <u>both</u> miss <u>and</u> false alarm errors. Recently completed research (Ref 15) which investigated vigilance type task discussed here, indicates that both the miss errors, and the combined miss and false alarm errors would be Weibull distributed; however, each would have a different set of shape and scale parameters. Clearly the reliability predicted from expression of equation (13) would then be different valued for each of the two cases cited. The prediction of the mean time to first miss error, and the mean time to combined miss and false alarm error would also be different valued. Each, however, would be determined from the expression of equation (9). For the Weibull this leads to (Ref 16):

$$m = \alpha^{\frac{1}{\beta}} \Gamma\left\{\frac{1}{\beta} + 1\right\} \tag{14}$$

where the function $\Gamma(\cdot)$ is the gamma function which is defined for every $v > 0$ by

$$\Gamma(v) = \int_0^\infty x^{v-1} e^{-x} dx \tag{15}$$

the values of which can be obtained from standard tables.

DECISION MAKING:

Extending the example of the vigilance task further, one can relevantly pose the equestion: how long should a subject be expected to perform the given task before fatigue sets in causing the error rate to increase. Alternately, one can ask what is the expectation of the variates. Clearly the answer must come from expression of equation (9). The decision of subject replacement in the performance of the given task will, of course, be dependant on the cost functions associated with the miss error and the false alarm error. In either case it is important to stress that the isolation of the probability density function underlying the miss and false alarm error is a primary requirement before a decision model can be formulated. The minimax, the maximum likelihood, and the Bayesian are but three examples of decision models which can be formulated and are potentially useful. These models have been subject of extensive research, and are adequately covered by a number of sources (Ref 17 and 18).

REFERENCES:

1. Williams, H.L., "Reliability Evaluation of the Human Component in Man-Machine Systems," ELECTRICAL MANUFACTURING, April 1958.

2. Swain, A.D., "A Method of Performing a Human Factors Reliability Analysis," Sandia Corporation MONOGRAPH SCR 685, August 1963.

3. Meister, D., "Methods of Predicting Human Reliability in Man-Machine Systems," HUMAN FACTORS, December 1964.

4. Metrolpolis, N., and Ulam, S., "The Monte Carlo Method," JOURNAL OF AMERICAN STATISTICAL ASSOCIATION, Vol 44, No 247, 1949.

5. Regulinski, T.L., RELIABILITY ENGINEERING, NAFI, Indianapolis, Indiana, 1965.

6. Von Alven, W.H., RELIABILITY ENGINEERING, Prentice-Hall, Inc., 1965.

7. McCormic, E.J., HUMAN ENGINEERING, McGraw Hill Book Co., 1964.

8. Verdier, P.A., HUMAN ENGINEERING, N.Y. Exposition Press, 1966.

9. Barlow, R.E. and Proschan, F., MATHEMATICAL THEORY OF RELIABILITY, John Wiley and Sons, Inc., 1965.

10. Gumbel, E.J., STATISTICS OF EXTREMES, N.Y. Columbia University Press, 1958.

11. Charnoff, H. and Lieberman, G.J., "The Use of Probability Paper," ANNALS OF MATHEMATICAL STATISTICS, #27, 1956.

12. Regulinski, T.L. and Askren, W.B., "Mathematical Modeling of Human Performance Errors for Reliability Analysis of Systems," Report AMRL-TR-68-93, WPAFB, Ohio, 1968.

13. Freund, J.E., MATHEMATICAL STATISTICS, Prentice-Hall, 1962.

14. Massey, F.J., "The Kolmogorov-Smirnov Test for Goodness of Fit," JOURNAL OF AMERICAN STATISTICAL ASSOCIATION #42, 1951.

15. Regulinski, T.L. and Askren, W.B., "Mathematical Modeling of Human Performance Reliability," ANNALS OF ASSURANCE SCIENCES, January 1969.

16. Regulinski, T.L., "Systems Maintainability Modeling," ANNALS OF ASSURANCE SCIENCES, February 1970.

17. Pierre, D.A., OPTIMIZATION THEORY WITH APPLICATIONS, John Wiley and Sons, 1969.

18. Breipohl, A.M., PROBABILISTIC SYSTEM ANALYSIS, John Wiley and Sons, 1970.

DISCUSSION OF REGULINSKI's PRESENTATION

Tolcott - All of the relevant questions that could be asked of the data in this bank relate to time. It appears to me that many of the tasks that we're concerned with don't fall into that category.

Regulinski - No, absolutely not, however, the modeling domains listed on the blackboard are perfectly valid whether time and/or space are discrete or continuous. In the former case you are not computing R(t) as a function of time, but rather R(x) as a function of x. But if your formulation of the kind of data that you may want to collect is discrete, time category may be unnecessary. Beautiful!

Lamb - But you also said that we have to listen to you, before we really know what kind of task, because what we may be doing, as a discrete task may in fact be continuous. So we may think it's typical - applicable in a discrete situation, when in fact you would think it continuous.

Swain - Yes, but I would say that many of the systems engineers that you're talking about, although they know things are time-continuous and so on realize they don't have the data and so they, at least in my experience, most of the systems engineers I work with, deal with discrete things, and man they're happy with point estimates. They use point estimates. All the tables of defect data they use have fudge factors applied to them because they weren't tested under the appropriate environmental domains.

Regulinski - You are obviously pointing out that fact that amongst the blind, a one-eyed man is king.

Meister - Eventually we're going to have to deal with continuous data. It's well known, nobody has to tell us, that we have a hard time handling continuous data at present which doesn't mean that eventually we won't be able to handle it. Certainly it shouldn't mean that we shouldn't attempt to handle it. The tracking people had done a fair amount of work, I believe, with probability density functions, although I don't think they invented the application to reliability aspects.

Regulinski - The tracking people do have a very interesting model that is being used at Cornell University. You might be interested because it is in your domain. Input to tracking is taken directly off the muscle. Say you're doing hand tracking, and, the input is taken from the muscle. The muscle body generates certain frequencies. These frequencies are then measured and what is known as power spectral density analysis is performed. The power spectral density function will also lead you to reliability. There is a relationship.

Meister - It seems to me that what you're pointing out is that one must know exactly for what use the data you're collecting is going to be applied.

Tolcott - Yes, you can't answer that question only by looking at the task. You have to look at the objective of the man who is asking the question as Dave pointed out earlier. We can take your system and not worry about how long a man ought to be left in a spot before you yank him out because you might want to ask a completely different question. For example, what is the maximum number of targets to be shown on any one display. Maybe you can divide the area into sectors, and reduce the load on each man, and in that case your probability density function will be quite different with these different parameters.

Regulinski - That is correct, for example take marksmanship. The Army found out some time ago that a marksman who shoots the rifle, his distribution of bullseye miss distance is Rayleigh distributed. The same holds true on range, if you shoot a missile down range the distribution of the missile miss distance is also Rayleigh distributed. Once you have the probability density function over the distribution then the next domain friends is prognosis, prediction. That's the key you see. You're absolutely correct.

Tolcott - You were asked a question earlier which you said you were going to answer. How does the system reliability engineer predict the reliability? What I was getting at this morning was that the data bank may not be the only tool we have to predict reliability. We may want to get other data in the laboratory when necessary.

Meister - All the data that you get from your simulation tests eventually goes into the data bank. I would be the first one to recognize that it will be very unlikely that we would get answers to every question. That would be an ideal situation.

Tolcott - It will be "ideal" only when you have gotten measurements of peoples performance under every conceivable situation which you want to know.

Regulinski - You have an example, the Navy started 10 years ago with the Air Force on MIL-HDBK-217. I think you can learn from their failure. There are many many really gross errors, yet to this very day it is worth its weight in gold.

Meister - At the same time MIL-HDBK-217A only deals with electronic components. All the engineers swear bloody murder because there is nothing equivalent to it that they can use for non-electronics.

Regulinski - I think you can place faith, trust, in that 217A is good, not only because the process can tell you what probability density function is. In short, let me say this. You get to learn how to model adequately equipments which constitute the total system.

Meister - Of course, it may be that one has to go through this process from the more molecular component to the larger function. In other words, what I'm trying to say, it may be unavoidable.

Regulinski - Let me ask you this, Dave, I think this might be a relevant question. Do you really want to break down these molecular tasks or can you look at the human function as a system?

Meister - It's a very relevant question, although possibly not in systems engineering terms. Unfortunately, or fortunately, the people that one works with, design engineers, often ask questions which are not formulated in terms of major function limits. The engineer may well ask you what is the affect of changing a location of switch from one position to another. It's a very molecular kind of question. I would suspect that we would be further along if we never had to answer that kind of thing, but if we say that we should address ourselves to the kinds of questions design engineers ask then unfortunately, unless he gets better educated, as you suggested, than he is in process of being, then we do have to respond to that kind of thing. I would be very, very happy to eliminate the necessity for dealing with extremely molecular tasks. It's a real pain.

Your major problems of combining predictions stems just from that aspect alone, the molecular aspects.

Siegel - I would argue in a couple of ways. I'm beginning to worry about what's going on here. First, I would argue that I think that you underestimate the ability of some of these fellows to do arithmetic. I think you'll find out that most of them do pretty good arithmetic (and calculus, too), but that's not the point. In your presentation you are essentially saying to the human reliability people, "Get on my math model, get on my analytic technique." Now, your math model is merely some type of symbolic representation, that's all any model is whether you put it in letters, or not. I don't think that I for one am ready to be seduced into jumping into your math model. I do agree with you that while we're getting Beta, we want to know the nature of the distribution. I don't think any person here at this table would argue with you when you say, with a good deal of assurance, that the vigilance function is exponential; I think everyone here would argue that they could change the nature of that distribution just by superimposing a different confidence level on the observer. The point is that I think that we ought to be very wary about jumping on anyone's bandwagon or emulating any specific discipline. We do need help from wherever we can get it in terms of this pursuit of an optimum method for predicting human reliability.

Regulinski - I agree with Art but when he gets up tomorrow to present his paper, he's going to use these very methods in his model. Tomorrow we will hear a very fascinating paper .... Art Siegel's.

Meister - Jim has suggested earlier that the way the Navy functions, the human factors man doesn't get into the design loop until after the basic design parameters, the design concepts have been developed. Presumably the kinds of questions you, as a systems engineer will ask, would relate to the very, very early formulation of the design concept.

Jenkins - What I would like to see from the programs which we have, is how to make use of your point of view. I know the fairly advanced system modeling generally does not take place until after the fact. Operational requirements are related to missions, as to perform a 60 day mission or launch X number of airplanes. They don't state it in terms of the system function, only after the fact, when they attempt to show to the scientific community that they have a pretty good system or why they don't have a pretty good system and are justifying the existence of what they have.

Meister - There is apparently a continuity, as I said before, of system development functions. One extreme is represented by systems engineering kinds of questions. The other extreme is represented by the extremely molecular detail design of the individual control panel or the individual component. Perhaps we are more able to address ourselves to the more molecular kinds of design questions because of the data we have. Of course this doesn't mean that we should not consider attempting to answer the other extreme which you represent.

MAN-MACHINE MODELING: SOME CURRENT DEFICIENCIES AND FUTURE NEEDS

R. E. Blanchard and R. L. Smith
Integrated Sciences Corporation

MAN-MACHINE MODELING: SOME CURRENT DEFICIENCIES AND FUTURE NEEDS

R.E. Blanchard and R.L. Smith
Integrated Sciences Corporation
3122 Santa Monica Boulevard
Santa Monica, California 90404

Prepared for:

Navy Human Reliability Workshop
July 22-23, 1970
Washington, D. C.

MAN-MACHINE MODELING: SOME CURRENT DEFICIENCIES AND FUTURE NEEDS

I. INTRODUCTION

Justification for developing man-machine performance prediction models, particularly operability (reliability) models, has been based on the premise that human errors play a major role in reducing system effectiveness. Interestingly, there is only limited empirical data available on the frequency, nature, and effect of such errors, undoubtedly because of the difficulty in obtaining admissions by personnel of their errors and the difficulty and expense associated with collecting such data by observation techniques in field situations (e.g., Chapanis, 1959; Shapero, Cooper, Rappaport, Schaffer and Bates, 1960; Irwin, Levitz and Freed, 1964). Nevertheless, there seems to be considerable anecdotal evidence that human errors are a major consideration in systems effectiveness. The recent anouncement that the failure of Apollo 13 was due to human error represents a dramatic case in point.

Mathematical modeling of man-machine systems is a relatively new discipline, having spanned only a single decade. Judging by the output to date, it appears that greatest emphasis has been placed on developing maintainability models. (For reviews see Rigney and Bond, 1964, NSIA, 1967; and Smith and Westland, 1970.) Only a few operability models, designed to evaluate multiman systems, have been developed (Swain, 1963, 1964a, 1968; Pickrel and McDonald, 1964; Meister, 1969; Smith Westland and Blanchard, 1969a,b), although intensive work has been focused on single operator models (e.g., Siegel and Wolf, 1961), and many models of the human operator in continuous tracking systems.

Our recent work in the areas of both operability (Smith et al., 1969a,b) and maintainability (Smith, Blanchard and Westland, 1970; Smith and Westland, 1970) modeling leads us to conclude that thu state-of-the-art has not yet achieved a level which justifies general utilization of models. We conceive two basic needs: (1) increased research on the generation and acquisition of human performance model input data; and (2) a thorough delineation and (at least theoretical) resolution of serious problems inherent in current man-machine models.

A. HUMAN PERFORMANCE DATA

In contrast to the efforts expended in developing models, relatively little work has been directed toward the generation and acquisition of valid, relevant, model input data. As a consequence, none of the existing models, regardless of potential applicability, can be exercised with confidence. This is not unlike the hypothetical situation of building an aircraft airframe without consideration or availability of an engine. The aircraft obviously cannot fly without an engine and existing engines may not "fit" the aircraft.

What is the current state-of-affairs regarding human performance data? Assessing human performance on a variety of man-machine tasks has been a progressively increasing endeavor since World War II. Almost all of such studies have been conducted in laboratory settings. With the exception of a very few studies (e.g., McKendry, Corso, Grant and Scheihing, 1960; RCA, 1960, 1961; Retterer, Griswold, McLaughlin and Topmiller, 1965), all have also been concerned with equipment operations, rather than equipment repairs. It is perhaps natural to suppose that data from these studies can be used directly or indirectly (via some transformation process) as model inputs. Such a supposition apparently provided the impetus underlying the monumental data extraction study by Payne and Altman (1962) and their coworkers (Munger, Smith and Payne, 1962). Although some reservations were expressed regarding the validity and utility of their resulting "Data Store" (e.g., Payne and Altman, 1962; Swain, 1964; Meister, 1964), there appeared to be a general and ready acceptance of these data by the human factors community.

Having been reasonably familiar with much of the literature used by Payne and Altman in developing their Data Store, and having an immediate need for such data in our newly developed operability model, we undertook a similar data extraction program in the hopes of providing more comprehensive data. However, it soon became apparent to us that the various laboratory data were not amenable to meaningful manipulations and transformations and that little could be gained by further attempts to extract data (Mitchell, Smith and Westland, 1967). We concluded with little hesitation that virtually no confidence could be placed on laboratory data as model inputs. We made such a conclusion, it may be noted, with the realization that it left us without a means of exercising our model.

Although it cannot be said that the human factors community -- in particular, individuals involved in modeling man-machine systems -- has expressed unaminity with respect to the lack of utility of human performance literature, such a view tends to be supported by default. That is, the use of laboratory data has occasionally been given verbal support, such as "its better than nothing", but we have yet to observe any concrete, empirical evidence. Perhaps the icing on the cake was exemplified in a recent article by Chapanis (1967) who questioned the relevancy of laboratory results for <u>any</u> real-world purpose. Chapanis' critique formalized what many of us believed for a long time but perhaps feared to publicize. Human factors and so-called applied behavioral research has long drifted into the domain of the theoretical, as opposed to the real world for which it was initially established to investigate.

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 whose presumed applications are for a diverse assortment of military systems.[1]

[1] Most of the available data are associated with maintenance performance. However, our review of these data suggests that they are too few in number and most are questionable in terms of validity and generalizability (Smith and Westland, 1970).

An adequate statement of the performance data problem, followed by general recognition and acceptance, is a basic requisite for seriously advancing the state-of-the-art. Unfortunately, all of these antecedent conditions do not appear to exist within the human factors community. A recent RFP is a case in point. Aften ten years of futile attempts to extract data from laboratory (and field) studies, this RFP nevertheless expressed more than hopeful optimism that perhaps yet another attempt might be successful. We believe that failure to fully recognize the data plight has been a prime reason for the lack of any real progress in the application of man-machine models.

In summary, it seems clear that either stronger emphasis must be focused on the generation of appropriate human performance data or development of quantitative man-machine models ought to cease.

B. MAN-MACHINE MODELS

While we would not suggest that significant progress has not been made in the area of man-machine modeling, we do feel that such progress has not been commensurate with efforts thus far expended. There has been considerable redundant work and continual "rediscovering" of the same basic problems. We have not chosen to advance the state-of-the-art through applications, test, and refinement (or rejection) of the models that were available. Each new modeling problem has been viewed as an opportunity to explore inidividual concepts for elegant, new approaches rather than one of taking advantage of prior work and progressing from there. The intrigue and attraction of mathematics and the variety of directions possible tends to distract the researcher from the actual real-world problem. As Karush (1962) observed, it is perhaps understandable that we try to find ways of avoiding the tedious and less elegant means for accumulating knowledge, but we are in error when we grasp on approach out of weakness rather than the strength of knowledge.

Of particular importance has been the acceptance of various and usually untenable assumptions. The acceptance of such assumptions is often made to expedite completion of model methodology and is equally often forgotten during the ensuing process. Statements are made to the effect that "the assumption was made in order to make the model work." However, such a statement is clearly false; a model based on and in need of an untenable assumption cannot "work".

Examination of existing man-machine models, when stripped of all methodological and mathematical lingo, reveals that they are agonizingly simple. One is immediately struck by the paradox inherent in the implicit assumption that a simple model can adequately describe quantitatively the obvious complexities of man-machine systems. Occasionally, an insightful analyst presents an undeniable case in this regard (e.g., Rigney and Hoffman, 1962; Quade, 1962; Schaeffer, 1962). However, their "pleas" apparently have been either ignored or brushed aside to date. Since we cannot express more eloquently their discussions on the fundamental problem of modeling, the following quotes are offered:

> "The most obvious weakness of mathematical modeling and simulation is what might be called an inherent schizophrenia. It is all too easy to model or simulate a world that does not exist. Validating data are often hard to find, or are nonexistent, forcing the model-maker to substitute assumptions and oversimplification for data. And, he may become so fascinated by it all that he retreats from the real world altogether, electing to spend his time polishing abstractions." (Rigney and Hoffman, 1962.)

> "It is a pitfall to be more interested in the model than in the real world. Technical people with specific training, knowledge, and capability like to use their talents to the utmost. They like to reduce the problem to one they can handle. It is easy for analysts to focus attention on the mechanics of the computation or on the technical relationships in the model :uther than on the important assumptions of the study. When this is done, they may find out a great deal about the inferences that can be drawn from the model, but very little about the question they set out to answer." (Quade, 1962.)

Schaeffer (1962) noted a desire for "dogmatism" on the part of many analysts rather than an orientation toward rational problem solving. When this happens, Schaeffer maintains that:

> "...the analyst tends to begin his analysis with the abstractions science offers, rather than with the problem as given. Where this occurs, we deal neither with good science nor with good systems analysis, but with irrelevant facts and theories which more charitably are known as oversimplified analyses and solutions. The producer, and thus defender, of these irrelevancies tends to justify himself by claiming that if he had more time and money he could have gone into greater detail and produced a more meaningful analysis. However, an examination of his methods and techniques usually does not bear out his assertion. What he did was to force the specific problem to fit his generalized analysis, rather than fit the analysis to the problem."

Closely related to the present issue are the comments of Harkins (1969) which form an ideal epilog for the above quotes. It is Harkins opinion that there is considerable evidence that past approaches have not resulted in successful programs or effective systems and that change is in order. He observed that men must want to change before change can occur. Elaborating further, Harkins suggested the following:

> "...Since men respond to the value system of the organization in which they function, the easiest way to accomplish change would be for the value system of the organization to change. To accomplish such a task would require the highest levels of the government to accept the reality of the present dichotomy and determine to change it. Unfortunately, such a change may be a long time in coming. In the meantime, practitioners of the system effectiveness disciplines in both government and industry must accept the ineffectiveness of their present behavior, reject the lure of the twilight world in which they have drifted, and brave the psychological threat of interaction and communication." (Harkins, 1969.)

To those of us who are faced with immediate system effectiveness analysis problems, it may seem justified to characterize the above comments as ivory tower philosophizing which frequently "demands" requirements that transcend the practical. However, ignoring the problems expressed is tantamount to ignoring the validity of our models. Unfortunately, such problems will not go away as a result of the mere passage of time.

As in any area of research in which literature becomes depressingly prolific, continuity of thought in the man-machine modeling discipline has tended to dissolve through the years and a true measure of the current state-of-the-art can only be assessed by individuals persistent enought to accumulate, analyze and integrate previous work. Tc paraphrase Harkins (1969), what appears desperately needed is a thorough critique of the literature and, very importantly, a delineation of the many problems that must be resolved in order to develop models with practical utility. Then efforts must be directed toward resolving the problems, rather than toward perpetuating them.

II. RECOMMENDATIONS

As indicated in the foregoing, we believe that research in man-machine modeling should be shifted to emphasize input data and practical solutions to methodology problems currently limiting the utility of models.

A. AN IN-DEPTH STUDY OF MAN-MACHINE MODELING

It is suggested that literature related to man-machine modeling research and theory is in urgent need of a comprehensive review and analysis. However, "just another review" is not suggested. Rather, efforts should clearly describe the state-of-the-art in all of its ramifications and delineate concisely all of the theoretical and technical problems whose solutions will permit at least a major milestone in modeling to be achieved. Such a review should not pull any punches in the sense that it underplays weaknesses and overplays strengths. Progress is best served by describing briefly what we have and discussing in detail what we don't have but need.

In addition to presenting a thorough appraisal of the state-of-the-art, the suggested review and analysis should also discuss constraints which serve to delimit, delay or prevent altogether solutions to important problems. Costs, time and practicality are obvious considerations. However, we perceive the inertia of the "system" itself as the primary deterrent to progress. For example, we have noted time and again that many RFP's request products that cannot be realistically developed with funds made available within the time frame desired. However, industry nevertheless responds positively and optimistically to such RFP's because it needs the business and frequently has no alternative project choices. The vicious circle is complete when the products (reports) of such projects cloud weaknesses of results and propose further research to achieve "even better methods". Unable to keep up with the massive literature produced by contractors, customers cannot scrutinize methodologies and results carefully, are often deceived into false optimism, and agree to the proposed further research. Thus, the vicious circle is preserved and perpetuated -- at least until judgment day when the product is applied to the real world.

Our perception of the current state-of-the-art leads us to believe that man-machine models will not be generally applicable to multiman systems in the very near future. However, planning and conducting relevant research in the near future will at least assure that practical application will not be deferred indefinitely. The suggested review should focus on identifying specific areas of research and providing preliminary designs for studies whose results would be of greatest benefit to facilitating model development and application. One of our greatest current weaknesses in modeling man-machine systems is identifying and accounting for the effects of feedback mechanisms and reinforcing factors which act to facilitate human performance and cause our overall estimates of human reliability to be grossly underestimated. Laboratory studies of individual tasks to obtain performance data typically have not considered such facilitating factors.

We also deal at length with display and control design "principles" with little or no information on the sensitivity of human performance to such design guidelines. The human operator (or maintainer) utilizes all sorts of cues in the actual system environment, many of which are subtle and not directly involved in the specific tasks he is performing. Such cues are used to guide his task performance and thereby reduce his error rate and to detect and retrieve a certain proportion of the errors he does make prior to effect on the system. Simply stated, we need to know more about human behavior in specific system environments, and derive a means for identifying and quantifying the real effects of performance shaping factors which interact to facilitate human performance.

Hopefully, such studies would result in a set of principles or modeling guidelines which could be generalized at least within a particular system context. We might also be able to identify certain classes of tasks which tend to interact in a particular performance-shaping manner. At any rate, conduct of such studies is considered to be a necessary next step to evolving the science of man-machine modeling.

B. COLLECTION OF HUMAN PERFORMANCE DATA

We have indicated that there is a paucity of human performance data applicable to man-machine models. Data for operability models are almost totally derived from laboratory research while data for maintainability models are almost totally derived from field studies. And very little of both sets of data have any real value.

As everyone knows, the basic problem of obtaining field data is cost. There is no getting around the fact that the accumulation of sufficient quantities of empirical data would be very costly. Unfortunately, available "short-cuts" for reducing costs do not appear justified as quality of data will be unacceptably reduced.

In view of the traditional difficulties in obtaining funds for collecting empirical data, and the apparent fact that they will not likely become available in the near future, we have, for some time now, stressed research on subjective judgments of experts as potential means for generating relatively inexpensive, large amounts of data. We have thus far conducted three studies in which we collected human reliability or maintenance data (Blanchard, Mitchell and Smith, 1966; Mitchell, Smith and Blanchard, 1967; Smith et al., 1970). The judgment technique used to obtain reliability data was the paired-comparison method which generates an interval scale of z score values. Thus, the scale had to be transformed to a reliability scale with the use of a hypothesized transform function and two empirically derived data points. With respect to the maintenance data, transformation was not necessary since absolute judgments were required. Between judge agreement was relatively low for the reliability data, while it was extremely high for the maintenance data.

As is so often the case, funds were not available for validating the judgment data. It cannot be said, therefore, that a major step has been taken in investigating the potential utility of such data. On the other hand, considerable knowledge was gained which can be applied to future studies.

All things considered, we feel very optimistic about using subjective techniques for obtaining equipment maintenance times, but somewhat pessimistic about using them for obtaining reliability data. Since the judgment task involved in the former seems to be far easier to judges than that in the latter, it is perhaps not so surprising that our feelings are so directed. However, a great deal more work needs to be done before any decisive conclusion can be reached.

Complex problems are seldom resolved as a result of conducting a few studies. Even though we were not able to perform validation work, which theoretically represented the prime basis for evaluating our data, the results of each of our studies revealed unanticipated findings which facilitated the design of subsequent studies. It is important, therefore, to recognize that the judgment task and process must be thoroughly explored systematically before we can meaningfully accept or reject the technique. We suggest that a series of well-planned studies whose designs do not call for validation efforts, could nevertheless provide a "model" of the judgment process. Once we understand how and why an expert will give specific judgments in a reliable manner, we will then be in a position to determine the ultimate potential of subjective techniques. Of course, validating the data of each study would be highly facilitating in that the direction and magnitude of judgment errors could be determined. However, the point here is that it is apparent that some fundamental work needs to be done, independent of validity considerations. For example, factors in need of investigation include experience level of judges, judgment techniques, instructions, various voluntary and involuntary judgment biases, etc. All of these factors and more are known to affect significantly an individual's judgment. Clearly, they must be systematically explored and related.

In summary, we have reluctantly concluded that the human performance literature is devoid of significant quantities of valid and relevant data for use in man-machine models. Moreover, current research trends do not appear particularly optimistic with respect to facilitating this state-of-affairs. Since field data collection studies are costly and will probably be performed only infrequently, it would appear that the use of judgment techniques is the on!y untapped and potentially practical recourse. In view of the possible rewards that may derive, it would seem illogical not to devote appropriate attention to such techniques.

REFERENCES

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.

Chapanis, A. Research Techniques in Human Engineering. Baltimore: The John Hopkins Press, 1959.

Chapanis, A. The relevance of laboratory studies to practical situations. Ergonomics, 1967, 10, 557-577.

Harkins, J.A. The real world of system effectiveness. Annals of Assurance Sciences, 1969. Proceedings of Reliability and Maintainability Conference, New York: Gordon and Breach Science Publishers, 1969.

Irwin, I.A., Levitz, J.J. and Freed, A.M. Human reliability in the performance of maintenance. Proceedings of the Symposium on the Quantification of Human Performance, University of New Mexico, Albuquerque, New Mexico, 1964.

Karush, W. On mathematical modeling and research in systems. System Develop-Corporation Rep. SP-1039, Santa Monica, California, November 1962.

McKendry, J.M., Corso, J.F., Grant, G. and Scheihing, F.A. An experimental investigation of equipment packaging for ease of maintenance. U.S. Naval Training Device Center Tech. Rep. NAVTRADEVCEN 330-1-1, Port Washington, New York, April 1960.

Meister, D. Methods of predicting human reliability in man-machine systems. Human Factors, 1964, 6, 621-646.

Meister, D. Use of a human reliability technique to select desirability design configurations. 1969 Annals of Assurance Sciences: Proceedings of the Eighth Reliability and Maintainability Conference. New York: Gordon and Breach Science Publishers, 1969.

Mitchell, M.B., Smith, R.L. and Blanchard, R.E. Test application of TEPPS on a Navy CIC subsystem. Dunlap and Associates, Inc., Santa Monica, California, August 1967.

Mitchell, M.B., Smith, R.L. and Westland, R.A. Development and test of a technique for establishing personnel performance standards (TEPPS): Phase IV - final report. Dunlap and Associates, Inc., Santa Monica, California, August 1967.

Munger, S. J., Smith, R.W. and Payne, D. An index of electronic equipment operability: data store. American Institutes for Research Rep. AIR-C43-1/62-RP(1), Pittsburgh, Pennsylvania, January 1962.

NSIA (National Security Industrial Association. Report of NSIA Maintainability Prediction Task Group. April 1967.

Payne, D. and Altman, J.W. An index of electronic equipment operability: report of development. American Institutes for Research Rep. AIR-C-43-1/62-RP(1), Pittsburgh, Pennsylvania, January 1962.

Pickrel, E.W. and McDonald, T.A. Quantification of human performance in large complex systems. Human Factors, 1964, 6, 647-662.

Quade, E.S. Pitfalls in military systems analysis. Paper presented at the Electronic Industries Association Symposium, San Francisco, California, November 1962.

RCA Service Company. Maintainability measurement and prediction methods for Air Force ground electronic equipment (Phase II report). Rome Air Development Center Tech. Rep. RADC-TN-60-221, Griffiss Air Force Base, New York, September 1960.

RCA Service Company. Maintainability measurement and prediction methods for Air Force ground electronic equipment (Phase III report). Rome Air Development Center Tech. Rep. RADC-TN-61-141, Griffiss Air Force Base, New York, 1961.

Retterer, B.L., Griswold, G.H., McLaughlin, R.L. and Topmiller, D.A. The validation of a maintainability prediction technique for an airborne electronic system. Aerospace Medical Research Laboratory Tech. Rep. AMRL-TR-65-42, Wright-Patterson Air Force Base, Ohio, May 1965.

Rigney, J.W. and Bond, N.A. Maintainability prediction: methods and results. Department of Psychology Tech. Rep. No. 40, University of Southern California, Los Angeles, June 1964.

Rigney, J.W. and Hoffman, L.S. Human factors research in electronics maintenance: An analysis of recent trends, with some suggestions for the future. Department of Psychology Tech. Rep. No. 35, University of Southern California, Los Angeles, July 1962.

Schaeffer, K.H. The logic of an approach to the analysis of complex systems. Air Force Office of Scientific Research Rep. AFOSR-2136, Washington, D.C., April 1962.

Shapero, A., Cooper, J.I., Rappaport, M., Schaffer, K.H. and Bates, C. Human engineering testing and malfunction data collection in weapon system test programs. WADD Tech. Rep. 60-36, Wright-Patterson Air Force Base, Ohio, February 1960.

Siegel, A.I. and Wolf, J.J. A technique for evaluating man-machine system designs. Human Factors, 1961, 3, 18-28.

Smith, R.L.,Westland, R.A. and Blanchard, R.E. Technique for establishing personnel performance standards (TEPPS): Technical manual. Personnel Research Division Rep. PTB-70-5, Vol. I, Bureau of Naval Personnel, Washington, D.C., December 1969a.

Smith, R.L., Westland, R.A. and Blanchard, R.E. Technique for establishing personnel performance standards (TEPPS): Procedural guide. Personnel Research Division Rep. PTB-70-5, Vol. II, Bureau of Naval Personnel, Washington, D.C., December 1969b.

Smith, R.L., Blanchard, R.E. and Westland, R.A. Subjective judgment as a means for generating corrective maintenance time data. AMRL Tech. Rep., Wright-Patterson Air Force Base, Ohio, 1970 (in press).

Smith, R.L. and Westland, R.A. The status of maintainability models: a critical review. AMRL Tech. Rep., Wright-Patterson Air Force Base, Ohio, 1970 (in press).

Swain, A.D. A method for performing a human factors reliability analysis. Sandia Corporation Rep. SCR-685, Albuquerque, New Mexico, August 1963.

Swain, A.D. THERP. Sandia Corporation Rep. SCR-64-1338, Albuquerque, New Mexico, October 1964a.

Swain, A.D. Some problems in the measurement of human performance in man-machine systems. Human Factors, 1964b, 6, 687-700.

Swain, A.D. Some limitations in using the simple multiplicative model in behavior quantification. Sandia Corporation Rep. SCR-68-1697, March 1968.

DISCUSSION OF BLANCHARD's PRESENTATION

Jenkins - I'd like to ask Bob a couple of questions. What I wonder about is the approach. You begin with a data bank analysis or other kind of human performance analysis which requires a molar approach, because of the tremendous difficulty in getting specific task associated data. I question whether from the approach you use results in specific recommendations related to design or whether you can derive specific and reasonably good predictions, because more detailed studies would have to be made. I wonder whether your recommendations for analyzing rater reliability and the validity of judgement have not already been done so, that we could look at ratings or judgements now and say this technique will buy us this and so for valid assessments of performance.

Blanchard - My reaction to the molecular/molar question is more in terms of the research that is being done. That is, I don't think we're going to get a great deal of payoff in the long run looking at just one or two or three tasks. I think that the research has to be done at a more molar level.

Meister - Doesn't this automatically mean, though, that we're forced largely to go into the operational environment or some sort of reasonably faithful simulation of cperational environment? This again puts the bite on the military.

Mills - In this research that I am doing I set up what I like to call a molecular strawman. I feel that before we reject work at the molecular task level we have to demonstrate that it's not relevant particularly in the operational environment, and that we don't need it, really before we can go ahead. What I feel should be done is that these tasks need to be built up from the task element level and/or molecular level. Set these things up as strawmen and knock them down with research as you go up and try to determine what task level can we really deal with and would be acceptable? Now this was the approach that I decided to take and I may be wrong and I may be wasting a lot of time. I could perhaps really do the same type of thing by immediately establishing a higher order task and working at it at that level.

Blanchard - I didn't mean to imply that we could jump in a molar level. I think what Bob is doing is fine. It's great that we're beginning to worry about research specifically designed to investigate this problem. I think that sooner or later, when you get to the point of your study where you are beginning to get feedback, you are going to automatically evolve to higher order tasks. Your point of establishing a series of hypothesis which you are testing is great. It seems to me that one of the things that you're also saying there is that you must first build your strawman in order to tear him down.

Connery - In science negative results are often times as valuable as positive results. One of the things that occurs to me is that this kind of approach today might be better invested in the academic community because in the military you can't afford it.

Meister - When we start spending what little money we've got today to get negative findings, we're out of business.

Mills - Here is the dilemma and it is that a certain amount of research is required. Everybody says this, everybody agrees with this. These guys need money and we don't have very much money, so we go for the high payoff immediate type of item, we say data bank now. They can't develop a valid and applicable data bank without having the research. This is our dilemma. So I say, okay until somebody either performs the basic research or research for thes. kinds of questions, until our government starts funding these projects, we're going to have to do it on our own. They can't have a data base that they can really operate with unless we're able to establish the funding environment for them to develop an operating data base. The Human Engineering Division of our medical research laboratory is supposed to serve this function. We are supposed to be a research laboratory. In fact, we have a task which comprises about thirty or forty percent of our budget and don't hold me to that but it's a sizeable task which is devoted to basic research.

Connery - The only point I wanted to try to make here was that don't forget that we're concerned with advanced development, not research. In this area we've got to produce.

Coburn - I agree with much of what you said but I question this scale which you mention about finding the payoff. I think we should look to an earlier payoff. If we can begin to do something practical, let's begin to do it soon.

Jenkins - You've given essentially a point of view which is in variance with what others have been saying. You said by going to a molar approach, by taking judgement as the means by which we assess task performance, we would be able to derive effective predictions and measures of human performance and reliability. Now, if we ought to go that route, we are forced, because of time alone, to ignore almost any other technique, because it would require the resources available. This worries me because we have to obtain performance prediction and design requirements.

Blanchard - As I tried to explain, my concept of molar approach relates to type of experimentation. In fact, an approach like Bob's sooner or later will reach that. I'm not suggesting that we necessarily have to take a molar approach right now.

Jenkins - I think we probably have to deal at a somewhat higher level than the task elements.

Meister - Bob, aren't we being a little bit overly pessimistic. I can agree with everything you say, because we have a dismal lack of knowledge and therefore we need more data and we need more research. Nobody can disagree with that, but I get the implication from what you say is that in effect we're at time zero in this kind of work, and I question whether this is actually the case, because we really have not attempted to assimilate all of the work that has been done in the past to be able to say whether we are at time zero or time 10 or 20 or what have you.

Blanchard - That's what I've suggested by my task of looking at what we have and I'm trying to define what it consists of.

Meister - That I strongly agree with.

Blanchard - What I'm saying is that we've got to get out of the mental set that we've been using for a number of years. We've got to begin to acknowledge that there are basic problems that sooner or later must be solved before we are going to achieve any significant progress. Obviously, there has to be an interim approach. We have to utilize what we have now, but we can't continue with that level. We can't continue to collect tiny pieces of human behavior and try to fit them together.

Meister - What you're saying in effect is what we need is a systematic way of attacking the problem.

Blanchard - I think what you're suggesting in your paper, Dick, is an excellent interim approach.

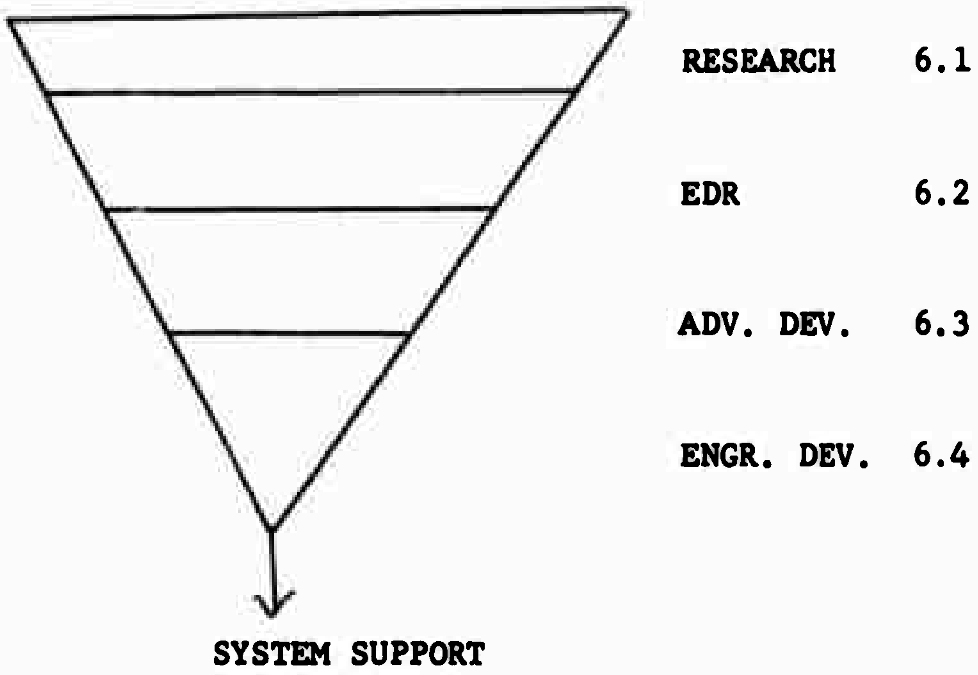


FIGURE 1. SIX POINT MONIES

Connery - In the six point monies, we're funding in research, here's our big area right up here, research. Theoretical philosophy is that we come down this track, that we learn and discover up in here we advance on this area, those areas in which indicate the military payoff. What we're talking about here today is in the advanced development area where we are supposed to, at this level try to capitalize, on what we have learned up here in here in terms of technology and take those, the best of what may be available and try to answer some questions. Then with what comes out of this, most promising part, we bring into engineering development. One of the difficult things about this whole pyramid is that in the definitions of these it's equipment oriented, purely a hardware type definition. We do not have a comparable set of definitions for software, for behavioral science, human factors, we don't have comparable definitions.

Blanchard - My point is, fine, within the concept of advanced development objectives you're looking at what we have on the shelf trying to take it, optimize it and use it, I would propose that.

Connery - Then you're saying, let's stop right here stand back and take a look at what we've got.

Blanchard - That's right. Let's don't keep thundering down the paths that we have been pursuing during the past ten years. Now at the same time we're doing that, somebody has got to find some money and do something about evolving our methodology, evolving the state of the art through 6.1 or 6.2 or whatever. Without some research coupled in at some point in this program, we're not going to get very far.

Connery - It's hopeful now that Dr. Fields, and Dr. Tolcott here with some of the development going on in the field of mathematics will be able to start up here and invest a little more effort in the human factors area.

Blanchard - What I would be validating are not the judgements but the techniques. I would not propose to validate the data each time that technique were applied, once the utility of the technique had been established.

Meister - That seems to be the name of the game to look at the various approaches and decide where we are lacking, where they stand in terms of potential payoff and proceed from there. This would be an excellent idea but in the behavioral sciences area we get practically none of this sort of decision making, if you want to call it that. Frankly, in the case of human reliability, human performance qualification, it means that you have to be brutally frank about the money wasted in the past and be brutally frank about your inadequacies and unfortunately governmental customers don't like to admit their inadequacies, just as we are reluctant to admit ours. That would seem to be the way to go. At the same time I have a feeling that our situation with regard to limited goals, not necessarily mathematical modeling or human performance quantification in their totality, but limited goals of predicting a limited set of task situations, we're probably not that badly off. Although we probably don't really know how well, or how badly off we are. Would anybody care to comment on that?

Swain - About two weeks ago some man called up from Omaha and he wanted to determine what the probability of human error was in certain aircraft operations in Navy airplanes dropping "X" and I was able to tell him that. We had data we could give him, namely, predictions of the probability that the "X" we deliver to the Navy would do everything it's supposed to do, assuming zero human errors in the future. I was able to tell him that had the military some system like the one I'm recommending here and put all this data that is available in some usable form than he could have answered his question or gotten somebody to answer the question for him. That's the thing that bothers me and why I'm recommending that we start a data store pretty quickly because the data is available, it's just being lost all the time.

Jenkins - I don't want to cut off the discussion, if there are more questions. Tomorrow we're going to present the program which we've started. You've each presented your ideas on which way an organized program in reliability should go and I'd like you to review what you hear in the light of your particular orientation and opinion. If there is concurrence on the part of the group here, as to any changes or modifications, I'm sure that they can be made. It is obvious that we don't all agree. I think there are certain key points. Is the approach that Dave Meister presented the model which we should take for the next five years? Should we take something like the data bank method which Alan Swain has suggested, regardless of the problems of task taxonomy? Contrast that with what Bob has said as compared to Dave's position and Alan's position, tying in with whether we should go into system design analysis as Regulinski was talking about. This is the approach that I would like to take tomorrow in terms of being helpful to us. We need some structure to our organization in this program.

Meister - It would be very illuminating. Not that I think we have the time to do it but it would be illuminating to try to determine those things that we actually do agree on as I indicated to you some months ago, I think that if we were all of us to get together and try to determine those things that we actually agreed upon, you would find a very substantial body of agreement, because there is not really a critical difference, say, between Blanchard's point of view and my own, or between my point of view and Alan's. We disagree perhaps on the detailed strategy, how to implement certain things, but I have a feeling that in terms of basic assumptions that we do agree, we may even agree on certain basic definitions and even God-help-us certain basic approaches. The details will differ, so I think that until we have a chance to examine the commonality of interest, you won't really know whether this type of a conference could be productive or not.

Mills - I think that the biggest problem here is simply scoping the effort to get tc the final product in 3 or 4 or 5 years.

Jenkins - We do have to focus on all the things we agree on for this one program. It has large ramifications as you were saying. At the moment I think that comes as the number two problem.

Coburn - There would be some advantage in having more time after all the presentations to see what is in common, what do we modify and what we have scheduled here isn't enough time.

Jenkins - The idea was that after the last paper we'd have this discussion, both in terms of the specific programs which we're talking about as well as the areas of concern. The summary and conclusions are to wrap things up.

Swain - I would be willing to, when we get back to our shop just briefly write our comments on various parts of the papers. I've done that already in rough notes. Some of my comments I had in the margin I found out are irrelevant or misinterpreted.

RELIABILITY AND INDEPENDENCE

Jerry C. Lamb

Naval Underwater Systems Center

RELIABILITY AND INDEPENDENCE

Jerry C. Lamb
Naval Underwater Systems Center
New London Laboratory
New London, Connecticut 06320

Introduction

This project is concerned with the testing of various assumptions about the independence of a series of sub-tasks which comprise a total maintenance task. It also considers the effect of the violation of this assumption on the models which have made these assumptions. The purpose of this is to determine at what level and how the human reliability data should enter in the overall system effectiveness equations.

The Navy's PAU model of system effectiveness has three elements (NAVMAT, 1967) Performance, Availability and Utilization. That is

$$E_S = f(P,A,U) \qquad (1)$$

The availability of a system to perform its mission is the greatest determiner of overall system effectiveness after design is complete. The factors which determine the availability are the reliability of the equipment and man modules. A method is needed for predicting the reliability of the human which is equivalent to that of the equipment reliability. "The factors associated with the man-module(s) in the systems are not now quantifiable; they must be quantized(...) and indexes or figures of merit must be used to synthesize availability." (NAVMAT, 1967, p. B-9). To this end, a large amount of work has been done on developing models to be used in predicting human reliability, particularly in the critical maintenance area.

A common assumption of these models has been the independence of some level of the elements[1] which make up the tasks being analyzed. It is generally recognized that this assumption is not strictly true but is convenient because:

(1) it simplifies the development of the probability trees used to predict reliability, and

(2) the available data base does not contain precise information even on independent tasks let alone on conditional task performance.

[1] For this paper we are concerned with the reliability of the operator, not the time element consumed. For recent research on time to failure for the operator see Askern & Regulinski, 1969 and Siegal and Miehle, 1967.

Similar assumptions are usually made for the independence of one operator performing the same task repetitively, or for two operators working together. One of the purposes of this research is to test the independence assumption and determine how well it holds at each task level.

A second characteristic of the models is that they have moved consistently from fine-grained analysis to a more molar view of the operator's task. This step toward a more molar view is necessary if we are to eventually have a model which will allow predictions early in development and yet will be flexible enough for revision during system design. Therefore, another purpose of this study is to determine which level of the proposed type of model (molar or molecular) most adequately fit the data and where the deviations from predicted performance occur.

The following two sections discuss relevant research in each of these areas, while the remaining sections describe aspects of the proposed research.

Independence of Tasks:

The assumption of independence of tasks, i.e., there is no interaction of tasks, means that performance (success or failure) on one task is not affected by performance on a second task. Mathematically,

$$P(A \text{ and } B) = P(A) \cdot P(B) \qquad (2)$$

or

$$P(A/B) = P(A) \qquad (3)$$

Equation (3) is the product rule which is normally used in determining the reliability of a hardware system (Pieruschka, 1963). The probability of completing all of a series of tasks is simply the product of the probabilities of each task taken separately.

Meister (1964) has noted that multiple interaction (common in large systems) may produce "significant changes in either direction from a hypothetical 'average' error rate." (p. 643). He gives examples to illustrate the type of situations which could lead to changes both to reduce and increase the error rate. It is generally recognized even by those who make the assumption of task independence that, as Meister points out, the assumption may not be true. Several alternative approaches have been worked out. Basically, they fall into two categories, corrections to the product rule for repetitive tasks or collection of conditional probability data to more adequately model each task.

Swain (1963) in discussion of the THERP technique indicates that for low time stress tasks (those where time to repair is not critical) p(a) on trial 2 can be obtained by squaring P(A) on trial one. If the task is high time stress, $p_n = p_1 \cdot 2^{(n-1)}$, $p_n \leq 1.0$.

For the case of two operators working together, the redundancy reliability, R, is (Meister, 1964)

$$R = \frac{1 - (1-R_1)\ (T_1) + R_1\ (T_2)}{T_1 + T_2} \qquad (4)$$

where R_1 is the reliability of a single operator and T_1 is the percent of time the second operator can observe the first and T_2 is the percent of time left. Swain (1963) notes that the reliability of the second operator in detecting the first operator's failure is about 0.85.

Swain (1967) has shown that the THERP model can be used to predict system reliability if the appropriate conditional probabilities are known. He notes that the product rule can be used safely only if the interaction effects are small or large errors in prediction can be tolerated.

Therefore, while the nature of the interaction problem has been well known and various attempts made to correct for it, the degree to which it occurs in typical maintenance operations is not known. Another question remaining to be asked is where does the interaction occur, at the part task (molecular level), the whole task (molar level), or both. The answer to these questions have implications for model development and for the number of terms necessary to describe the human in the system effectiveness equations.

Approaches to Reliability Prediction

The first approach to the quantification of human reliability was the data store approach (Munger, et al, 1962). The data store consists of task-oriented reliability and time data for a variety of equipment operations such as length of toggle switches, angle of throw, etc.

The basic procedure for using the data store is to perform a task analysis on the equipment under consideration, break down each operation in the task analysis according to data store terminology, and multiply the reliability values from the data store to obtain task reliability. Because the task analysis must be so fine grained, this may be referred to as a "molecular" approach.

Swain (1967) in analyzing the data store for Monte Carlo simulated tasks found that the average reliability score (of all the data score) could be used with little loss of final accuracy. Given the independence assumption and a typical task requiring approximately 40 or greater steps, the failure for any task for the task would be approximately

$$Q = (.001)N \tag{5}$$

Therefore, all that is required is to count the number of steps detailed in the task analysis. No conversion to explicit data store terms and table look-up is required.

However, in the same article, Swain argues for a more molar approach oriented around tasks as opposed to task elements. Siegel and his colleagues have approached the development of a molar model from a different viewpoint than that of building up a task data store. They derived through factor analytic methods, average reliability scores for electronic technicians on nine types of job activity. Computation of task reliability consists of:

(1) Task analysis only to a level to specify the task type, and,

(2) combination of the appropriate task type probabilities.

This approach seems to have several advantages, including

(1) being usable before selection of all hardware components,

(2) requiring less task analysis effort, and

(3) requiring a relatively small data store.

Both the molecular and molar approaches to prediction utilize the product rule. One facet of the present research will determine which approach makes a better predictor when if the assumption of task independence underlying it have been violated.

Experimental Plans

In order to test the assumption of independence between tasks, an experiment will be conducted using a sonar system simulator. There will be three maintenance tasks involved, either separately or in combinations of two and three tasks. The failures will occur in the display (BTR) portion of the system. S will be required to

(1) diagnose failure symptoms,

(2). locate fault, and

(3) repair the fault.

Selection of faults will be taken from the results of an extensive reliability analysis done on this system during test and evaluation. This same test and evaluation data will also be used to validate the reliability levels obtained during the experiment.

Each S will serve as his own control and will be in 7 experimental conditions; 3 with one fault each, 3 with 2 faults, and one with 3 faults.

S's will be allowed to repeat task portions in order to simulate realistic procedures. S's will be ET's. A complete record of all S's activities will be kept.

Independence Analysis

Two measures will be computed for each condition:

(1) (R) reliability $= \dfrac{\text{number of successes}}{\text{number of attempts}}$

(2) Time to complete task.

If the assumption of task independence is correct, then

$$R_{TOTAL} \text{ (Overall Reliability)} = R_1 . R_2 . R_3 \qquad (6)$$

If $R_{TOTAL} \neq R_1.R_2.R_3$, the paired comparisons will be analyzed to determine the nature of the interactions. Also, conditional probabilities will be developed from the pairwise data and if there is no 3 way interaction, R_{TOT} will be the product of the conditional probabilities.

Note that the assumption of task independence may extend over many levels, that is $P(A/B) \neq P(A)$ is one possibility, $P(A/B \text{ and } C)$ is another, etc. While this can be carried to many levels, it is reasonable to assume that the most potent determiners of behavior on the task under consideration will be those immediately preceding it. Therefore, the present research allows a test of first and second order independence by using single, pairwise and triple combinations of conditions.

Model Analysis

Predictions will be made using the molecular approach with the assumption (Swain, 1967) that each task element can be represented by the average of all data store points (.9990). A similar prediction will be made using Siegel's (1967) reliability scores for Navy ET's.

The two analyses will be compared for every level of task combinations against each other and against the collected data. Analysis will be made of any discrepancies.

Potential Problems

At present, there appear to be two problems which could affect the results.

First, a relatively small n (compared to total number of systems in use) is necessary because of experimental conditions. This could lead to incorrect estimation of reliabilities, this is not expected to be a serious problem since there is test and evaluation data for comparison.

Secondly, the specific tasks chosen could determine the amount of interaction. Task selection will be done with the aid of personnel who conducted the reliability tests in order to minimize this difficulty.

Expected Results

The data obtained should point the way toward determining the level at which human reliability estimates can be entered into the reliability estimates for a system, e.g., task-element, task, or subsystem. Also, the degree to which predictive models will have to incorporate task dependencies will be evaluated. The manner in which the quantification of human reliability, both measured and predicted, should be included in the system effectiveness equations will be determined.

Finally, a comparison between laboratory estimates of reliability and system test elements will be evaluated.

References

1. Askern, W. B. and Regulinski, T. L., Quantifying Human Performance for Reliability Analysis of Systems. Human Factors, 11 (4), 1969, p. 401-406.

2. Meister, D., Methods of Predicting Human Reliability in Man-Machine Systems. Human Factors, Dec. 1, 1964, p. 621-646.

3. Munger, S. J., Smith, R. W., and Payne, D., An Index of Electronic Equipment Operability: Data Store AIR-C43-1/62-RP(1). American Institutes for Research, 1962.

4. Naval Material Command, Navy System Performance Effectiveness Manual, May 1967.

5. Pieruschka, E., Principles of Reliability, Prentice Hall, New Jersey, 1963.

6. Swain, A. D., Altman, J. W., and Rook, L. W., Human Error Quantification: A Symposium Report SCR-610. Sandia Corporation, April 1963.

7. Swain, A. D., Some Limitations In Using The Simple Multiplicative Model In Behavior Quantification. In Symposium on Reliability of Human Performance In Work. Askren, W. B. (Ed.) May 1967.

DISCUSSION OF LAMB's PRESENTATION

Blanchard - One typically looks at time measures where the primary systems effectiveness measure is Mean Time to Restore (MTTR). If you can specify precisely a best, a most correct troubleshooting route then you're able to score this cat in relation to whether or not he is correct or incorrect. This hinges very heavily on your being able to say that there is clearly a best route.

Lamb - There may be in fact a good technician who can take a shorter route.

Blanchard - There may be several good routes. This is what happens in maintainability because the nature of the troubleshooting process itself is highly complex. People should be given all the feedback information that's available - not just purposely disregard all feedback information.

Lamb - We are measuring the time, yes.

Blanchard - The reason is that time often turns out to be the only usable measure of our performance.

Siegel - I would almost say offhand that when you get involved with time as a measure of performance, you've got trouble, no matter what you're measuring, you've got trouble once you start with time. If someone tries to tell me I'm interested in anything, and I'm using time as a performance measure. I say, "Oy vey, you got trouble fella."

Blanchard - We may have to accept that trouble, Art.

Mills - Unfortunately, we work with time.

Meister - In terms of the limited objective that this study has, I wonder if we're not attempting to overcomplicate it by emphasizing the fact that maintenance is, in reality, such a horrendous kind of thing. Basically, you're doing essentially what Bob Mills did with his more laboratory oriented task. It seems to me that if what you're attempting to do is to determine whether a performance time or error prediction for an individual task will, when combined with a prediction for other tasks, predict the overall performance. There really shouldn't be that much difficulty, should there?

Blanchard - I'm not sure that he's got the best context to study the problem. There is a difference between what Jerry is doing and what Bob did, because Bob was controlling, he was able to control the dependent variable. This is confounded within variables.

Lamb - Well in one part, in the repetition it's confounded, but in the task it is not, because it was 1,1,1; 2,2,2 and 3; so I have measures of each contingency.

Blanchard - Yes, but within any one task, and I'm not familiar with the equipment or what the guy has to do, conceivably there could be a lot of little feedback loops and cues that could be given to him.

Lamb - Yes and the one part is the testing of the independence of repetition. The testing of the independence of the task, I feel is fairly straightforward because I am measuring each of them and all the possible pair combinations, all the possible signal combinations, and then total.

Mills - It's just that the kind of things that you're dealing with, the nature of the game is difficult.

Lamb - It may be that in fact I cannot effectively given the experimental constraints, measure repetition but this was a followup of the way that we designed the study because we want to allow them to do essentially as they would in the real world. We very carefully separated out the tasks so that we have these independently and in combination and we're hoping to get the repetition test also. If we don't then I will feel badly that it is not the major purpose. The major purpose is to test the task in dependence, because that's the way we want to combine the data. We want to say that if he did this repair and this repair and this repair can we just put them altogether? Bob Blanchard suggested we can't or we have to find some rule for modifying our independent probabilities and that's the basic purpose that we're looking at.

Mills - Are they going to know that this is the standard procedure? He's experienced and he's likely to deviate from this procedure. This is saying that his expertise is really being used against him in terms of math modeling procedures.

Lamb - It is being used against him in one sense, because at the molecular level that's supposedly the best procedure. There may in fact, be a better procedure but that will have to be judged.

Swain - In the total sense it won't work against him, because (a) he'll get it right, and (b) he'll also get it right very quickly. This is why I was suggesting that you might want to look at this other technique for grouping, in troubleshooting we've used it by asking them to voice what they were doing.

Mills - Well, you might bring each one in and somehow establish a baseline for best performance among, say if you're going to use 10 subjects, you'd have all 10 come in.

Meister - I don't think it's really necessary. I'll tell you why. After the man goes through the procedure and if has deviated and you can't tell by observation why he has deviated you can always ask him, "Why the hell did you do it, this way?", and he will tell you. This is something that we found out by our observations of maintenance at SAC bases, he will tell you the procedure is not updated and so I used my own version. That is not really an error. You can exclude that.

Lamb - You're talking about a whole other experiment just to get the baseline data to work when it has been done in terms of operating procedures.

Swain - Baseline data is the most important thing that you can collect to provide the criterion to judge here if the rest of your results are mixed.

Lamb - On this one piece of equipment there has been an enormous amount of baseline data analysis done. More than 1 can do experimentally. Now if I do find my technicians deviating during the experiment, I have to do something like Dave suggested. I've got to revise my scoring procedures.

Meister - You are going to talk to these people after the experiment?

Lamb - Oh yes, I'm not going to say, "Come in, do it and go home, no".

Mills - Another alternative would be to establish your criterion procedures, your procedure as a criterion and simply deal with those people who do not bow to the procedure which leads to poorer performance than that. I don't mean eliminate the analysis or whatever.

Meister - He's already got a criterion procedure: the written procedure is the criterion. A man can vary from this; if he varies because he has found a more efficient way, and you can determine this, then he can still use his data.

Mills - Here is another alternative too, aside from what I was suggesting about the baseline, and that is what Dave may be saying. You don't have to analyze the data until after the study is over so you can modify your procedures then if you want to. I'd like to ask another question. How many steps are we talking about generally to location?

Meister - The number of procedural steps?

Lamb - I don't really know I mean I haven't gone through any kind of an audit - around ten or twenty - fairly short and I'm going for fairly easy failures.

Mills - What I was getting at was, is the number of steps equated across conditions?

Lamb - I have a choice of possible failure modes and the final decision on which ones to use depends upon which ones tend to fail, typically I don't care if they perform exactly because we're measuring within each subtask or each fault.

Meister - Measuring within the maintenance procedure.

Tolcott - Jerry, would you be able at the end of this to kind of validate your results against some other piece of equipment? You're really trying to find out whether the combinatorial rules that you get here could be applied to some other piece of equipment which contains some of these steps.

Lamb - Assuming we collect the data with all the experimental difficulties that we talked about today. There are two things to do. One is obviously to find out whether it holds, if it holds then we can feel strength in what we're doing. If it doesn't hold, we want to see where is the grossest violation, molecular or the molar level? Successive failure on one task destroys the next task, effects the next task very badly. Then, I would like comparison of the kind of data that we get in the lab with the so called operational situation. This is the best way to start answering the questions which you're asking. If they seem to be the same as in the laboratory then you can do two things. You can go back to the laboratory which is infinitely more controllable and measurable than the kind of thing I'm trying to do. Also we can start to be sure that we have some set of rules for combining.

Tolcott - I would be surprised if you came up with the same kind of rules for maintenance tasks of this kind. The ultimate test of what you're doing might be the ability to see whether it calls for another kind of maintenance.

Meister - You're suggesting that the combinatorial rules may be specific to at least a class of equipment that required a certain class of behaviors.

Mills - I think a possibly more realistic hypothesis would be that it's got something to do with the redundancy, the number of redundant steps in performing a task.

Meister - This is certainly the kind of study that we need to have done.

Siegel - It's the only kind of study, if we're ever going to get any operational people to accept what we've done.

They're going to say, "That's great, all these nice elementistic data which are collected all over in your nice shiny labs and so on, but does this correlate with something that looks like real life?" Obviously when you get closer to real life you get closer to situations which are difficult to control, difficult to collect data in, have much more non-predictable variance, much more non-controllable variance, but that is the situation you're in. We find, for example, that much of maintenance time is just attributable to things you would never predict by looking at a job. For example, we found in one case, that much of maintenance time was accounted for by the amount of time it takes the technician to find the key to the spare parts cabinet, because the key to the spare parts is kept by the chief and the chief is only available 8 out of 24 hours on a watch situation. The technician has to walk out of the maintenance room, down to the chief's quarters, pick it up, and come back to the job. These things were accounting for a lot of his time. This is the real world.

Meister - Of course, you're concerned, even in the operational situation, with just the time actually spent exclusive of administrative down-time. This is the reason why you can't use environmental timing.

Siegel - We're predicting MTTR, and we come in and we say, "15 minutes boy! this is wonderful." Now equipment gets on the ship and the Captain comes out, we tell him - 15 minutes MTTR this is the greatest piece of machinery you've ever gotten Captain and three months later he says, "I just had a malfunction down there and it took this technician 45 minutes to complete the repair. What's with this 15 minutes?" You explain to him. He now feels that we have let him down.

Regulinski - I sympathize with the Captain. I would not want to be given the mean time if I were on a ship. If I were, say, on a submarine and developed a bad control system, I would not want to know how long, on the average, it would take to repair the system. Rather, I would want to have some reasonable assurance that the system can be fixed within some maximum time. That is all I would want to know. Not the mean, but rather the maximum repair time would help me decide better whether to hold off the enemy or to make a run for it.

Tolcott - If the Captain wants information like that he's going to have to tell you where he keeps his spare parts, what the organization of the ship is, and how you are going to get there, if that's what he wants to know.

Meister - I don't really believe that the Captain is so stupid that he doesn't understand that there are such things as administrative delay times. I don't consider that to be a crucial factor.

Regulinski - Jerry, a brief comment that has nothing at all to do with the Captain and his problems. The Air Force systems C5A and the F-111 faced pretty much the same repair dichotomy dilemma. The repairs, whether corrective or preventive, I suppose, are analogous to what the Navy is facing. From experience they have learned that breaking repairs down to the type of levels that you, Jerry, are talking about is really a horrendous task, In modeling repair, what they ultimately did was to assume what Dave was earlier talking about, namely, the gross task approach. All they are asking now, is, how long will it take to repair the system. This obviates the blasted conditional probability problem concomitant with task dichotomy, because when time measurements are taken in the continuous time domain, the conditional human responses manifest inherently in the time recorded. In short, questions asked involving the denumerability in time-discrete domain are not generally relevant in time-continuous domain because of the non-denumerability of the random variables. In time-discrete domain it is relevant to ask what is the probability of task A occurrence, and jointly the probability of task B occurrence, whether or not conditional probabilities are involved. In space-time-continuous tasks, neither time nor tasks are denumerable. Whether the C5A and the F-111 systems repair modeling experience will succeed, time will tell. This is, however, the direction the Air Force is taking.

Meister - This is actually irrelevant, really, in terms of Jerry's question which is a crucial one, The tasks he's dealing with are perhaps molecular troubleshooting tasks, but the principles that he's trying to get at are the rules of combination, whether they are independent or dependent. These combinatorial rules are applicable not only to integrating molecular tasks into grosser tasks, but also to the integration of gross tasks into complete functions. Moreover, you just can't say, I want to know what's the probability only of the man being able to fly a bomber; that's not good enough, that's too gross. The basic principle he's getting after is a crucial one even though you might decide to throw away the more molecular tasks in terms of your predictions. Presumably if combinatorial rules hold for the simpler tasks, this will give us perhaps greater confidence that they hold for larger, more molar tasks. We don't have that kind of information.

Mills - I'd like to add something too. We mentioned C5A which reminded me that the MADAR system on that aircraft which is a microfilm projection, supposedly procedurized, troubleshooting guide that the flight engineer uses in combination with the automatic failure detection system. If a failure occurs on the aircraft in a module, he gets a signal indication that there is a failure in the aircraft. This is in flight. It also can be done on the ground by the maintenance technician. At this time the flight engineer then goes to his microfilm projection and begins to call up the routines for isolating the failed module. The format used is a typical computer type format flow chart and the flight engineer is told step by step what to do and what to test, what to read and the instructions are in the form of read voltage from such and such meter and he continues through this thing until he theoretically isolates the failure. The concept is that this is supposed to decrease the turn around time of the aircraft because the flight engineer then radios ahead and they have the replacement part ready.

Swain - Does this really work?

Mills - It's a terrific system and of course when we first encountered it, we believed it was really something that was going to make a tremendous impact on maintenance of aircraft and so we became quite involved in trying to develop an experimental program because the format that was not chosen, (the flow chart format was chosen because it just so happened that a person in the project had been a computer programmer and liked this format on the basis of performance criteria. At any rate we had some discussions with them in terms of the fact that they needed human performance research on this subsystem. We tried to generate for them the kinds of programs they needed for doing research on this system, for example on the flight engineer's capabilities. However, it never did get off the ground. I'm still thinking about the problem but I have no support to do anything in it. The point I wanted to make here was the fact that his performance, the kinds of instructions that are dictated to him in his procedures are in the form of task elements. The maintenance technician is also permitted to use this system when it is on the ground. His instructions are also in this form. The instructions that John Foley over at Human Resources Laboratory uses in his proceduralized aids which are going over well and being tested in Southeast Asia now are of a similar nature. They're a little more general but they're still at the kind of level that we're talking about when we talk about subtasks and task elements.

Meister - What's a plug in and plug out? To plug something in or remove it? That would be a simple task, the simplest level task. It's not a task element.

Lamb - Because the elements are unscrewing the bolts and pulling the thing.

Regulinski - Well, the C5A and the F-111 have similar electronic components, however they are mounted on printed circuit board so that if one goes bad, you simply plug-out the bad and throw it away, and plug-in the replacement. I think what you are saying is that when the wire-head of this plug assembly is out, this calls for the type of repair action which may best be modeled by an exponential maintainability function.

Jenkins - Many of the tasks of the BQS-6 are quite similar. The technician is given very specific instructions through his tech manual and his maintenance repair cards. "Read voltage ±2. If more than 2 do this, if less do that."

Mills - Back to this question of whether or not it works. I understand that it is not working very well.

Lamb - I think that we've pretty well covered everything else that I was going to say.

Meister - I think that everybody agrees that this is the kind of study that's needed. All we're really talking about is the details of how you carry out the study. Am I correct about that?

Jenkins - Do you see any particular category which you think in the task description, either from the paper or from what Jerry said, would lead us in the wrong direction, or in a less parsimonious direction?

Meister - As far as I'm concerned you definitely have to investigate these combinatorial processes. Otherwise you would never get any further than collection of simple task data.

Harris - I see some little problem with the ratio of failures, whether this is based upon different individuals or are these measures of repeated actions of one person? The question is really, "Is that ratio based on the number of different people or is it derived from the number of steps times the number of people?"

Lamb - For the molecular level, yes, but we're measuring success. Did he get the system working again independent of how many times, how long it took him and how many faults he made. But in that we are measuring the number of steps that he makes, and if he makes the same step two or three times, each one of those is an attempt. Some of our analysis is going to depend on his failure and how it comes out.

Mills - This scientific mystique about having to specify all your hypotheses in advance is oversold. Obviously you don't go about a study without having some idea of what kinds of hypotheses you want to test generally, but you find out too many things in the study itself to limit yourself, to restrict yourself.

Swain - That's why I'm suggesting that when you do the study, you have the people verbalize what they're doing every step of the way.

Meister - No, no.

Siegel - I tend to feel that some type of introspective report is necessary. For example it occurred to me that while I was sitting here, did Newton write down F = MA and then say, "Gee whiz, now I'm going out and collect data and verify it," or through some mystique did it occur to him that F is probably related to MA in some way. Then he worked a little bit and came out with the answer. I suspect it was the latter case, although I haven't spoken to him recently. My point is that just to say rigidly, "I'm going to test the empirical procedures, I'm going to apply some empirical procedures," may not be entirely correct, and as you'll see when I speak, that is the procedure I'm following, but it may not be entirely correct, I'm prepared to admit that. One would argue in favor of Alan.

Lamb - I think, Dave is arguing that introspective data was in fact necessary, the question is whether you collect it all in a continuous stream of thought or in some measured way.

Meister - Our experience has been, and I'm basing this solely on experience, that when you ask a man to continuously verbalize while he is doing something, it's almost like a concurrent interfering task for him. What we typically do, I think every experimentalist does this, is after the man has completed the task, then we interview him very intensively to try to find out what went on. If it's a very, very long procedure, you may have difficulty, but if it's a relatively short procedure, the man generally can tell you what essentially he was going through, what the reasons were for his actions and so forth. You may have to yank it out of him, but generally speaking you can get a fair amount of information in that way. I wasn't really objecting to Alan's concept of how we often get additional information, I was objecting simply to the method of getting that information.

Swain - We've used the method, the one that I proposed, and it seemed to work real good.

Meister - I know, there are strange things going on at Sandia.

Regulinski - May I summarize what the system engineer would probably look at? I think Dave's observation is perfectly valid. Certainly for our own satisfaction if nothing else, we should know whether in fact these tasks are dependent or independent. The systems engineer, when he finds the tasks are dependent, has at least tools to solve the problem. This is not an unsolvable situation, you understand, he has used them for years. But, more important to us is when we are working in time-space-continuous domain, in which case we do not worry about time dependence or independence because that's inherent in the data. If you are working with discrete levels as I detect these levels down below to be, then you should absolutely worry about such things. This is a perfectly legitimate point of research.

Jenkins - Could you give us an idea of when you will finish the study, about a year from now, two or what?

Lamb - With everything we should be finished, and I'm always optimistic, within a year.

Meister - Is this particular area a continuing program of research or just a single study?

Lamb - What I'm particularly interested in doing is finding out how to improve in the system engineering equations the discrete data. In one aspect of this model to model should it be task element, should it be task, should it be function, where are we? One of the questions that came up was, "we've always said that we're going to combine independently and the first step is making sure we're putting in the equations correctly is to test this independence. The long range goal is to incorporate them appropriately in the system engineering.

Meister - Are you going to continue doing further studies, perhaps complicating the situation, adding new parameters? Or after you finish this study, will that be it? I can think of n number of parameters that affects this independent relationship and I would hate to see just a single study run on this.

Jenkins - It's not our intention to ask the laboratory to stop at the conclusion of this study. The next phases have to be discussed and they're going to be quite dependent on what we find here. Ideally, what we would like to do is if we verify our hypothesis, is to prepare a handbook of a set of rules to include in Navy contract requirements so that for all current and future systems we would have a way of predicting the total system reliability, so far as maintenance actions are concerned.

Meister - What you're talking about is some sort of data bank handbook. That's going to be the output of your entire ADO project, will it not?

Jenkins - I think the output of this project will be a number of things. One of them will be hopefully a number of types of data banks. One of them will be this kind of design engineer's handbook.

Mills - How does this fit into the ADO? In light of what Cdr. Connery said yesterday, does this mean that if you did obtain your funding you would support this kind of research?

Jenkins - How will you do research? You could do it many ways. From this point of view we can't cast the results of research as basic research.

Mills - What I'm asking is simply would you put out an RFP for a research program of this nature? One of the things that I think we'll have to discuss is given these projects what are the next steps? Such as, maybe it is an RFP.

Tolcott - If research is necessary as part of this ADO implementation, could this research be supported out of 6.3 money?

Mills - I make a distinction between the in-house type of research and between support of contractor research. Basically you are not generally going out to do these kinds of things. You contract to solve a specific problem. We get these projects to solve these specific problems. The solution of the specific problem is dependent upon some sort of research base. It's never established, some sort of data base is never established. I contend that this is one of the problems in this area. We go out with a contract, we say we'll build this data system. We'll do it on this amount of money. When what really should be done is, we'd like to have a data system sometime in the future, but it is obvious that we need a certain amount of basic fundamental research whether it's operational or laboratory research.

Fields - There are funds for these purposes.

Mills - As I understood Cdr. Connery yesterday, the ADO is not directed for this.

Jenkins - This program is directed toward solving the problem; how we do it is our business. We have to turn out a product, that's all he's interested in.

Mills - If you're going to sell a piece of research to the government, you've got to include some sort of a concrete output that they can see. Research in general is something most people can't feel, so you've got to sell something in addition.

Jenkins - You interpret it too literally.

Fields - In solving an ADO you can do anything you've got to do to solve it, as long as you solve it.

Blanchard - I'd just like to present our reaction to the problems of dependence. It is our experience that we've never got into a position yet where we could make the assumption of independence. I think it's fine to research the question and I'm all for that. There are ways of handling dependent events in a system and the way we got at it in our own primitive way was through the technique of graphic modeling. Any model or any problem like this usually requires some form of mapping technique. The technique usually evolves to a mathematical model. The big problem that we found was in mapping a system and being able to identify and account for these dependencies and feedback loops and so on. As I mentioned before, a lot of these relationships are very subtle. What we did, when we had a feedback condition that we could define, was to look back at preceding tasks and assign them a probability of 1.0 because realistically speaking, as far as when that system is operated, that task probability-wise would be completed perfectly. You might make an error the first time through, but because of a cue or some form of feedback later on, the error would be detected and retrieved before the consequences of the error could occur. What we're after is a prediction of the operability of that system which is sensitive to the effects of dependencies.

Swain - In your case it just happened to work out that way. When we do it, we use different values if less than that is appropriate.

Blanchard - This is a perfectly retrievable case I am using as an example. Obviously, they might not all be so. Then you must consider detection probability and retrievability probability. We have found in most cases at fairly micro level of analysis that it was highly unlikely that the error if it were made would not be detected and corrected. I don't think under most circumstances when you're concerned with an applied problem and really wrestling with it that there aren't very good ways, simple ways, of handling such problems. This assumption of independent is archaic. I really don't see it as being meaningful.

Lamb - I can imagine levels at which it is minimal.

Siegel - I think that when you start going global, your independence is subsumed within your total number. Now if your argument is that, well I'm worrying about repetition of this particular block, that is easy to manage. We can calculate success given one trial, given two trials, given 36 trials. So that's no problem. We can handle that aspect of it.

Meister - I'm forced to agree with both Bob and Art, but what I object to is the fact implied statement, "Okay, we certainly can handle these things, I'm sure we can on the basis of assumptions, but we have no empirical research to validate these assumptions." That is the whole point of this line of reasoning.

Siegel - And what if we get into too many dependencies, we end up with as many reliability prediction methods as there are systems in the world and we're not going to zero in on the problem, we're just going to open up a Pandora's box.

Meister - I would hope it may not occur, but I hope that research such as Bob's and such as Jerry's would allow us to be able to say that a certain type of dependency at certain levels could in effect be eliminated from consideration because of one reason or another; this would in fact simplify the total problem. The point is when we make assumptions and operate on them, all we're doing is playing some sort of game, and I personally would have more confidence in the various models and modeling techniques if we had more empirical data on which our assumptions could be based.

Regulinski - You make an excellent point. Telstar was designed so that you could isolate difficulty by following the Bayesian probability estimation. For example, if subsystem A were to go out, the probability of subsystem A going out is equal to the conditional Bayes probability. This conditional Bayes probability governs each subsystem, and data is telemetered.

DEVELOPING A HUMAN RELIABILITY PREDICTION METHOD

Arthur I. Siegel

Applied Psychological Services, Inc.

The Applied Psychological Services' Program Plan for Developing a Human Reliability Prediction Method

Arthur I. Siegel

Applied Psychological Services, Inc.
Science Center
Wayne, Pennsylvania

This paper presents the background to and methods involved in the Applied Psychological Services' program for developing a human reliability prediction technique. Such a technique would make a significant contribution to:

1. predicting the maintainability of future systems
2. the provision of significant design verification information, not otherwise available
3. the development of preferred methods of maintenance and use of equipment by operational commands

There is, at present, a set of specifications which prescribes the analytic determination of equipment reliability (mean time to failure) during the equipment development cycle. However, there is no parallel specification in the field of human reliability. Thus, although an early statement of the probability of hardware failure is sought, there is no parallel statement available in regard to human reliability. It is self evident that total system reliability is a function of both the equipment and the operator reliability.

Concepts and Considerations

Any technique which purports to yield information regarding the reliability of the human component in a system must possess a number of attributes if the technique is to be useful. First, the technique must yield a numerical estimate of predicted reliability. Moreover, the numeric which is yielded must be amenable to compounding with an equipment reliability determination in order to allow the determination of a total system reliability. Thus, it should be possible to combine the human reliability prediction directly (or with a simple transformation) with the equipment reliability prediction. This requirement indicates the need for a human reliability statement in terms of a probability number which indicates probability of successful human performance.

Second, the determination should state not only that a system possesses a given reliability, but it should also allow determination of what technician completed sequences (components in hardware reliability) were instrumental in causing the derived reliability. That is, it is not sufficient to know that a system possesses a given reliability. The designer wants to know where his system is weak. Only through this knowledge can he improve his predicted reliability. Stated alternatively, the technique must yield subsequence (subsystem in hardware reliability) as well as total sequence (system in hardware reliability) reliability predictions.

Third, the technique must be applicable early in the system development cycle. If the required human reliability prediction fails to become available until late in the design cycle, the cost impact of any indicated design changes could be excessive.

A fourth requirement involves technique practicality. Practicality infers cost minimization as well as ease of application. A technique which can be employed by a minimally trained analyst is held to be more practical than one which implies excessive mathematical or other sophistication. Similarly, a technique which is compatible with hand calculational or desk calculator methods is considered to be more practical than one which rests on the availability of high speed digital computers.

Fifth, the technique must be applicable to a wide variety of systems, i.e., the technique must possess generality. We note that a technique which is too broadly based may lack veridicality for any specific situation. On the other hand, a technique which is highly specific, while possessing considerable relevance for one situation, may fail to be relevant for other situations. Accordingly, a middle road, which will optimize the generalizability of application of the technique, is sought.

Sixth, the technique should be fully compatible with specified end products which emerge from human factors analyses which are currently performed during system development. Moreover, the technique should impose few analytic requirements other than those imposed by actual technique application. More specifically, if task or operational sequence data are required by the technique, the data requirements should be directly based on information which is customarily made available during the equipment developmental cycle.

Seventh, the technique should be valid. Validity in the present sense means predictive validity as well as content and construct validity. Validity further, in the present context, relates to the mathematical procedures which are involved and to the reasonableness of the mathematical assumptions.

The concept of psychometric reliability represents an eighth requirement. Different users should obtain the same answer when applying the technique to the same system. And, the same user should obtain the same answer when he applies the technique on separate occasions.

Finally, the technique should yield a statement of the time that it will take a technician to complete a given task as well as the probability of successful task completion. Time to completion is an important ingredient for military tasks and it seems, in the case of mean time to repair estimates, to be an aspect which cannot be ignored by any technique which purports to be at all inclusive.

Background

It is Applied Psychological Services' contention that a considerable body of knowledge, relevant to the problem of predicting technician contribution to weapon system reliability, has been developed in the past several years. It is our goal to build a technician reliability assessment technique on the basis of the firm foundation provided by these prior studies. These prior studies fall into four areas: (1) multidimensional scaling, (2) technician reliability estimation, (3) mathematical analysis, and (4) computer simulation. The recent developments in each of these areas are reviewed categorically below. Then, our plan is presented for weaving these prior methodological developments into a technician reliability predictive scheme which will meet the requirements outlined in the "Concepts and Considerations" section above.

Multidimensional Scaling

Multidimensional scaling analysis is a comparatively recent technique for defining or structuring an unordered universe. Originally developed by Richardson (1938), this expansion of basic psychophysical scaling has recently been studied and extended in some detail by several of Gulliksen's students and a few other research workers. Gulliksen (1961) summed up his feelings about the value of the methods involved by saying that multidimensional scaling:

> ...is a rather powerful technique for investigating a wide array of situations. The basic experimental question is a very simple one. Despite a superficial appearance of difficulty and unreasonableness, one can get consistent answers and can come up with rather interesting conclusions--some of which verify the results of unidimensional scaling and others of which go beyond (p. 17).

The two central problems in multidimensional scaling analysis are the determination of: (1) the minimum dimensionality of a given set of stimuli and (2) the scale value of each stimulus on each of the dimensions. The specific experimental and computational procedures used have been described in detail by Torgerson (1952, 1958), Messick (1956a, 1956b), and others.

As Gulliksen has pointed out, the basic judgment upon which the whole structure of multidimensional scaling analysis rests is very simple. In order to obtain estimates of the "psychological distances" among the various stimuli in a set, most experimenters have asked the subjects (judges) merely to indicate in some manner the degree of over-all similarity between each stimulus pair. The methods for obtaining and scaling these distance judgments are generally analogous to the classical psychophysical scaling techniques.

If the obtained scale values can be taken as measures of the interstimulus distances in a Euclidean space, the analytical problem then becomes the determination of the number of axes in that space and the projections of the stimuli on these axes. In these final stages multidimensional scaling analysis uses factor analytic methods. As in factor analysis, for example, the pattern of scale values (loadings) of the stimuli (tests) on each dimension (factor) presumably enables the experimenter to attach meaning to, and so to name, the dimensions.

There are a number of technical problems involved in multidimensional scaling, such as the choice of method for obtaining the inter-stimulus distance estimates, the choice of spatial model to represent the distances, the determination of the constant required to set the distance estimates on a ratio scale (Messick & Abelson, 1956), and the decision as to whether a transformation of the basic data is required (Helm, Messick, & Tucker, 1961). Basically, however, multidimensional scaling involves the steps of: (1) obtaining a matrix of inter-stimulus distances and (2) determining the dimensionality of the space containing the stimulus points.

The techniques have been applied to a wide variety of problems. The early work on colors by Richardson (1938) and on relations between nations by Klingberg (1941) has been followed more recently by applications to such areas as attitudes (Messick, 1954, 1956a; Abelson, 1954), personality (Jackson, Messick, & Solley, 1957), jobs (Reeb, 1959), and facial expressions (Abelson, 1962), among others, in addition to further work in color (Torgerson, 1951, Messick, 1956c).

Multidimensional scaling differs from unidimensional scaling in one very significant respect. In the typical unidimensional experiment, the scales or dimensions are presented to judges who are asked to order the stimuli on the dimensions as defined by the experimenter. In multidimensional scaling no such a priori assumptions or definitions are made. Rather, the purpose of the analysis is to discover the number and characteristics of the underlying dimensions which may be justified by the empirical data.

In areas where the variables are complex and the dimensions unknown or doubtful, it would seem particularly appropriate to delineate the variables through multidimensional scaling analysis rather than to establish the dimensions arbitrarily. The research in areas of fairly well established dimensionality, particularly color, has been cited as evidence of the validity of the methods. Messick, in particular, after completing some of this work, concluded that "since multidimensional scaling procedures yielded structures which correlated highly with the revised Munsell system, it would now seem reasonable to apply these procedures for purposes of exploration and discovery in areas of unknown dimensionality" (1956c, p. 374).

Siegel and Schultz (1963) and Schultz and Siegel (1962) performed a multidimensional scaling analysis of the job of the Naval electronics technician. As a result, nine basic factors were isolated. These factors can be employed to describe completely the work of the Navy personnel concerned with electronic maintenance. The factors are: (1) using reference materials, (2) instruction, (3) equipment operation, (4) electro-safety, (5) electro-cognition, (6) electro-repair, (7) circuit analysis, (8) equipment inspection, and (9) personnel relationships. It is these factors which will form the basis for the proposed work.

Technician Reliability Estimation

The factorial based, empirically derived taxonomy described above yields a simple basis for describing all electronic maintenance tasks. It provides a structure which is manageable and which is free from excessive cumbersomeness. However, the question of the availability of data regarding the Navy technician's ability to perform the functions subsumed by these factors arises. Applied Psychological Services has already collected such Fleet data on the proficiency of Naval electronic maintenance personnel in each of these factors. Moreover, these data are in a form which is directly compatible with the mathematical analytic technique described below. The data were collected in 1969 and hence are applicable to the current Fleet technician. Moreover, the methods and techniques for collecting such proficiency data have now been proven. Accordingly, additional or updated data of this nature can easily be derived. However, it is believed that the presently available data are sufficient for the next several years.

The present data store is based on the following ships: USS Roan, USS Dyess, USS Sperry, USS Basilone, USS Ingraham, USS Page, USS Fiske, USS Eaton, USS Cony, USS Hank, and USS Conway. A total of 533 technicians is involved in the sample. These include technicians in the following Navy rates: EM, ET, FT, IC, RD, RM, ST, TM. It is believed that these rates include all persons concerned with electronic maintenance in the Fleet at present. Moreover, data are available by pay grade within rate. On the basis of these data, it is now possible to state the probability that each of the factors in our taxonomy will be performed by: a. technicians on a given ship within a given rate, b. technicians on a given ship but across rates, c. technicians across ships and rates, and d. a through c above by pay grade.

Thus, the methods to be employed by Applied Psychological Services do not need to await the development of appropriate input data. The input data are already available. Moreover, acceptable answers are already available to questions regarding the nature of the distributions underlying these reliability values, the factors involved in deriving these estimates, the psychometric reliability of these estimates, and the ability of these factors to predict other criterion data (Pfeiffer & Siegel, 1966). These other criterion data include such items as amount of school training required, training aid requirements, GCT and ARI requirements, school scores, and amount of inservice training required before acceptable proficiency can be anticipated.

Mathematical Analysis

The mathematical analytic techniques for achieving an integrated technician reliability value on the basis of the taxonomy described above have also been previously derived by Applied Psychological Services (Siegel & Miehle, 1967). Moreover, the applicability of these mathematical analytic techniques has already been demonstrated for two Navy systems, the IDNA system (Siegel & Burkholder, in press) and the HINDSIGHT system (Miehle & Siegel, 1967). The technique is fully described in the Siegel and Miehle (1967) report. However, it is briefly reviewed below. The reader is referred to the Siegel and Miehle report for a more complete elaboration.

The satisfactory performance of a task may require the satisfactory performance of some, or all, of certain activities.

Let:

s = satisfactory task performance

r_{mn} = satisfactory performance of job activity m by technician n

R_{mn} = reliability of technician n on activity m

$P_r[r_{mn}]$ = probability that statement r_{mn} is true.

Thus, $P_r[r_{mn}] = R_{mn}$ and $P_r[s]$ = reliability of task performance.

Suppose performance of a task involves technician b on three job activities: 3, 4, and 6, and technician g on three activities: 3, 5, and 8. Both technicians perform activity 3. The condition for satisfactory task performance is:

$$s \rightleftarrows (r_{3b} \vee r_{3g}) \wedge r_{4b} \wedge r_{6b} \wedge r_{5g} \wedge r_{8g}$$

$\vee$ is a symbol for inclusive <u>or</u> (inclusive disjunction)

$\wedge$ is a symbol for <u>and</u> (conjunction)

$\rightleftarrows$ is a symbol for "is equivalent to."

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. The negation of r_{ij} is r'_{ij}.

Series Activities

If all activities must be performed satisfactorily, the condition is expressed by joining all statements by "conjunction" ($\wedge$), $s \rightleftarrows r_{1a} \wedge r_{2a}$. This might be called a <u>series</u> task.

$$P_r[s] = P_r[r_{1a} \wedge r_{2a}] = P_r[r_{1a} | r_{2a}]P_r[r_{2a}]$$

$P_r[r_{1a}|r_{2a}]$ is a conditional probability which is read "the probability of r_{1a}, given r_{2a}." It is the probability that r_{1a} is true under the condition of r_{2a} being true. When the truth of r_{1a} is independent of the truth of r_{2a}, we say that r_{1a} and r_{2a} are independent statements. In this case $P_r[s] = P_r[r_{1a}]P_r[r_{2a}]$.

Let $s \rightleftarrows r_{2a} \wedge r_{3d} \wedge r_{6a}$

$$P_r[s] = P_r[r_{2a} \wedge r_{3d} \wedge r_{6a}]$$

$$= P_r[r_{2a} | r_{3d} \wedge r_{6a}]P_r[r_{3d} | r_{6a}]P_r[r_{6a}].$$

If all statements are independent, this reduces to:

$$P_r[s] = P_r[r_{2a}]P_r[r_{3d}]P_r[r_{6a}].$$

Parallel Activities

When a task is performed satisfactorily if either one or another activity (or both activities) is performed satisfactorily, this is expressed as:

$$s \rightleftarrows r_{2e} \vee r_{3e}.$$

This might be called a parallel task. In this case, job activities 2 and 3 are involved and the task is performed by man e.

When the same job activity is performed by two men and acceptable performance of either man will constitute acceptable performance for the team, the condition is expressed as $s \rightleftarrows r_{2a} \vee r_{2c}$. This might also be called a parallel performance.

Here, activity 2 is performed by men a and c.

$$P_r[s] = P_r[r_{2a} \vee r_{2c}] = P_r[(r'_{2a} \wedge r'_{2c})'] = 1 - P_r[r'_{2a} \wedge r'_{2c}]$$

$$= 1 - P_r[r'_{2a} | r'_{2c}]P_r[r'_{2c}]$$

$$= 1 - (1 - P_r[r_{2a} | r'_{2c}])(1 - P_r[r_{2c}]).$$

If the statements are independent:

$$P_r[s] = 1 - (1 - P_r[r_{2a}])(1 - P_r[r_{2c}]).$$

Let $s \rightleftarrows r_{1b} \vee r_{1c} \vee r_{1g}$.

$$P_r[s] = P_r[r_{1b} \vee r_{1c} \vee r_{1g}] = 1 - P_r[r'_{1b} \wedge r'_{1c} \wedge r'_{1g}]$$
$$= 1 - P_r[r'_{1b}|r'_{1c} \wedge r'_{1g}]P_r[r'_{1c}|r'_{1g}]P_r[r'_{1g}]$$
$$= 1 - (1 - P_r[r_{1b}|r'_{c} \wedge r'_{1g}])(1 - P_r[r_{1c}|r'_{1g}])(1 - P_r[r_{1g}]).$$

Both the series and parallel formulas can be extended to larger number of activities or performers. These formulas can be written in many different forms if conditional probabilities are involved.

Computational Examples

Let us select a number of the job activities performed by the Naval electronics technician:

1. using reference materials
2. instruction
3. equipment operation
4. electro-safety
5. electro-cognition
6. electro-repair
7. electronic circuit analysis
8. personnel relationship
9. equipment inspection

Example. Assume that Task A is performed by technician c and that activities 1, 7, 3, 4, and 9 are involved, that performance will be considered to be satisfactory if, and only if, either (or both) activities 1 or 7 are performed satisfactorily, either (or both) activities 3 and 9 are performed satisfactorily, and activity 4 is performed satisfactorily. This is symbolized by:

$$s \rightleftarrows (r_{1c} \vee r_{7c}) \wedge (r_{3c} \vee r_{9c}) \wedge r_{4c}^{*}$$

$$P_r[s] = P_r[(r_{1c} \vee r_{7c})|(r_{3c} \vee r_{9c}) \wedge r_{4c}]P_r[(r_{3c} \vee r_{9c})|r_{4c}]P_r[r_{4c}].$$

* This expression might be read as follows: this maintenance task will be successfully performed if either the use of reference manuals or an electronic circuit analysis is completed successfully and either an equipment operation or an equipment inspection is performed acceptably and safety precautions are observed throughout. All activities are performed by technician c.

The first two probability expressions would require further expansion to remove the expression $r_{1c} \vee r_{7c}$ in the first term and the expression $r_{3c} \vee r_{9c}$ in the second term. This would produce a very complicated appearing expression, still containing conditional probabilities, for $P_r[s]$. However, it can be argued that such expansion is not warranted.

In the study of systems reliability, it is generally assumed that the proper operation of one component does not depend on the proper operation of another. This assumption does not always hold. For example, suppose that two beams are used to support a weight. If one beam fails, the whole weight is then placed on the other which will now possess a greater probability of failure, although each beam was designed to hold the whole weight. This is the "domino effect." Another example is the recent extensive power failure in the northeastern United States.

For independence to hold, a failure of one component must not influence the operation of another. If there is a cause for failure of one component, that cause should not operate on the other components. If, for satisfactory overall performance, all components must operate properly, then when one fails, the whole system fails. In this case, it is irrelevant whether other components also fail as a result of the failure of the first component. Here, the reliability value of interest is the conditional probability of proper functioning, given that all other components function properly. Usually a component is tested in isolation, and it is assumed that when combined with other components, its reliability will not be influenced. Otherwise, each component would have to be assigned as many reliability values as there are systems in which it is used.

This consideration may also hold for persons on a job activity. If reliabilities were not independent, then a single value like r_{2b} would be useful only if technician b worked on job activity b all by himself. If he performed several activities or worked with other technicians, then his reliability would have to be determined on this particular task under a variety of "given conditions." In that case, overall task reliability may be determined directly rather than on the basis of the component reliabilities.

Assuming independence (the success or failure on one job activity does not affect the probability of success on another activity), the formula simplifies to:

$$P[s] = P_r[r_{1c} \vee r_{7c}]P_r[r_{3c} \vee r_{9c}]P_r[r_{4c}]$$

$$P_r[s] = \{1 - (1 - P_r[r_{1c}])(1 - P_r[r_{7c}])\}$$

$$\{1 - (1 - P_r[r_{3c}])(1 - P_r[r_{9c}])\}P_r[r_{4c}].$$

Assume the following values: $r_{1c} = .88$, $r_{7c} = .82$, $r_{3c} = .82$ $r_{9c} = .89$, $r_{4c} = .88$.

Then:

$$P_r[s] = \{1 - (1 - .88)(1 - .82)\}\{1 - (1 - .82)(1 - .89)\}(.88)$$

$$= \{1 - (.12)(.18)\}\{1 - (.18)(.11)\}.88$$

$$= \{1 - .0216\}\{1 - .0198\}.88 = (.9794)(.9802)(.88)$$

$= .845$ as the overall task reliability.

Example. Assume that a task involves technician j, who performs tasks 1, 2, 5, and 7, and technician e, who performs task 6. Tasks 3, 4, and 8 are performed jointly.

$$s \rightleftarrows r_{ij} \wedge r_{2j} \wedge r_{5j} \wedge r_{7j} \wedge r_{6e} \wedge (r_{3j} \vee r_{3e}) \wedge (r_{4j} \vee r_{4e}) \wedge (r_{8j} \vee r_{8e}).$$

$$P_r[s] = P_r[r_{1j}]P_r[r_{2j}]P_r[r_{5j}]P_r[r_{7j}]P_r[r_{6e}]$$

$$\{1 - (1 - P_r[r_{3j}])(1 - P_r[r_{3e}])\}\{1 - (1 - P_r[r_{4j}])(1 - P_r[r_{4e}])\}$$

$$\{1 - (1 - P_r[r_{8j}])(1 - P_r[r_{8e}]\}.$$

Allow the following r values: $r_{1j} = .82$, $r_{2j} = .79$, $r_{5j} = .91$, $r_{7j} = .77$, $r_{6e} = .86$, $r_{3j} = .81$, $r_{3e} = .76$, $r_{4j} = .87$, $r_{4e} = .80$, $r_{8j} = .88$, $r_{8e} = .75$.

$$
\begin{aligned}
P_r[s] &= (.82)(.79)(.91)(.77)(.86)\{1 - (1 - .81)(1 - .76)\} \\
&\quad \{1 - (1 - .87)(1 - .80)\}\{1 - (1 - .88)(1 - .75)\} \\
&= (.82)(.79)(.91)(.77)(.86)\{1 - (.19)(.24)\} \\
&\quad \{1 - (.13)(.2)1 - (.12)(.25)\} \\
&= (.82)(.79)(.91)(.77)(.86)(1 - .0456)(1 - .026)(1 - .03) \\
&= (.82)(.79)(.91)(.77)(.86)(.9544)(.974)(.97) = .352.
\end{aligned}
$$

This value seems too low to be considered to be acceptable performance, and training of the technicians in a number of the job activities seems indicated.

Activity Repetition

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, in the end, the whole task is performed unsatisfactorily. It is often possible to improve the result by repeating a process or by calling on someone else to correct deficiencies or "touch up" the result. This is equivalent to parallel operation which gives a reliability factor of: $1 - (1 - R)(1 - R) = 1 - (1 - 2R + R^2) = 2R - R^2 = R(2 - R)$, instead of R itself. Thus, if $R = .8$, the new factor is: $.8(2 - .8) = .8(1.2) = .96$.

The expected number of attempts, E, is a function of the maximum number (n) of attempts permissible or the number of attempts necessary to give a specified resultant reliability.

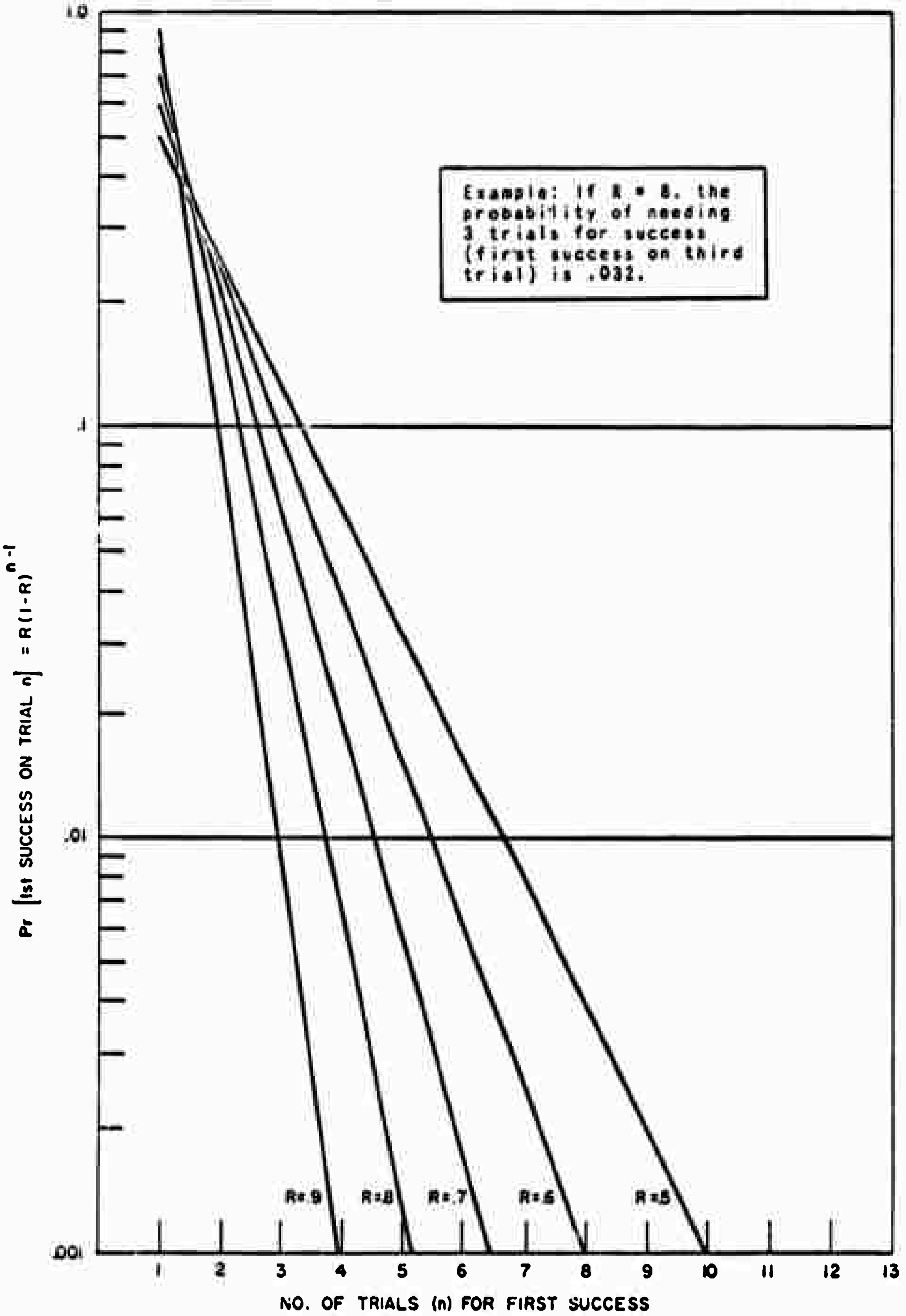
$$E_n = \frac{1}{R}[1 - (1 + nR)(1 - R)^n + n(1 - R)^n$$

where n is the maximum permissible number of trials. Figure 1 presents values of E for various R values.

In the limiting case, as n increases indefinitely, E approaches $\frac{1}{R}$. Thus, if $R = .8$, E approaches 1.25. This means that, if many trials are allowed, or equivalently, if the required reliability must be close to 1, then for $R = .8$ the average increase in the number of trials is not more than 25%. For $R = .6$, E approaches 1.67.

Let R_n be the reliability attained by allowing up to n trials:

$$
\begin{aligned}
R_n &= R + R(1 - R) + R(1 - R)^2 + \ldots + R(1 - R)^{n-1} \\
&= 1 - (1 - R)^n.
\end{aligned}
$$



Example: If R = .8, the probability of needing 3 trials for success (first success on third trial) is .032.

Figure 1. Probability of first success on trial.

Computer Simulation

Computer simulation represents a fairly recent technique for representing concepts which do not permit analytic solution in general form. As such, it yields numerical solutions to problems which do not lend themselves to deterministic solution. Computer simulation has been employed to investigate the behavior of people investing in the stock market, plant flow, and social behavior. It has also been employed in test of a number of military systems from the man-machine interactive point of view. Digital computer simulation has been made possible by the advent of the high-speed digital computer and is held to possess the following advantages:

1. Computer simulation allows consideration of the idiosyncratic and variable aspects of human performance.
2. Computer simulation is less costly than physical simulation.
3. Computer simulation allows test of hypothetical procedures and systems.
4. Computer simulation allows consideration of a myriad of variables in interaction. Consideration of a multiplicity of variables in interaction is not possible through other techniques.

Applied Psychological Services, Inc., has been in the forefront in the development and application of computer simulation techniques to represent behavioral processes. While Applied Psychological Services has developed a number of such models, the specific computer simulation model of interest in the present context is the model which has become known as the Siegel-Wolf two-operator man-machine simulation model. The model considers variables such as operator stress during task performance, operator waiting and idle time, effects of random environmental events, time allowance for mission completion, operator level of aspiration, and random extrinsic and emergency events.

Over the years, the model* has been successively revised, improved, and validated. Initially prepared to simulate the actions of a single operator (Siegel & Wolf, 1959), the model was later ex)anded to allow the simulation of a two-operator man-machine system. New variables have been added and computational routines constantly modified and updated as the model has matured. Similarly, the model has been continuously updated in terms of the use of more modern computing equipment and in terms of more symbolic programing languages. Most recently, the model has been adapted for running on the GE635 computer using the FORTRAN IV language.

The most thorough and up-to-date description of the technique is found in Siegel and Wolf (1969). Additional documentation, expansion, and elaboration of the technique has been performed by the Boeing Company, Honeywell, the General Electric Company, Autonetics, the Naval Air Development Center, Sperry, and the North American Aircraft Corporation. Currently, it is being employed by the Air Force for predicting weapon system vulnerability/survivability from the point of view of the human operator. It has also been specified for use in such advanced developments as the B-1 and the F111 weapon systems.

The model's predictive validity has been demonstrated in a wide range of applications. These studies, completed at Applied Psychological Services, have included carrier landing, air-to-air missile firing, inflight refueling, air intercept, sonar employment, and a series of simulated man-machine interactive situations. Additional validity studies have been completed by the Boeing Company (Outcalt et al., 1966) and by Honeywell Incorporated(Lane et al., 1966). In all of these validational studies, the ability of the model to predict independent outside criteria data was tested. With the exception of one of these tests, the results of all of the validation efforts indicated adequate correspondence (principal differences which are not statistically significant) between the model's prediction and the criterion data. Additionally, certain of the model's internal constructs have been validated, and the ability of the model to predict part-task success has been verified. The model has been adapted and successfully employed by a number of industrial and governmental organizations. Tnus, the model has withstood reasonable tests of validity and utility. It has also been employed in maintainability analyses for the PAIR system and its use is included in the specification for the IDNA system of the Navy.

* The word "model" as employed here is defined as a logical mathematical representation of a concept, system, or operation programed for solution on a high-speed digital computer (Martin, 1968). As such, digital simulation models are separate from replication (analogy) and formulation (mathematical) models.

The purpose of the model is to simulate a man-machine mission consisting of a series of operator and/or equipment "tasks" (alternatively termed task elements) involving one or two men. The major features of the model, as programed on the digital computer, are calculations for the following variables for each simulated task:

- stress based on time pressure. The model is largely, but not exclusively, time oriented

- task element execution time, stochastically determined from specified normal distributions using a Monte Carlo technique and dependent upon stress levels, and the operator speed parameter, F_j

- task element success and failure based upon both performance, stress, and probabilities supplied as input

Approximately 200 task elements are simulated in each second of computer time utilized. The ratio of real time to computer time is dependent upon the time the actual operators require in performing the task. For some recent sonar maintenance task simulations, the ratio for a single simulated mission (called a computer iteration) to actual mission time has exceeded 1000 to 1.

The following important items are also calculated or considered in the model:

- task element precedence (variable sequencing of task elements)

- maximum stress encountered

- operator interaction (waiting for a partner)

- joint task elements (performed by both operators simultaneously)

- equipment delays

- operator decisions

- skipping of nonessential task elements

- operator cohesiveness

- idle time spent waiting for a prespecified event before working
- time allotted for the mission
- time precedence (idling until a given time occurs)

In the case of a one-man task, the items, operator interaction, joint task elements, and operator cohesiveness are not applicable.

In order to simulate intra and interindividual differences in performance, the simulation of any individual task element is based, in part, on a random process. The total simulation process is a repetitive process, i.e., task elements are simulated sequentially to comprise a mission and the missions are repeated many times to obtain averages of the data generated by randomization techniques.

It is our contention that human reliability values which are derived from a combination of the mathematical and computer simulation techniques, described above, will yield a technician reliability estimate which is most defensible. The mathematical technique will yield probability of successful performance values for the series, the parallel, and the series-parallel maintenance cases. The computer simulation technique will take the same input probability values and yield predicted time values. It will yield such predictions as the time required by technicians at various skill levels to achieve a given probability of success and the technician skill level required to achieve a given success probability.

The research plan for the development of a human reliability determination system which integrates these two techniques and which is based on the factor analytically derived maintenance taxonomy, described above, is presented below.

Research Plan

The research plan, based on the prior discussion, which will achieve a technique for predicting reliability in modern weapon systems is presented below. The methods to be employed are based on a previous research investment of 10 man years. This prior research has been thoroughly documented. Hence, the savings to be accrued to the planned program are considerable. Moreover, the techniques to be employed have been "proven." Hence, a "minimum risk" program is involved. Briefly, and by way of overview, Applied Psychological Services' research program will involve four work phases. Phase I will involve selection of two available Navy systems to provide a test bed for the emerging technician reliability determination scheme. Phase I will also involve a deterministic calculation of the maintenance technician reliability for these systems. Phase II will involve a revision, expansion, and reprograming of the available computer simulation model so as to allow accommodation within the model of the type of data employed in the phase I analysis. Phase II also includes application of the computer simulation to a series of the tasks included in the phase I analysis. At this point, there will be available predicted time (from phase II) and success probability (from phase I) values for the maintenance tasks in two current Navy systems. Phase III involves the collection of actual technician performance time and success data for the two systems involved and comparison of these data with the predictive data yielded by the work of the two prior phases. In phase IV, a complete report describing the logic, methods, procedures, and results of the total study will be prepared. A separate user's manual will also be prepared in phase IV. This manual will represent a tutorial text for future users of the method. The specifics of the work to be performed in each phase are described categorically below.

Phase I--System Selection and Mathematical Analysis

As an initial step in the proposed work, two representative Navy systems will be selected to serve as a test bed for the current developmental effort. These systems will be selected, in coordination with Navy program representatives, on the basis of their representativeness of electrical and electronic systems within the Navy and on the basis of the availability of the systems for inclusion in later aspects of the present study. Having selected the systems, the maintenance repair actions to be considered within the present work will be selected. A sample of around 25 maintenance tasks for each system is seen as sufficient, i.e., a total of 50 maintenance tasks. These 50 tasks will be selected according to such criteria as frequency of occurrence, criticality, and representativeness of various types of electrical system (power supply, filter, etc.).

Having selected the systems and tasks, technician reliability will be calculated for each task in each system and for each system in toto. In these calculations, the series, the parallel, and the series-parallel cases will be considered. The calculational methods employed will be those described fully in Siegel and Miehle (1967) and outlined in Section I to this report. The input probability data for these calculations will be the probability data available at Applied Psychological Services and also briefly described in Section I of this proposal. The structure of the analysis will rest on the factors (also described in Section I of this proposal) derived in the prior Applied Psychological Services' factor analysis of electronic maintenance. Specifically, each maintenance task will be broken into its subelements and the factors applied to each subelement. Then the subelement reliability will be derived. This will be followed by the calculation of the total reliability on the basis of the subelement reliability values. The end results of these calculations will be a probability of completion of each of the 50 tasks selected for analysis. Moreover, for each task, a probability statement will be available regarding which subelement(s) contributed most to low human reliability prediction values.

These calculations and analyses will be completed for the series, the parallel, and the series parallel situations.

For a subsample of eight tasks from each system, two other analysts will independently complete the same analysis. Comparison of the results obtained by the various analysts will enable a statement of the between analyst agreement. A computational example follows:

> Consider a sonar system in which one column of the bearing time recorder's printout is blank. The first step of the repair action involves the use of built-in test equipment. This step may indicate a switch (or switches) to be defective, a delay line to be defective, or wiring to need repair. One or two technicians may be employed to complete the repair. The use of the built-in test equipment entails primarily equipment operation (EO). All three repairs involve equipment repair (ER). The second technician is used (if at all) only on the electro-repair activity. Three cases will be calculated: one man and no job activity repetition, one man and one repetition, and two men and no job activity repetition. Let $R_{ED} = .89$ and $R_{ER} = .83$.

One Man and No Job Activity Repetition

$$s \rightleftarrows r_{EO} \wedge r_{ER}$$

$$P_r[s] = R_{EO}R_{ER} = (.89)(.83) = .739.$$

One Man and One Repetition of EO and ER

$$R_{EO} = 1 - (1 - .89)^2 = 1 - (.11)^2 = .9879$$

$$R_{ER} = 1 - (1 - .83)^2 = 1 - (.0289) = .9711$$

$$P_r[s] = (.9879)(.9711) = .959.$$

Two Men on ER and No Job Activity Repetition

$$s \rightleftarrows r_{EO} \wedge (r_{ER1} \vee r_{ER2})$$

$$P_r[s] = R_{EO}[1 - (1 - R_{ER1})(1 - R_{ER2})]$$

$$= .89\,(1 - .0289) = .864.$$

Phase II--Computer Model Elaboration

Phase I will have yielded sets of overall and subelement probability values for each of 50 maintenance tasks as drawn from two current Navy electronic systems. However, the technique described above, while yielding statements regarding the probability of completion of the various tasks, fails to yield statements of the time to perform the tasks. We intend to rely on the computer simulation technique, described in Section I of this proposal, for these time values. To this end, the Siegel-Wolf computer model will be modified and reprogramed so as to accommodate the factorially based probability values. Currently, the model employs an input probability value for each subtask along with input time and standard deviation values. Several methods exist for modifying the computer program so as to allow consideration of the factorially based input values. The most direct approach would consist of the development of a subroutine which will modify the present input probability and time values as a function of the factorial probability values. If treated in this manner, the subroutine is essentially a preprocessing subroutine and would add only minimally to computer processing time. Moreover, the internal validity of the model would remain unaffected. The boxes enclosed by the dashed lines in Figure 1 represent this method for modifying the computer program.

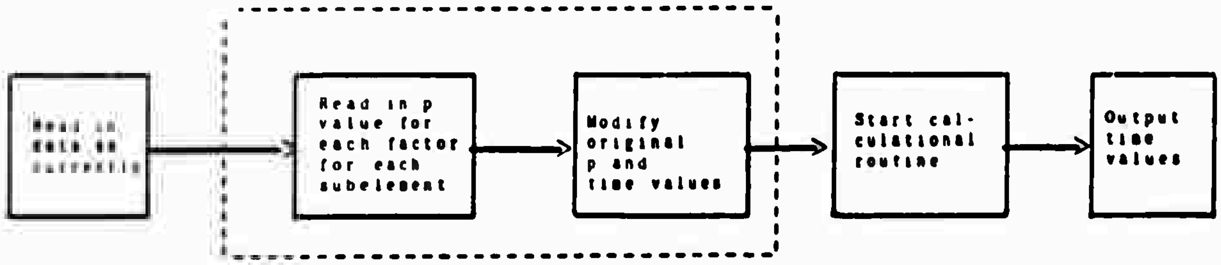

Read in
data on
Read in p value for each factor for each subelement
Modify original p and time values
Start calculational routine
Output time values

Figure 1. Overview of modification to computer simulation program.

Having developed the complete logic for this computer program modification, reduced the logic to flow chart form, and programed the logic, the simulation will be applied to a sample of ten of the tasks analyzed during phase I.

At this point, there will be available the probability values derived during phase I and, for ten of the tasks, predictions of the time to complete each electronic repair.

Phase III--Validation

In phase III, we plan to verify the probability and time data, as derived in phases I and II, against actual criterion data. A sample of maintenance technicians will be asked to perform each of the ten maintenance tasks for which both probability and time values are available. The test situation will be of the practical performance test nature and, if possible, conducted aboard ship using bugged components. It is believed that a sample of around 60 technicians will be required. The malfunction repair acts of these Navy technicians will be scored in such a manner as to allow the derivation of probability values which are directly comparable to the probability values yielded by the analyses of phase I. Similarly, the time required for task completion will be recorded and compared directly with the time estimates yielded by the computer simulation of phase II.

For these validations, only the series and the parallel circumstances will be considered. Twenty of the technicians will be involved for the series validations and the remaining 40 will be involved with the parallel validations.

Tests of statistical confidence will then be applied to both the probability and the time data sets, as yielded by the predictive method and as yielded by the actual measurement, in order to determine the statistical reliability, if any, of the differences between the two sets of data. At this point, the validity of the derived probability and time predictive techniques will have been established, at least in a preliminary manner.

Phase IV--Report and User Manual Development

In phase IV, a complete and detailed technical report will be prepared. The report will include the details of the methods employed and the results obtained. The implications for technician reliability determination in the Navy will be fully discussed.

As a second aspect of phase IV, a user's manual will be prepared. This manual will represent a simple tutorial text for the use of others who may wish to apply the derived technician reliability determination methods and procedures.

Review of Program Plan

The program starts with the firm foundation provided by prior research. It proceeds systematically through the derivation of probability of successful task completion to the derivation of task completion time. Then, the validity of the various procedures is tested against actual Fleet data. Finally, a user's manual is prepared along with a technical report which fully describes the program. The methods and techniques proposed for employment have been largely demonstrated in prior studies. Moreover, the data store required to support the factorially derived maintenance task structure is already available at Applied Psychological Services. Thus, a successful study and a usable end product may be anticipated at a high level of confidence.

The planned methods and techniques may be evaluated against the set of requirements delineated earlier. This evaluation is performed in Table 1. We note that, of the nine requirements, the proposed Applied Psychological Services' technique meets eight. One requirement is only partially met (requirement 5). However, a taxonomy which is so broad as to cover all types of electronic, electrical, mechanical, and electromechanical systems will probably be so broad that it will lack precision. Like a blunderbuss, it may scatter its shots broadly but lack precision for any given target.

Table 1

Comparison of Proposed Methods and Techniques with Requirements

Requirement	Met or Not Met	Remarks
1. Yield numerical probability estimate	Yes	Compatible with hardware reliability estimate. May be directly compounded with hardware estimate.
2. Allow statement of work sequences yielding low (high) reliability	Yes	-
3. Applicable early in developmental cycle	Yes	-
4. Practicality	Yes	Reliability calculation requires only simple multiplication. Computer application is automatic.
5. Generality	?	Proven applicability to Navy electronic systems. Applicability to mechanical systems remains unknown.
6. Compatible with other human factors techniques; minimum additional analytic requirements	Yes	Required data may be drawn directly from task analysis and operational sequence diagrams
7. Validity	Yes	Internal validity already demonstrated. Validity of computer technique already demonstrated. Predictive validity to be demonstrated during program. Mathematical assumptions not excessive and procedures are same as those accepted elsewhere.
8. Psychometric reliability	Yes	-
9. Yield time as well as probability value	Yes	-

REFERENCES

Abelson, R. P. A technique and a model for multidimensional attitude scaling. Public Opinion Quarterly, 1954, 18, 405-418.

Abelson, R. P. Multidimensional scaling of facial expressions. Journal of Experimental Psychology, 1962, 63, 546-554.

Gulliksen, H. Linear and multidimensional scaling. Psychometrika, 1961, 26, 9-25.

Helm, C. E., Messick, S., & Tucker, L. R. Psychological models for relating discrimination and magnitude estimation scales. Psychological Review, 1961, 68, 167-177.

Jackson, D. N., Messick, S., & Tucker, L. R. Psychological models for relating discrimination and magnitude estimation scales. Psychological Review, 1961, 68, 167-177.

Klingberg, F. L. Studies in measurement of the relations between sovereign states. Psychometrika, 1941, 6, 335-352.

Lane, D. S., Schaefer, R. W., Schwartz, H. F., & Braley, G. F. Man system criteria for extra-terrestrial roving vehicles. 12504-ITR3, Honeywell Inc., 1966.

Martin, F. F. Computer modeling and simulation. New York: Wiley, 1968.

Messick, S. J. The perception of attitude relationships. A multidimensional approach to the structuring of social attitudes. Unpublished doctoral dissertation. Princeton University, 1954.

Messick, S. J. The perception of social attitudes. Journal of Abnormal and Social Psychology, 1956, 52, 57-66. (a)

Messick, S. J. Some recent theoretical developments in multidimensional scaling. Educational and Psychological Measurements, 1956, 16, 82-100. (b)

Messick, S. J. An empirical evaluation of multidimensional successive intervals. Psychometrika, 1956, 21, 367-375. (c)

Messick, S. J., & Abelson, R. P. The additive constant problem in multidimensional scaling. Psychometrika, 1956, 2, 1-15.

Miehle, W., & Siegel, A. I. Maintenance personnel subsystem reliability prediction for the HINDSIGHT system. Wayne, Pa.: Applied Psychological Services, 1967.

Outcalt, N. R., & Loughlin, J. One versus two man crew study, general purpose attack aircraft, Vol. IV, crew performance computer model. D6-16276-3, Boeing, 1966.

Pfeiffer, M. G., & Siegel, A. I. The functional relationship between job complexity and a number of electronic maintenance training variables. Wayne, Pa.: Applied Psychological Services, 1967.

Reeb, M. How people see jobs. A multidimensional scaling analysis. Occupational Psychology, 1959, 33, 1-17.

Richardson, M. W. Multidimensional psychophysics. Psychological Bulletin, 1938, 35, 659-660.

Schultz, D. G., & Siegel, A. I. A comparative multidimensional scaling analysis of the job performance of Naval aviation electronics technicians. Wayne, Pa.: Applied Psychological Services, 1962.

Siegel, A. I., & Burkholder, G. Operator decision reliability for the IDNA system. Wayne, Pa.: Applied Psychological Services, 1969.

Siegel, A. I., & Miehle, W. Extension of a prior personnel subsystem reliability determination technique. Wayne, Pa.: Applied Psychological Services, 1967.

Siegel, A. I., & Schultz, D. G. A comparative multidimensional scaling analysis of the tasks performed by Naval aviation electronics technicians at two job levels. Wayne, Pa.: Applied Psychological Services, 1963.

Siegel, A. I., & Wolf, J. J. Man-machine simulation models. Performance and psychosocial interaction. New York: Wiley, 1969.

Siegel, A. I., & Wolf, J. J. Techniques for evaluating operator loading in man-machine systems. Wayne, Pa.: Applied Psychological Services, 1959.

Torgerson, W. S. A theoretical and empirical investigation of multidimensional scaling. Unpublished doctoral dissertation. Princeton University, 1951.

Torgerson, W. S. Multidimensional scaling. I. Theory and method. Psychometrika, 1952, 17, 401-419.

DISCUSSION OF SIEGEL's PRESENTATION

Meister - When you use the factor analytic approach you get different factors, depending on the different types of questions you ask.

Siegel - I'd be happy to get into a discussion of factor analysis and when we stop factoring and how you know what percent of variance you've accounted for, and even what you estimate the commonalities with and what I estimated the commonalities with. I don't really know whether that would take us anywhere. All I'm saying is what has been done. We did a factor analysis of the universe of maintenance tasks and, by the criteria of factoring we employed and by the input data we employed, we derived this specific taxonomy. Now, in answer to the questions asked, I say to you, it's very easy for a guy to sit back there and pick away but that's not a fair way to criticize. If you've got a better factorial structure, come in here with it; come in with some data and then we'll compare the two sets of data.

Meister - The question is not to imply criticism, though you're sensitive to this.

Siegel - No, I'm not sensitive to it, I think at this level, we're trying to move ahead, and, certainly, at every stage of the procedure, there have been assumptions made. In answer to your question, Dave, there have been judgements made. We've collated data from different sources. We are trying to move ahead with measured deliberateness. There could be mistakes in there, there's no doubt about that.

Meister - Well, let me ask this question, Art. Do you envisage when your methodology is fully implemented, fully worked out, that you will have fairly close relationships between your various factors and at least a limited set of equipment characteristics?

Siegel - Yes, as a matter of fact. Let's assume that the technique works. I could sit down and say factor 1 equipment reflected characteristics are and enumerate right down the line.

Swain - I was just curious because the IOC operates different equipment generally than an FT so how would you know what of your variance was due to just their differences and the equipment differences. Is there a technique for doing this?

Siegel - I would not be able to say - this is not a design tool. I think we have to be careful. It's not a human factors checklist.

Tolcott - One of the questions that has been underlying these two days of discussion has been, "Do we have someplace in our literature data that can be used, or does this program have to be implemented by people going out and collecting new data by one means or another? You've got some data there, the discussion is addressing the question of what are the characteristics of the data? Are they generalizable enough so that we can use it.

Siegel - I would say the data are generalizable.

Meister - Is it usable in terms of the kinds of design development questions which may arise?

Siegel - What question are you asking? Is system A better than system B? You can answer that.

Meister - What equipment characteristics should I include in my design in order to improve?

Siegel - You can answer that.

Meister - You can?

Tolcott - You can if the factors are universal and can be related to performance.

Meister - That's the missing step.

Lamb - Isn't that the question you asked him - that when he was finished and if it can be related and I would think it can, then yes it can do the design thing.

Tolcott - Can you, Jerry, for example use those factors in what you're doing and show some relationship between measures that you're getting?

Lamb - Yes, I think I can, but not at my grosser level, It's not really the equipment, it's the task you're performing on the equipment, And then I think you have to take it one step further to get to the equipment itself. It's not really a one stage translation. We're talking about two steps in the translation, but it can be done.

Siegel - We want to remember that we're trying to predict technician reliability. That's the major goal.

Meister - Why are you trying to predict technician reliability?

Siegel - To determine the contribution of the human error to the total system error. Not a design tool.

Meister - But you see, that's the problem. I can accept your technique, Art, in terms of the limited uses that you specified the technique will serve, but then there is a whole set of other questions which your technique will not handle which suggests to me that we need an additional approach.

Siegel - I agree with that and my general philosophy in the world is that as you build a technique that gets broader and broader, more general and general, to answer all questions to all people, it gets less valid to any individual approach or usage.

Tolcott - I'd like to rephrase what you've said. Maybe it will point us in the direction of what has to be done. I'd like to say, you've got some data, you developed for a somewhat different purpose than the data bank that we've been talking about. What has to be done next to find out whether your data can ideally be used?

Siegel - The question then becomes, "do these data that you are presenting involved in this combinatorial technique, cut cheeze?" and, "How can we find out?" This is what we're going to do, we're going out and set up a validation experiment.

Meister - Validation, or an experiment to establish relationships between your factors and equipment characteristics?

Siegel - Validation.

Tolcott - That other step has to be taken, somebody has got to address that question.

Jenkins - Let me clarify something. Our thinking was this: if this technique will give us a better approximation of human reliability, not the best, but a better one than anything that we've got going (which is almost nothing), then let's try it and see whether it is. The purpose is to take Art's approach and apply it to two systems for which the equipment reliability is very well documented. One will probably be a communication system and one would probably be an aircraft electronics system, and validate the technique. After we have it, we would then be able to get to the, if in fact it does work, we would then be able to give to the systems designers a first cut at the early design stage for gross system problem. For example, the kind of problem which we have continually is what they call the body and maintenance philosophy and what they're saying is to what extent should we have an automatic fault location system? Right now the automatic fault location system is costing as much as the operational system. How deep do we have to go to automate for fault location, particularly with microcircuit cards? Nobody really knows. Our technique would answer at a gross level that kind of question.

Meister - But that's not the only question.

Jenkins - Agreed! That's the purpose.

Meister - I have great confidence it will work. It will be validated. Art always validates his data, right?

Jenkins - I think you have a deep objection to something, and I don't know what it is.

Meister - No, I do not object to Art's technique. Art knows that I'm a great admirer of his work. What I'm saying is that apparently Art's technique deals with a subset of the total number of questions which a human reliability technique must use, and that's fine. It works for what it works. I would like you not to forget that there are other subsets of questions with which it will not deal.

Mills - This is very important, especially for your program for which you must determine limitations. I don't want to use that word in a derogatory fashion, only that each technique as you say, one technique spread out over more problems will become less valid. You have to recognize what questions it will answer, what questions it will not answer. That's why it is so important to first of all, at least one of the first things is to try to determine what kinds of questions will be asked. Those that you cannot answer then you try to go out and find the answer.

Jenkins - We see this essentially as something that may be useful. System development is a series of steps. Here you have math analysis, then you have computer simulation, then you have simulation and so on. We see this as a technique that can be used at the conceptual level for helping clarify the design concepts. Now we don't say this is the answer to all questions. We say there are other techniques needed, other approaches needed, other methods which will and should be used at various phases.

Mills - This is interesting because some of this may be taking some shape. For example, the developmental questions we might be asking at the conceptual stage of system development just may be perhaps an ideal place for Bob Blanchard's type of technique. Alan Swain's may operate at a little lower level for some more specific questions. There's not so many differences here, after all, and these are the kinds of things I think we want to try to find out. Try to put the blocks in order.

Lamb - We've been talking for the last few days, everybody thinks we have two different and not necessarily convergent goals, and one is the prediction aspect which we have to help the system engineering people with and the conceptual designer, the other is the design phase itself where we have the problems of actually using some sort of prediction perhaps, of a different nature in order to answer the kinds of questions that Dave has posed in his presentation.

Regulinski - I assure you I have no quarrel with the mathematical model. It's an excellent model indeed. However, I think it would be less than honest to say that there are some things that do not bother us. You ask Dave Meister, there's something that's bothering him. I hoped you detected what is bothering us all.

Jenkins - Dave made the point yesterday that we have to look at all the various models and determine what are the critical assumptions, where and how far each go. I can't deny that at all. I don't think that this technique is the answer to predicting human reliability. I think it will predict some elements which are important. The approach that Jerry Lamb is going to take, hopefully will help us at a much finer level, or what Bob is doing. I don't believe there's just one way of doing things. I don't think there is one kind of data bank that's going to solve this. So much then, for the math analysis. Do you feel in any way the approach which Art is taking rests on assumptions which should first be investigated prior to the application?

Swain - If the thing works, if it really predicts maintenance repair time, so what!

Meister - For the limited set of questions that the technique addresses itself to, it's perfectly reasonable to accept factorial analysis. Everything follows logically.

Swain - If it works, Dave, that's all that counts, really. It has no utility for me because I have different questions that I'm going to answer.

Jenkins - This is the point. This technique is not designed for Art Siegel to use. It is designed for human factors to use.

Swain - If they are willing to answer such questions as what is the probability of maintenance success and not go much further than that. In other words not be concerned with the type of equipment behavior configuration, than fine. It certainly is not a tool to do the kind of design work that some of us do in working with system designers.

Meister - There is really a whole series of studies after the one that Art does.

Regulinski - I had a conceptual system on the board. I'd like very much to have the answer to this question, "What do you predict?" Art's program will obviously not do this. I think what the Navy should want from Art's program is a reasonable correlation between the simulated model and the practical model and if the Navy buys this thing at some level of confidence, then you may have a working possibility.

Swain - How would you apply this to new systems, though? Don't we have to first of all make a judgement as to whether or not the types of behavior involved are fairly much the same as those of systems we're using?

Jenkins - This is one of the reasons we want to extend it to electronics systems. It's the electronic systems which have not been used before. We're essentially saying, "What is the generality of the approach?"

Swain - Then you really would have to do a study and say take some other system.

Jenkins - He has no data at all on the system he's going to apply this to. He doesn't even know what they are.

Swain - If it doesn't predict it, you might then want to know why then you might go back and say, "Well, the reason it didn't work is that the behaviors are different."

Jenkins - If we take the system as it has been applied to and the classes of people, it would seem to make a huge difference.

Lamb - The designer is asking the question, "If I use automatic test equipment then we have a certain reliability factor because it's ah____"

Meister - No, no, no that's a mistake, because Art's factors are not tied to a particular type of test equipment.

Lamb - No, it isn't, but if I include this then it leads automatically to certain kinds of factors.

Siegel - You mean automatic test equipment is equipment operation. If you're depending on ammeters and so on that's related to another factor.

Jenkins - How about degrees of automatic equipment?

Siegel - We will have to look at the task specifically. If it's semi-automatic, we may be getting into an equipment operation factor with an electronic circuit analysis factor, or an electro repair factor, or something.

Meister - That inference must be made by the man who essentially utilizes it, by the analyst.

Siegel - This is why the reliability question is under consideration.

Meister - You mean reliability between analysts. Did you say you had investigated this?

Siegel - No. I said, Dave, that this was part of the present design -- to have independent analysts do it, and this is why we want to do the analyses before the validation because if we can't show inter-user reliability, we've got to straighten that out before we go on.

Swain - This is an interesting point. May I interject something here? When we first started using the AIR data store, two of us would independently make predictions and see if we agreed, and we agreed very well. Then when we moved away from the AIR data store to a more molar approach, the task-step level, we found we had not nearly as much agreement. Obviously this was because we were no longer essentially counting. So now what we do is take two or three (usually three) analysts for some rough problem and we try to make a more or less independent judgement and then we arrive at a concensus about where we disagree. So this is a committee approach.

Siegel - Maybe that's the way the technique should be used, through a committee approach. It's a content analytical problem, maybe that is the way to do it. You make sure you don't go ahead until you can get 98% agreement between your analysts.

Tolcott - If Art can take his 8 or 9 factors, however many, and show that people can reliably characterize these factors it could simplify the job. Now you don't have to get down to the step by step task analytic procedure that we always have to go through. This is a tremendous step forward.

Jenkins - Art, you can give an idea of your schedule.

Siegel - Let me say this in regard to scheduling. This program is not funded yet, so we haven't done a thing except write a paper and I did that, on my own time. Hopefully, we go to contract, say by September. That would mean we would have the reliability analyses and the computer simulation technique modified and the time estimates done in a year. We hope to spend the second year on the validation, so we're talking about a two year program from about September.

Jenkins - The contract is not negotiated, it could have been changed, it could have been cancelled.

Meister - But obviously you're going to have to go a second route along with Art I would assume.

Jenkins - I'm not sure I follow you.

Meister - In view of the fact that we've ascertained that while Art's technique was competent to answer certain questions, to be able to answer other types of questions, you're going to have to adopt at least one additional methodology to be able to answer those other questions.

Jenkins - Right, and this is where the work that Jerry will be doing comes in.

Meister - No, not really. Jerry is answering a specific set of questions, but he is not developing a technique per se that would enable him to go ahead and make predictions that will answer these system development questions.

Lamb - No, neither of us is answering the system design kinds of questions.

Jenkins - You're right, we have a hole in our program.

Swain - My point of view is that we already have techniques for doing reliability prediction work of the analytic type that is needed in design work, and now what we need is more data. We need a data store.

Jenkins - One of the things I hope to do this afternoon is, after you finish with what we have now, is to identify specifically these holes and get recommendations as to, for example go Alan's way or some other way because we have the capability of doing so.

Meister - Alan's technique, the Therpian structure, may very well be effective, but I think that before you adopt another approach like the Therpian approach, it ought to be just as critically analyzed as Art's structure; there may be holes in what you're doing.

Swain - There probably are, there are probably gaps, but my point is you can use it now if you answer some important questions now but it would be even better if there were better data. Any other techniques will essentially use the same kind of data.

Meister - When you are able to answer specifically the criteria questions that I've exposed to you, then we can say that your technique is appropriate.

Swain - I don't understand. You haven't come up with anything, so far, that we are not able to do. We can have some trouble with say maintainability perhaps when you can't come up with a reliability estimate, a probability estimate. But for all the work that I do, we're able to come up with probability estimates.

Meister - I'm not convinced, myself.

Blanchard - I can't stress too much that I think there needs to be a systematic, thorough, evaluation soon of current technology. This needs to be done with factors such as Dave's, which need to be amplified and studied very carefully. As I said before, such an evaluation has to be cast against a set of objectives. You have to know where you want to be, what you want to be able to do at various points in time and then lay this on available technology.

Jenkins - I think everyone agrees with that at the moment. To be very explicit this might be the first new task of the program.

A HUMAN PERFORMANCE DATA BANK FOR COMMAND CONTROL

Richard Coburn

Naval Electronics Laboratory Center

A HUMAN PERFORMANCE DATA BANK FOR COMMAND CONTROL

Richard Coburn
Head, Human Factors Technology Division
Naval Electronics Laboratory Center
San Diego, California

Prepared for
Human Reliability Workshop
Naval Ship Systems Command
Washington, D. C.
22-23 July 1970

A HUMAN PERFORMANCE DATA BANK FOR COMMAND CONTROL

by
Richard Coburn
Head, Human Factors Technology Division
Naval Electr›nics Laboratory Center

The title of this paper may be a little misleading because I cover quite a bit more than just the data bank idea. Included is a discussion of our position on human reliability in systems development, the implications of this position for data bank development, and what we will be trying to accomplish in the near future.

First, to introduce our position on human reliability I would like to step out of my field and talk about the manned space program. Going back to the Mercury program, we note that the original design philosophy was to make all critical functions fully automatic, with man along primarily as passenger and observer. What could be more logical than this, once one realizes that man's reliability could never approach that of the machine? The record shows, however, that during each of the Mercury orbital flights man had to intervene because of one or more failures in the automatic systems. In succeeding space programs greater human involvement has been deliberately designed in from the beginning.

Now to me this seems to raise a question as to the applicability of the reliability concept when it comes to human functions. If reliability is "the probability that an item will perform its intended function for a specified interval under stated conditions," we see that our Mercury spacemen were indeed "unreliable." Their intended purpose was simply passenger/observer; but they freely abandoned this purpose to become intimately involved in operation and maintenance functions. Now even though this change of roles is commendable it bespeaks unreliable behavior in the strict sense.

Looking at human performance generally the one thing which is most characteristic--its extreme variability--is the very thing which makes man look pretty bad when it comes to reliability. In this connection it should be noted that man is capable of generating erratic performance at any time. Unlike machine performance which will usually stay within nicely defined tolerance limits until old age or some other catastrophe sets in, man's performance can and does bounce out of tolerance quite unexpectedly, and as often as not will then resume immediately within tolerance. Again this is different from the pattern of machine behavior, which once out of tolerance generally stays that way until repaired or conditions are altered.

If, as is typical, variances for machine performance are small, the system designer can add sizable safety factors so performance remains within tolerance until systematic degradation occurs. With human functions this cannot be done, because the real-world variances in human behavior are too large considering the impact of all the psychological, physiological, and environmental factors affecting human performance.

In view of these kinds of differences we question whether we really make any headway by trying to apply equipment-based reliability concepts to the human operator. Of course, if we could do this successfully, it should make it possible to generate meaningful total system reliability figures. This would certainly be a desirable objective, but we feel it is simply out of reach for the present anyway.

We are assuming that when ADO 43-13X speaks of human reliability, it is not doing so in a strict and narrow sense, but rather is concerned with the predictability of human performance in every meaningful way. The really important thing, of course, is to end up with systems which do the job they were intended to do whether or not they have "reliable" people in them. This means we must have good design methodology and good data to use in our analyses. The methodology must include means of selecting between design alternatives as well as means of predicting performance.

With this preface about our views on human reliability, I would like to move on to what we propose to do. This is to undertake work leading to a human performance data bank for command control. With the thinking and advance work we have done about such a data bank we feel there are practical things which can and should be done without getting in over one's head. Perhaps the first point which should be made is that we feel it essential to look both ways before we commit ourselves to a specific data bank concept. That is, we need to look beyond the data bank at the way in which it could actually be used in a practical sense; and, of course, we must plan for only that data which is feasible to obtain. Moreover, we believe that the data bank and the methodology for its use must be concurrently developed in order to keep from getting lopsided results. And, needless to say, we can quite properly accept somewhat limited goals as a starter and then expand them on an iterative basis as appropriate.

How could a data bank be used by the practicing human engineer? Our preliminary thoughts on this question might be characterized as follows:

1. The human engineer does not attempt to predict man-machine reliability--he admits he doesn't know how to do it now and may never be able to do it.

2. Instead he makes appropriate analyses and advises the designer of:

a. The best man-machine performance which could reasonably be expected from the proposed system.

b. The way in which performance may be expected to change as load builds up.

c. What kinds of errors might be expected if the concept is implemented as planned.

d. How the design could be altered to minimize occurrence of error, to maximize probability of error detection, and to minimize system perturbations due to uncorrected errors.

e. Reasonable safety factors for time and accuracy of performance.

So far as the data bank is concerned then, it should contain the minimum data which will enable the human engineer to perform as indicated above. Generally the needed data are probably not available in a form ready for use, so it will be necessary to obtain it.

Two approaches to obtaining the needed data may be considered--empirical measurement and judgments of experts. A combination of the two approaches is seen as the only practical way to go, for we obviously cannot directly measure everything on which we might want data. On the other hand even if judgmental techniques prove to be very effective some empirical data are needed to anchor the judgmental data. Whether by measurement, judgment techniques, or a combination of the two we need to obtain the following types of data as a minimum:

1. Time-of-performance data for each critical function performed under each mode of operation at each operator station within each type of command control system to be included in the data bank. The operator to be measured will be the one recognized as the best in the business, and he will operate under the most favorable of conditions. This will provide the upper limit human performance data.

2. Variation in performance time and error with load for each of the same functions used in (a) above, as performed by an "average" operator. Error data should include both type and frequency of error. We regard type-of-error data to be particularly important for it is absolutely essential in order to be able to minimize impact of error on the system through design. Obtaining these data for an "average" operator as opposed to the "best" operator is more productive because more error data will be generated, and the impact of load may be more apparent. (It will be particularly important to identify those functions for which the operator becomes disoriented under load and generates high error rates.)

Thus it is proposed that initially we might take only two cuts at performance measurement for a given function: (1) the best operator operating under ideal conditions to obtain best cut and (2) the dynamic characteristic of an average (modal) operator under load. These two cuts will, of course, not completely define the range and characteristics of performance, but they will get us started. We consider it overly ambitious to attempt to gather definitive data on the impact of the endless list of situational variables such as operator capability and training, equipment state, shipboard morale, sea state, type of ship, etc.

Our early efforts will be a 3-pronged approach. First we propose to do more thinking about just what kind of data we wish to take, keeping in mind the considerations mentioned above. Second, we will look specifically at the feasibility of instrumenting NTDS to record selected human performance data as mediated by the console controls. Third, we will investigate the reliability, validity, and utility of a judgmental approach to establishing human performance bench marks and correction factors to apply to our empirically-derived data.

With respect to NTDS performance we envision the possibility of recording at sea by tapping terminals in the central pulse amplifier which samples all console outputs. These outputs are in the form of 30-bit parallel words, and it is proposed to convert them to serial bit streams and record them on magnetic tape for transmittal back to the laboratory. At the computer center the tapes would be processed as necessary for input to and analysis by IBM 360/65. What would go into the tape would be the words generated by button-pushing actions which would identify the particular function code, the originating console, and its mode of operation. Clock pulses could also be recorded on this tape if they should be needed. Since much of the buttonpushing activity is target-oriented, it would be necessary to relate track number to the console output words. This we believe would be available in the computer messages to the auxiliary readout which contains track number information. Presumably the recorded events could be related to the major exercise event times, thereby allowing us to determine elapsed time between events in the real world and the man-machine response to these events.

Once we obtain the data for the two cuts I described earlier (performance ceiling and average operator load characteristic) we will need a way to make it more universally applicable. This is one place we are hoping the judgmental technique can help. Specifically we will attempt to establish correction factors for a number of important situational variables using judgmental techniques. These variables would include such things as amount of training, type of ship, sea state, and miscellaneous operator characteristics. Needless to say, our initial work will include specific attention to evolving a workable performance-prediction methodology applicable at various points in the life cycle for various types of development efforts. The methodology should be capable of being readily used by either Navy or contractor human engineering personnel.

In summary, we are advocating deferral of a comprehensive, all-inclusive attack on the human reliability data bank and methodology in favor of a shorter term effort with limited objectives and in the command control area. We propose to work with simple time and error data obtained under two cuts (a static "best" and a dynamic "modal") using the most effective combination of empirical measurement and judgmental techniques. Applicability of data would be extended by "correction factors" for selective parameters, again derived by the judgmental technique. If this modest approach is successful, then we may feel it would be warranted to extend the effort to other types of systems and to attempt a rigorous tie-in with system effectiveness formulations.

DISCUSSION OF COBURN's PRESENTATION

Siegel - I question whether we need more research into these matters which have been looked at by Stevens and the psycho-physicists and infinitem -- to the point of where we don't have to deal with it.

Harris - I'm definitely not proposing more research into the business of method, however I'm interested to find out whether we can get reliability of the relative frequency of occurrence of certain kinds of errors. I think so. There are several requirements that have to be filled. What is the measure going to be, what could it be? The kinds of measures that we end up with are for equal values of an equal interval scale. My question is, what more do you need? For the purpose that I'm talking here, let me make this clear. I'm talking about for the use of the design engineer.

Meister - Then you have already explored potential uses for this data bank, and you have determined that this measure which you have not yet fully explored will in fact answer the questions of this potential user?

Harris - Certainly, I don't know this; if I did I would be more dogmatic, and I would present the evidence. What I've determined in the use of this method is that in a certain situation it applies and it applies very well. I have satisfied myself and I think it will satisfy all of you about the reliability and the validity of the methods in a particular application.

Siegel - Why did you go to this ranking method, as opposed to a ratio estimation or a magnitude estimation?

Harris - One major reason is it's easc of application. People can do it and they don't mind doing it.

Siegel - Temporarily, it may be okay, but when you're in the ranking method the perceptual task of the rater is to keep k items perceptually in front of him. When you're in a comparison, the perceptual task is to keep two items in front of him.

Harris - We all know Guilford's argument that people are going to respond to composite standards when they're ranking, a large number, and in fact he even further argues that they are really making a kind of paired comparison. Well, you know that they really aren't - they're not making all these paired comparisons. What they develop from viewing these situations are just composite standards and that these things are now ranked.

Siegel - Why doesn't he carry two methods along. Take category and magnitude methods and see where you go?

Meister - It seems to me that you're putting the cart before the horse. Presumably your goal is ultimately to develop a data bank, performance data of some sort. You have to start with the uses of the data bank, the elements, the assumptions, the definitions and so forth.

Harris - You can tell me right now what the uses of the data bank are, can you not?

Meister - I can specify certain potential uses. The question is whether the elements which you're going to throw in the data bank will in fact meet these uses. You're apparently already committed tc a methodology without having explored all of the preliminary choices.

Coburn - I think the only commitment, Dave, is that we don't want to explore the judgemental approach. I think, at this point, he favors to do this, but he's not committed beyond that. Another thing we do is plan to have an initial phase where we are looking at what we want the data bank for, questions we want it to answer, and the taxonomy and scoping the whole effort. Obviously we don't want to run off and start right today.

Swain - You'll definitely be coming up with an interval scale, a..d you can compare that interval scale with the ratio scale which you get on the actual frequency counts that you're able to make and that's what you may use for validation.

Meister - I'm not objecting to the particular rating methods, judgements, or what have you. All I'm saying is, that it seems a little premature before you've gone through the preliminary conceptualization of your data bank elements to have already decided on the experimental methodology.

Harris - I disagree with you, for this reason. We already have a task taxonomy.

Meister - Remember what I said before that your data bank is a microcosm of your entire technique. Here we seem to have a situation in which a data bank is being developed before its use and elements are explo ed. The methodology is being derived prior to the development of the preliminary generation of the elements of it.

Harris - You didn't attend to my first statement. What I view to be the steps on this kind of thing - first of all I'd better understand the system a great deal better than I do now, I must do a task analysis on the many operator and user functions that are involved in that sort of thing. From this we'd like to start thinking what our taxonomy is, classification of all this kind of performance. I think we have many different performances on all essentially same people that took them. We try to look at variations of performance as a function of equipment in this instance. We want to consider the task analysis and so forth, before we ever start any data collection. You've got to have a classification scheme, you've got to have an error scheme, you've got to have some knowledge of the kinds of errors and so on.

Swain - He's going to be describing them just like you would run an experiment. He'll describe all the various performance factors and so on that he thinks are relevant. It just all boils down to the great difficulty of generating an operational data bank. We have experimental methods for taking advantage of the studies that people have done - things that are in the literature. I feel that the collection of this data ultimately has to be organized.

Meister - This data bank will be a special purpose data bank in the sense that it will apply specifically and largely to command and control systems of the NTDS type?

Harris - That's correct, that's initially the case. That's the only system we're looking at.

Meister - I know that, but are you planning to expand this data base again using the NTDS as a model?

Coburn - That's a more difficult step and I think we're going to have to see how this first thing goes. We do want to handle the old command and control data problem later on but I think it's perhaps premature now to say whether we can extrapolate from the NTDS but it's a very good question. We've pondered this ourselves.

Tolcott - The purpose of the data collection is not to develop a data bank per se, what we've been talking about here is the development of the data bank in our area of concern to enable us to predict new situations for which you haven't collected data. Obviously the ones that collect data make this a lifetime career. But we really want to be able to predict. In your data collection effort there ought to be a way of dealing with the data which associates characteristics to it so that you could go to the next step and predict even if you're only limited to a command and control situation.

Coburn - There may be some misunderstanding here about the NTDS, too. It might sound like this is just an exercise to go and use NTDS because it is available. Remember this, that NTDS components form the basis of many new system configurations, so this data that is obtained from NTDS could be of value in these new systems which are going to use NTDS configurations. Now that doesn't answer all your command and control problems.

Meister - It will not answer the questions which will be raised about the totality of Navy equipments.

Jenkins - If I could make a comment on that. The NTDS system is not just a small unit in CIC. It does encompass all of your AIR, your ASW, your ECM, your total electromagnetic sensors. It's the total sensor information center outside of the hull of the ship. You can take data, for example, and apply it to a sonar system, or a radar system, or what have you, because they all form part of the total complex here. I think by attacking the NTDS, you're attacking probably somewhere around 80% of all of the kinds of operability problems.

Harris - You have to look at the performances which are involved in this thing.

Meister - Then I say, what you are implying is that ultimately the data bank for this kind of development will be capable of being utilized as a general Navy electronics data bank predictive structure. If that is the case, you have to look very carefully at the elements that go into it because you are assuming, when you do this, whatever human reliability methodology or structure is inherent in that data bank. You see, you're just not developing a data bank, you are accepting the methodology which is implicit in the data bank.

Blanchard - That's not generally true.

Harris - I view that the operational data bank is going to encounter problems in the operational system that are pretty much the operational problem. This generality is going tc depend on lots of things, one is the accuracy of the taxonomy, the classification of these performances. You would want to apply it to other similar systems. The goal is definitely a generalized data bank.

Meister - Then I come back to the same question that I asked before, Number one will you be able to associate your performance data with certain specified characteristics of equipment?

Harris - This is the intent - let me tell you how I'd go about it. When you apply this data let's say to error data, you have some confidence in what you have and you want to apply it to another, and maybe you would use some validation of it, but let's leave that aside for the moment - to the extent that you can do the sorts of thing you're talking about like making specific statements about design alternatives depends upon the nature of the system to which it is applied. Before you come up with design recommendations based on the characteristics of the data bank of the information that you obtain under these different applications, you need much analysis.

Meister - If you have a certain time, a certain error associated with the operators' pushing a button, let's say, you will, in your data bank, be able to associate live performance data with some characteristics of the button pushing activity in the equipment.

Coburn - It really isn't a button pushing activity. Button pushing only marks the end of an operation. Maybe it's identification, for example, entering an identification. The significant thing is not the button pushing.

Meister - Okay, well whatever it is.

Tolcott - It's the situational conditions under which the task is set. You have to at least be able to specify.

Harris - And this is the burden for the taxonomy we're talking about the molecular the molar and so on. This is obviously a problem of taxonomy. You've got to have the task identified at an appropriate level.

Swain - He's going to have to describe all the performance shaping factors, etc., as well as he can, and if I were designing the future CIC I'd love to have error rate and all the descriptive data that he apparently will be collecting. That would be better than what I deal with now.

Meister - You are also going to have to consider, although you may not deal with it directly, how you will combine items of data which apply to a particular stimulus, let's say in one sequence, how you combine that data with data reflective of performance relationships with some other stimulus. This is a serious problem which is also inherent in your project as well.

Harris - There's no question about that. It comes down to the question, How do you use that stuff when you get it? You've got to consider that before you start.

Jenkins - Perhaps this would clarify one point here. When we first began exploring command and control requirements of the data bank, NEL came in with a suggestion that basically said we would be passing off times to the computer and we kicked this thing around for awhile and we said that we really don't know all the ramifications involved in the task. Frankly, for the first six months or maybe even longer we're going to sit down and talk about potential approaches and see what are the assumptions we have to make and what are the implications. This is the first time NEL has come back and said here's a first cut; only some months from now will they say, "Here's what we're going to do." So this is not to be considered final.

Harris - There is one thing I would like to make clear. It may well be only when we get down to necessary things, now the element descriptions are the error descriptions, we may find that some kind of categorical judgement may be a better approach that is where you have a descriptive term for each category.

Blanchard - I want to make a suggestion, Bill. One thing that we used once very effectively is a mix between ranking and paired comparison techniques. One reason you'd like to use a variance technique like paired comparisons is because you can test the property of transitivity. You cannot do that with ranking. You want to make sure you don't have any specific variance from the stimulus itself being introduced into the situation. A neat thing to do is to first rank order your stimuli so you have a fairly good understanding of the ordinal relation and then set up a partial pair comparison scheme which eliminates pairing extreme stimuli. This eliminates a lot of the labor involved in complete pairing.

Meister - When we were talking about Art's approach and it was agreed upon that Art's approach would take care of a certain set of questions, we agreed that there would have to be another approach to supplement his. Is that what Dick's data bank will supply?

Jenkins - It's not necessarily intended to do so, but if it does fine. After this presentation is over I thought that we could lay out what is the agreement we reached on whether, in this fiscal year, we should do something else. For example, compare these various methods as you or Bob has suggested or should we hold back and say well let's see what we're going to get a year from now. Again we're talking about a 5 year program.

Swain - The results of this method are not restricted at all in the sense that the results of what Siegel is doing are restricted. They're going to come up with data that I, personally would like to have if I were trying to help design a system like this, or make reliability predictions for something like a CIC system - I could use it.

Meister - I'm not sure you could make that kind of judgement, Alan, because we don't know enough about the details of the system that they are developing.

Mills - You mean also that you could use it within this command and control situation.

Swain - Yes.

Mills - There's one thing that has been sort of nagging at me, but I haven't got the slightest idea why, and that is the use of this superman and then the average man. You're really not dealing with the average "man" in this sense, you're dealing with somebody considered better than average. Trying to design a system based on your data collected from a super average man. I don't know what it is but something about that bothers me.

Swain - Anytime you try to use the results of any study or experiment you look at whether these conditions generalize what you're trying to do.

Mills - The main point that I wanted to make was in terms of attempting to generalize any further than the command and control situation. The other point was that I cannot in my own mind, right now, figure out what it is that bothers me about it.

Connery - I'd like to respond to that if I may. I'll go back and review a little here. We've been talking about Human reliability, we've been talking about people reliability but nobody here in the two days that I've heard talked about sailor reliability. I propose for your consideration that sailors are a unique breed of cat. They are different from soldiers and they are different from airman in the Air Force. Bob was implying that but he didn't say it. I want to be sure that we are aware of it. To be sure soldiers, sailors and airmen are more alike than they are different but there are some real distinct differences among them as groups of people. A good part of these differences can be accounted for by their mission-oriented training. But this by far doesn't account for all of it. When we're concerned here about a data bank as you suggested, Dave, if you're going to build one, if it's going to be pertinent to the Navy, it's got to deal with sailors. One thing that disturbs me about all this talk about building a data bank is how we're going to go about doing it and implementing it. The reason I say this is because right now I've been sitting here trying to recall and figure out how many data collection systems we have ongoing in the Navy. We have the 3M system, FADAP, FODAP, the ILS system, the one at Pensacola. These are all efforts to build data banks about both operational and people performance. About two years ago CNO, Admiral Moore, put out a directive and as far as I know it still stands which says there will be no more data banks, we've got enough already. The fleet commanders are fed up having people coming out, collecting data.

Swain - That's an operational thing.

Connery - That's what he's talking about. However, an opening was left to enter any one of these current data systems, in existence. I propose for your consideration it may already well be much of the information that will be helpful to our group available in today's data banks. The 3M deals with maintenance of equipments, FADAP is the Fleet ASW Data Acquisition Program, FODAP is the Fleet Operational Data Acquisition Program. There is the Integrated Logistics System, there's a data bank. All of these in one way or another deal to some extent with people, (sailor) performance both individual and group. You know if they don't and you want something, it would be far easier to provide inputs to those acquisition programs and to have to retrieve it later than try to build a new data bank where you have to rely on fleet support to collect the data for you.

Swain - That may well be true but when you say data bank, we don't know what you mean. Are they collecting human error rate information, for example?

Connery - You can infer human error rate. You can definitely get equipment error rate. There is a lot of stuff that we would like to have and that includes the development. But rather than start a new system to be built by fleet people, it's better to get into one that already exists.

Coburn - I think the FADAP program has evolved into the OPDATS program has it not? Operational Data System. One point to note is that OPDATS is not funded for this year. That's a slight handicap, it just lost its money. A second point about OPDATS is in looking into it we couldn't see that you could get any fine grained, any really meaningful human performance data at all. It was at too gross a level for the kind of thing that they were doing. The other point is the way that we're talking about trying to get the NTDS data will impose minimum service upon the ship because we get in and we can tap off at the central pulse amplifier, it doesn't require modification of equipment, it doesn't require attendance by maintenance personnel, or anything of this sort, once it's done. If we can't do that, we would say it's infeasible because we know we can't go out and upset the ship's operation. I recognize that. That's why we're proposing to go this particular route. We think that maybe we've got a chance to do something this way. If we can go in and tap into the system without disrupting their operations, we can get our data and come back and analyze it.

Jenkins - We've got to establish a rapport with the working level technician. You're not there imposing yourself on the enlisted men who say there's sand crab been thrown on us.

Connery - It's not just sand crabs, even uniform psychologists go out and they get the same treatment.

Tolcott - Mike you're saying two things here which are mutually inconsistent. You're saying no more data banks let's use what we have, then you're saying these are no damn good. We know that's true of the 3M system. We've got a form to collect data on maintenance activities and it isn't even designed for example, to give you data on the level of personnel who did the maintenance action. You do know the allocation of time, but you don't know who did the maintenance action, and for several years now there have been recommendations to include that additional item of information. Not the person's name but the rate, the level and so on. It's not in the system yet. You can't use it.

Mills - Can I ask one thing here. On this one data bank that you have just described, I don't know just what the acronym for it was.

Coburn - OPDATS.

Mills - Yes, can you give me an example of the kinds of data that are in this system?

Coburn - The general purpose of OPDATS, as I understand it, is to be able to reconstruct an exercise. OPDATS involves taking data manually and automatically where possible during complex fleet exercises and then reconstructing the exercise after the fact. Unfortunately, OPDATS does not provide any very fine grained information about what's going on inside the ship, like the human performance which is involved. It just can't get down to that level. It's an operations research kind of thing.

Meister - I don't think anybody will disagree with the idea of a development data bank, on the basis of gathering performance data at sea - it's an entirely worthwhile one. I would suggest, that after you do your preliminary conceptualization of the structure of the data bank, however long it may take, three months, six months or whatever, that you come back to a group perhaps constituted such as this one to examine critically what structure you have imposed or what structure you are proposing for your data bank system. Obviously at this stage of the game you have a lot of spade work to do before you can begin to describe the outlines of these data bank structures. All of the requirements should be considered very carefully.

Jenkins - Before we close off this discussion, Dick do you have a schedule of this work?

Coburn - This preliminary phase where we're trying to scope the effort, review your taxonomy and what not should be three to six months, but more like six months. At the same time we can be pursuing the question of the feasibility of the NTDS. I would say by the end of six months we could certainly have that in hand and would know how feasible it would be to go ahead and implement it. A year of work, I would say, beyond would get an initial set of results.

GENERAL DISCUSSION

Jenkins - The initial objective of the program was first to validate a certain human reliability method, to integrate or determine the feasibility of integrating human reliability predictions with equipment predictions and finally to develop a data bank with emphasis on command and control. That has been the program we have presented and it's not intended to answer a number of responses that have come up. These were the initial objectives at the beginning of the program to start things off.

Regulinski - I thought prediction was the initial objective.

Jenkins - That's a global objective. We should spend a moment on the funding position we are in to accomplish initially these objectives and any others you might add. In fiscal 70 we had $75K, 25 went to NUSC, 15 went to NEL, 5 went to HFR and 20 went to APS.

Tolcott - Jim, are those the total numbers for the total ADO?

Jenkins - Only for the human reliability program. We had to compete with the other four elements of the human engineering ADO. In Fiscal '71' the initial allocation was 20 to NUSC, between 15 to 35 to NEL, 25 to HFR and none to APS. In '72' we don't know what they will be. It would be a total between $125K and $150K focused on human reliability with 25 going to APS. The reason being that this contract is not signed. It won't begin until fiscal '71' therefore I'm using fiscal '70' to really start '71'.

Meister - Have you got that money committed?

Jenkins - To where?

Meister - Just where you said.

Jenkins - We just started the fiscal year and I have not yet started the paper work. So far as '71' money is concerned, here in this area we can take $20K and start another effort. It would delay Dick's program to some extent. I would not like to do that, and yet if the group believes that we must start on another task, for example, comparison among the models, then this is the trade-off we might have to do. For the rest of the years the funding rapidly increases: $125 to $150 in '72' and $135 to $170 in '73', '74' is $175 to $225, then it gets astronomical $350 in '75' and $300 in '76'.

Meister - Is this just reflecting the rate of inflation?

Blanchard - Jim, was the $300K we heard for the total ADO for fiscal '71'?

Jenkins - In '71' the total ADO is $350.

Connery - There still remains a question as to whether there's deferral on that. I think not. There wasn't this morning.

Jenkins - This is the resources we have, to meet these and any other objectives.

Connery - Jim, I've got to have deliverable products out of that before 1976.

Jenkins - You will

Connery - When it comes to whether or not to divert $20K from NEL this year and start a new project, I would suggest that you concentrate on what you've got already. Once you've got the deliverable product we can start selling it and get that further money in the out years.

Jenkins - You get a deliverable product at the end of '71' from NUSC, you'll get a deliverable product in '72' if APS's work is completed; you'll get a handbook on how to do the validation which is written in terms of any system user, you get computer programs, you'll get a complete data bank of the kind of data and the nature of the data as to how it was collected. In '73' you would have a first cut from NEL and HFR on errors made with respect to time. From then on I can't identify the end product.

Connery - In 1973 you can deliver those to my relief.

Jenkins - Marty has brought up to me that we seem to have ignored the question of model generations almost completely.

Tolcott - No, that's not what I said.

Jenkins - All right say what you said.

Tolcott - It seems to me that the emphasis during the past few days has been on the generation of data banks of various types and the kind of data that will be going into them. In many cases I feel we're ignoring the objective of the data bank itself which to my mind is the prediction of reliability. I mentioned before that I thought that perhaps data banks are not the only answers in prediction of human reliability. Perhaps some thought ought to be given to just what other techniques are available. For example, you may want to predict human performance in new situations in which measurements have not been made. We might in the long run find that it's more economical to think of methods of sètting up tests to get a few measures in the laboratory and generating a distribution from that and we'd be in a lot better shape than if we spent years and years pouring data into data banks.

Meister - It's probably ancillary.

Tolcott - It's only ancillary if the objective is to develop a data bank. It's not ancillary if the objective is to predict human performance in your system.

Meister - There is no question that one can run quick and dirty studies. These have been done on system development projects for years now.

Tolcott - Data banks have been developed for years, too.

Meister - It's a far more complex problem.

Swain - There has never been a reasonable data bank developed from my point of view.

Meister - I can see your point of view.

Tolcott - There have been some very smart people trying to develop data banks, and if these turn out to be not reasonable, I might find myself in a position of saying well what makes us think that the next data bank is going to be any more reasonable.

Meister - The data bank that has been available has been useful within the limitations of its data. The problem has not been the data bank per se has been ineffective but that the content of the data bank has been lacking, because there has been actually not a sufficient amount of effort.

Siegel - I think that my feeling would be generally in terms of what Marty is saying. Maybe not as strong but the overemphasis on data banks as compared and contrasted (in terms of relative amount of time spent here anyway) on other tools, leaves me a little bit frightened in terms of the total global program. It's like saying there's only one type of research we need in computer aided instruction; software. We need machinery, we need software, we need methods.

Meister - That's not true, Art. The data bank is simply an output. How you do the research to get data inputs - there are many variables, subjective judgement, collecting data at sea, your own techniques. All will provide inputs to a data bank. As I said, the data bank is just a microcosm of an entire predictive structure. I suppose the reason people have fastened on the data bank concept is because it is relatively concrete - it's an output. There is a tremendous amount of preliminary work before one develops a very effective data bank.

Tolcott - The data bank is not the output. It is an output, but it is not the output that achieves the objectives.

Swain - I think there has been an under-emphasis on data banks. People in the human factors area have been hollering long and loud for many years for a data bank. We need it. Now if we're going to talk about some idealistic data bank, sure we need years of research, but my point, as a practical person, is we can start collecting data now. For example, in the work that HFR and NEL are doing, they can be developing a lot of very good work on showing what kinds of data need to go in the data bank within this CIC type structure. Great, I'm all for it.

Blanchard - Marty, I think what happened here probably was a manifestation of frustration on our part. People who have been involved in model development and who have tried to go out and apply these models have been constantly frustrated with the lack of good data to utilize them. My point before was that I think the application of models has been seriously hindered because we haven't had good data, even marginally acceptable data.

Siegel - With the modeling techniques we have, and the arithmetic we have available, and the assumptions these are based on, do we need that degree of finesse in our data?

Swain - I'm not advocating a lot of finesse, I just want some data.

Meister - All of this questioning really reflects the fact that there really has not been any systematic examination of the alternative approaches within this field, and frankly it strikes me as being rather extraordinary that one would attempt to develop an overall program before doing an analysis determination where we are in the state of the art. I don't think anybody knows. I think that such a comparative analysis of the alternatives might very well point out that, as you said, there are alternatives to what we have in fact been talking about. But we don't really know until we do this and I get the impression that everybody's got his own personal horse and is just riding down the road in sixteen different directions because it is his horse. If you look at this thing from a standpoint of a total program for the Navy or for the government as a whole, then we ought to take a total look at what we have and what we don't have.

Lamb - Along that line, I think we need to see what level of prediction would be required into the systems engineering equation.

Regulinski - I think that Bob's and Dave's comments also mirror the frustrations of the system engineer who has turned to Monte Carlo simulation methods for lack of human performance reliability data relevant to his models. He may have very well reasoned "I do not have your human reliability data, and since you can not provide me with the relevant data, I will simulate my own." It is a fact that he succeeded in simulating such data, and whether or not you eventually will provide him with a bank of data, he will continue to use Monte Carlo simulation because he has developed the ability and faith in the method.

Swain - Your're speaking from one point of view now about the system engineers of the breed that you describe. But the people that I deal with don't require that much sophistication; they don't have it themselves. They don't use it themselves.

Siegel - Are you saying, Thad, that we ought to or the Navy ought to in a general program, place a little more emphasis on Monte Carlo methods? The total research picture may include a greater emphasis on these Monte Carlo methods than has come across here. It may be that a useful input to the program would be some type of Monte Carlo simulation which could dovetail with some physical modeling or digital Monte Carloing of the system, so that at the end you do have a marriage of a human reliability and an equipment reliability Monte Carlo simulation. It's a worthwhile goal as a long range objective.

Lamb - We need some method of predicting that is divorced in some sense from the system and work on a functional level as opposed to an equipment level that may be an upgrading of the physical status of Monte Carlo or some other model, but we don't know what's required.

Siegel - Thad can speak for himself, I don't have to speak for him, but I think he would argue that we don't have to wait for the development of a data bank to go ahead with these other probes.

Meister - Again you're reflecting a particular position. My position is, and that's the one I'm going to reflect, we don't know enough about what we have represented in this entire room to be able to say yes, you're right and we should emphasize Monte Carlo, or Alan's right, we should emphasize data banks.

Siegel - Let's emphasize both.

Meister - Maybe both or maybe none.

Siegel - And other things.

Meister - The point is if you don't do some sort of analytic examination, then whichever road you're going to take is going to be based solely on the personal persuasiveness of the individual who likes that road and if you've managed to influence Jim, somebody else managed to influence Jim or Dick or whoever, that's the road that will be taken. That's not necessarily the optimum road. But before you decide on the road to go, you have to look at what you already have. I'm not pointing at your effort in any negative sense, I'm making a general statement. What we have seen today and what we have seen in other meetings of this sort is that everybody's got his own little horse and they're riding their horse and they'll kill anybody who gets in their way.

Blanchard - Do we really know what we already have?

Meister - I don't think we do. I don't know what you've got in your mind. I know very well that if I were to sit down with you for a day or two days and start asking question, I would have a dozen questions you couldn't possibly answer, and the same thing is true of Alan, and it would be true of myself, if one of the others would ask me those questions.

Harris - What is the source of this information. We don't know. Are you saying that we have not read the literature or what?

Meister - I'm saying that various approaches have over the years been developed. There have been an awful lot of implicit assumptions made, implicit methodology which you can't tease out at paper sessions because these people won't really tell you until you actually nail them to the wall and say, "you did this; now how did you really do it? There's an awful lot of that and it runs rampant throughout all of the methodologies that I've heard expressed to date. It's a natural tendency, we're all sons of bitches, and we all like to make sure that our private parts are suitably covered but that's not the point. If you're going to develop an overall approach and make a selection based on reasonable grounds then you'd better uncover the private parts and see who the hell's a boy and who the hell's a girl.

Jenkins - You're saying that rather than reflecting our effort in the development of refinement of methods of prediction we should give a good hard look at what we have.

Meister - I would say so. I recognize that there are practical problems since you have already started down a road with a number of efforts and I wouldn't suggest you eliminate those efforts.

Mills - There is also a problem with how you're going to go about doing this evaluation. If you take any one individual around the table here you're going to run into the problem of vested interest, not necessarily economic but ideological. What Dave is really saying here in a sense is that we sort of need a knock down and drag out fight amongst the people who have really developed these techniques and use them, not a superficial type of thing. Maybe split the contract up or call a one week workshop or something like this to reveal the best way to go.

Blanchard - You're trying to get objectively into this thing, to develop a methodology that everybody agrees on and we can look at, and make some judgement as to whether it's offering a reasonable solution. I think we need to establish some objectives. Somebody's got to write down where we're supposed to be in five years.

Lamb - Not just where you're going to be, but where you're going to go and to what level and with whom.

Blanchard - A detailed statement of objective first. Then we need an expanded version of what Dave has started on the board and perhaps more. The objectives would provide a basis for weighting the evaluation criteria. Then, when an individual makes the evaluation, he can present it and let everybody sit around and hack at it. "Say, look, you misinterpreted this or you misinterpreted those data." or "Hey, that's not right," or "Why did you rate those criteria so low?"

Mills - I want to add one thing here about the impartial observer type of thing - this is really my point. You get an impartial observer in here, one without idealogical interests and almost by definition he doesn't know the area and therefore, may not be competent to perform such an evaluation.

Meister - No, I don't buy that Bob.

Tolcott - Why do we want human reliability methods? What do we want to be able to do with them, once we have them? We're going to get a data bank for the command and control system within a couple of years, let's get some problems into that data bank and see whether that data that we're collecting can answer the questions.

Mills - That's exactly what we're doing on this one contract, and there's no difficulty. I'm talking about the problem that Dave is discussing, and that is, trying to evaluate the state of the art, and trying to determine what techniques woulc do what and what they won't do.

Meister - Undoubtedly any evaluator (and I assume he has to know something about the field because I can't imagine a complete ignoramus starting from scratch) but any evaluator is going to be somewhat biased by his experience and what he reads. You simply have to make your criteria and your methods of evaluation as objective as possible. I would say this, if your evaluator comes in and makes a horrendous error, we've got to pull the reins quick. That's perfectly obvious from what is happening in these two days. So he's not likely to get away with any sort of implicit bias, there are toc many people watching to make sure he doesn't have that bias.

Swain - I'd like to have the review be in writing, though, because I'm sometimes at a verbal disadvantage, here - I can't get a word in edgewise. If you can't do it in writing, it's not much good anyway.

Jenkins - Would you say then that what we should do this year is to go out on RFP, with our own sole source contract to one of you here or to someone else (but I don't really know of anyone else) and task them to establish objective criteria and methods of evaluation, and having done that, have those methods and criteria approved by this group, then apply those to the various prediction methods we have how and assess them, then come back together and see where we stand?

Meister - I'll buy that.

Lamb - I think that one of the things that we are getting to that this evaluator would need, is that you, Jim, would come up with more explicit goals out of the ADO.

Blanchard - That's right.

Lamb - That's the first step before you can go to an RFP or sole source contract or to someone, maybe an evaluator or maybe another way, and that's to say how we are going to use this. That set our criteria and everything else.

Jenkins - We can establish explicit goals such as this effort should produce for the design engineer and specific recommendations relating to hardware and software design. Second, the method or technique must relate to specific skills and knowledges for training or to selection or to task assignment. That's about as specific as you can get and that's pretty big.

Meister - Your objectives will become progressively defined, too, as the man or men who develop this evaluation start throwing questions at you as to what exactly you mean by what you say.

Lamb - You say, "available for the system engineer," for example, what do you mean? When? TDP time? Final design time? When? That's the type of question you're going to get asked.

Jenkins - Sometimes you are not going to know.

Meister - I would assume that the independent middle man would go to each of the people who represent a particular point of view and say, "I have a set of questions about what your objectives are, what you hope to achieve, how you hope to achieve it." I don't really think that you can take the published writings of the people in this room and extract that information, because as I said before, a great deal has been concealed by words and he is going to have to dig out many of the details of these methodologies which are not at all clearly evident today. It's probably never been made clear to anybody except the individual who developed the approach.

Tolcott - You might be asked to demonstrate if you can predict the performance of a system based on some minimum amount of data that you have available.

Swain - It's about time, but there should have to be a nuclear war to say whether we were right or wrong.

Meister - Then I would say that this would be a sort of a negative browny point.

Swain - If you never use these things, how in the world are you going to know?

Meister - Another aspect of comparative evaluation would be to look into such things as the capability to validate, that's one of the criteria. But to answer that question, you then have to ask yourself, what are the ways of validating? and What are the problems involved in validation? and how are these problems reflected in the initial approach? So there's a great deal involved in making such a comparative analysis. I know damn well, none of us has done this, not even for our own systems.

Swain - Well, I am sure that any one of these methods that you want to talk about can be validated somewhere, certainly on the CIC.

Meister - It's more than that Alan, it's the question of does the particular approach require particular kind of data which in turn requires particular kind of measurement, which in turn requires a particular kind of environment? and can you, in fact, get that? and what are the limitations imposed by an operational setting in terms of trying to get the validation? and this sort of thing. It's not simply saying I'll go out and see and validate or I'll go to a simulator and validate. I'm constantly being impressed by the complexity of the problems that we have to attack, and they keep getting covered over, very similar to the way a dog buries a bone.

Swain - I'm constantly being impressed by the overcomplexity with which we regard things because so often the answers that are needed in any human reliability effort are not nearly as tough as we make them out to be.

Meister - Your point is well taken, and part of this evaluation should be, as Jerry has indicated, just how precise must the data be that we have to have?

Swain - The engineers, they don't require this elaboration of prediction, they often predict with missing numbers or rank ordering models. No more than interval scales and forget the ratio scales.

Meister - Let's find out, because if that is the case it may be that we can cut years off this program. I don't really think it's going to be that easy, but it's a possibility which should be investigated.

Swain - No, I'm just saying that some applications ought to require that little amount of detail.

Mills - This is another point too and that is regarding precisely this question as to what kind of data are needed by the reliability engineer. This questionnaire that I distributed is not a project that I'm particularly associated with, but I did get into a discussion of it and the notion behind it is to first of all send the questionnaire out to Aerospace people and then follow it up with an on site investigation. My contention is that the on site investigation is needed. If it were conducted it would supersede anything pertaining to the questionnaire that's my personal opinion. But the on site investigation of trying to determine precisely, for various kinds of systems, various kinds of application problems, what kinds of data are useful?, or can be used. What kind of data would reliability engineers for example accept? This notion that Dr. Regulinski has mentioned, I don't think is that facetious. Quite frankly, it's got its merits, and an on site investigation, in other words across system applications is needed to ascertain the kinds of data systems engineers do want.

Swain - In other words if it were part of this survey or whatever you're talking about, I could say what kinds of information that the reliability engineer and the design engineer and the systems engineer, if you're talking about conceptual stages, want from the human factors man. What kind of data does he need?

Jenkins - I think you have a pretty good hold on that right now. Maybe not the kind but certainly the level just in the areas of pure human engineering data. Sometimes they want to know very precisely, and other times they want a very gross answer.

Tolcott - It's a question of what it is possible to give them when we talk about human behavior? We might not be able to answer all of the questions.

Mills - That's why the questions have to be determined, and they have to be determined specifically, not just in generalities, this doesn't get anybody anywhere. We have to determine precisely what these are, and try to put them into a taxonomy.

Swain - You have to have some kind of a weighting list with a proportion of each of these kinds of information like for example only one tenth of 1% of the time, you have to come up with a certain type of data and 90% of the time some other kind of data is okay.

Mills - Most of this is involved, for example, in the conceptual phase. If this occurred, we may be able to eliminate an awful lot of this drudgery work that we have been talking about. If we find out that the other case is just as true, and that is that these people really do want to know answers to questions at the microtask element level, if they do want answers to these questions, then we have really got to provide these data.

Swain - You've got to phrase your question in such a way that you will see if they can use it? You don't want to ask if they want it; certainly they want it.

Siegel - I question whether all of the questions we are asking really are human reliability? For example, I would think from what I hear, you are asking a number of questions. Question one is, "Will a system work at a given level in regard to a given set of objectives?, that's probably human reliability. Second question is, "If it doesn't, why doesn't it?" That's a little bit different from human reliability. Another question is, How to make it work? Now the how to make it work question is getting back into human engineering more, which is somewhat in my mind related to (and possibly calling on) but not human reliability per se. I think as you come down the tree, that these questions get further away from the concept of human reliability.

Tolcott - I think what has to be decided on is in what case will it work? I forget what we are talking about.

Siegel - The machine or the system.

Tolcott - Oh, not the prediction?

Siegel - No, I think from the point of view of the design engineer, we want to tell him will this kluge work?

Coburn - It goes back to the point that I made, I think that if you do take the position that you are not going to let catastrophic errors get into these systems, that is our business to get them out then, just how precisely do you have to know error frequencies and reliability data in order to be able to do this job of redesign to an adequate level to get it out? and that seems to me like a very good question.

Meister - Yes, it is.

Swain - I must admit, I don't have to use these numbers as often as I would like to use them.

Meister - I don't think you can second guess what the content of this comparative analysis would consist of. I am certain that it would be complex. There would be many questions to ask about these various approaches before you could come up with any sort of reasonable conclusion.

Lamb - I think the question that Art raised is a good question. What do we need for predicting reliability. We need the redesign that Dave talked about, we need the prediction for system error, we need all of these things. What techniques are going to go into the comparative analysis? This poor guy doing this thing has got a pension for the rest of his life if he has to get into every one of these areas.

Meister - No, no, no, no, no, it may be that he will use the Monte Carlo system, but I personally don't think that the task of doing such a comparative analysis is one which would be unduly prolonged although obvicusly it would have certain complex characteristics.

Harris - What are the criteria for such a comparative analysis?, do you get somebody else to get these up?

Meister - No, no, no, no, no.

Siegel - We set them up ourselves and then we evaluate our own criteria - you missed the whole point.

Swain - There's nothing wrong with that. What's wrong with that?

Meister - I am certain that we already agree on many criteria of an effective human reliability prediction technique. Bob's list corresponds largely to many things that I said and that Art said. Certainly the list that I put up on the wall yesterday needs to be amplified. I know it wasn't comprehensive. However, I don't really think that we have to worry about the kind of criteria that we would apply.

Harris - Let me just say one thing here, it seems to me that if indeed we have the kind of evaluated criteria in which we pretend to be objective, we pretend to be scientific or whatever, that surely we must behave that way. In applying these to whatever we ourselves are doing, not merely riding a particular horse because we like it. It's got to be related to real life criteria.

Meister - The whole point of the comparative analysis would be to be entirely objective.

Siegel - Well, isn't it true, Dave that when you do report writing you keep these criteria implicitly in mind and when you discuss your output you discuss it against the backdrop of criteria such as this?

Meister - No, I don't think so.

Siegel - I think most people do.

Meister - I don't think so - I think that if we did we would not be in the position we are in today.

Siegel - Then what does one put in the Discussion section of a report? What has one to say if he wants to write a discussion, unless he has some criteria, or he's talking about a method. I wouldn't know what to put in unless I did something like this.

Mills - You don't consolidate an entire area in the discussion section. You don't consolidate the magnitude of efforts that have already gone on in the field for ten years. Anyone can look back on the literature and read. I say to myself well, okay I've got a four year plan now what am I supposed to do.

Siegel - It seems to me that a literature review is something different than some guy playing God and judging everyone else's work.

Meister - No, nobody's suggesting playing God, I don't know anybody around here who would allow anybody else to play God. I would say that in the course of this comparative examination, it would involve a fairly reasonable degree of interaction with the people who are exposed to the particular points of view to make certain there was no playing God. That everything was in fact open and above board. I don't know that the comparative evaluation will answer all of the questions that need to be answered. I know that if we don't do something like that, ten years from now we will still be sitting around the same table asking the same questions, and this I might point out has happened time and time again in many other areas of human factors, engineering psychology, whatever you call it. If we are going to be quantitative and objective which is exactly what we pretend to be, then by all means let us be objective and scrutinize what it is that we are doing.

Swain - Of course you realize whenever we compare methods this means that we should be familiar with what all of these methods are. I don't know what all these methods are myself. I know Bob's, I know what you do, Dave, I know what I do, but I know very little about what Art does.

Meister - Well, isn't that very interesting that even with the limited number of people that we have here, who are supposedly experts, we don't really know the ins and outs of the various approaches. We've been so busy riding our own hobby horses. How can an independent agency such as that which Commander Connery and Jim Jenkins represent actually make any meaningful decisions, about which way to go, since they are outsiders looking in and we who are the insiders don't even know ourselves? All I'm saying is it's about time that we did, in fact, learn about the ins and outs of each other's methodology and present these to them for their decision. Then they can decide for themselves. Such a comparative evaluation would not necessarily specify a particular technique to follow, but at least it would point out to them, here are the alternatives as best we found out about them, you make the decision based on the objective criteria.

Aldrich - By the same token you can take into account Thad's position, I don't think you people know what he is saying. He's looking at the human in an entirely different way, treating him as a random variable. With all your methods and all your extensive development you treat the whole thing like a black box - the human all at once.

Meister - We are not arguing about whether we would or would not include Regulinski's point of view - we will.

Blanchard - Except that Regulinski has what he says is the most modern approach to reliability.

Mills - You said something in your paper about looking at other alternatives beside human reliability. I'd like to ask you what are your alternatives? Because even though you gave the paper, it seemed to me that you were still talking about relative frequencies which is exactly what Regulinski has said.

Harris - When I made the comment about reliability and probability it was based on what I view to be a long long way down the road before you can ever come up with anything that's going to be useful. That model that Regulinski presented (he called it modern and it's not modern) it's many years old. There have been many many experiments to plug in a human and try to treat him as just another element in a system.

Mills - Well, Alan Swain's branching network method does precisely that.

Harris - Well, of course it does, and the other question is, sure, you will end up with some model, some prediction model, some model of human reliability if you will. My only comment about the application of the hardware type probability reliability model that we have used, to human reliability is that it's difficult. It's been tried for a long time and it hasn't worked.

Tolcott - What do you mean it hasn't worked?

Harris - We don't have the estimates that are meaningful.

Meister - We are repeating ourselves, if I may say so. We are going back to prior discussion. That's not really the point of issue. I don't think you or anyone else can argue that the answers to the questions are explicit in what you are saying, simply, because there is no time at the moment for the comprehensive examination required to answer these questions.

Jenkins - Well, would you say that this comprehensive concurrent examination on models should be done in this year, or considering the total program, is it better to wait until '72' when our funds are greater?

Meister - Well, I'll tell you, when you originally asked me that question I said it should be done concurrently, but since you have problems in funding, it might be slipped a year or so. After listening to the conversation the latter part of this afternoon, I would say, "Man, you better get on it, fast, because the whole structure that we have been dealing with right now here seems to be becoming even more clouded than ever. Without such an evaluation you'll go down the road a bit further on and eventually you will find that what you are doing presently must be re-evaluated in terms of that comparative analysis.

Jenkins - That's true, but I think so far as the immediate objective and the program that we have to respond to these objectives, these are something which have been discussed and I think they are reasonable in the light of what we know now. Whether five years from now we will look back and say we shouldn't have done that, I don't know.

Meister - Then say I ask, Jim, why are you having this meeting in the first place?

Jenkins - No, I am saying that on the basis of what has been said about the programs that have been presented for your opinion and have asked you to take them apart, are they something that will meet the initial objectives, not the total objectives?

Meister - Well, they are not horrendous. It's not a bad program, it's just that it's a program that may not meet your objectives until you know what those objectives are.

Jenkins - Absolutely, we have to define additional routes - there's no doubt about that, and they have to be done concurrent with this review so that the review can make some sense. I don't want to mislead you, the purpose of this meeting was not to give some sort of blessing on the decision already made. That is not my intent, by any means.

Meister - I'm sorry.

Blanchard - I think that it would be extremely unfortunate to reach a point in the program where you are confronted with a need for a modeling technique not to have done this comparative analysis. In my opinion, before you spend a great number of dollars you have got to start applying something. The best available modeling technique for your purposes should be identified to fill short-term needs. This work then should be done first.

Jenkins - We'll do it.

Swain - I would like to make an observation that, with regard to starting a data store of that kind that I was suggesting, the work that you have going on by NEL and HFR is very amenable. They can be working on that while they are doing their problems. They should be, and instead of collecting data from all over the world as it were, we can at least be collecting the kind of data I think is required from the CIC.

Meister - The data will be good data, there's no question about it. Data are data.

Swain - They've got to be working on how to quantify all of these various performance shaping factors which has always been a problem for us. They can be trying the approach and other approaches to see what comes out.

Tolcott - I'm not sure the data will be good data, unless an analysis is done first to know what ought to be collected at the same time you're evaluating.

Meister - This is my feeling too. It's a question of how much analysis you do before you actually start your data collection efforts?

Siegel - I want to see a comparative analysis of data banks and data bank methods.

Meister - I don't like the way you say that, Art. The reason I don't like the way you say it is because there is a complete misconception here, that what we were talking about is a data bank. What we are really talking about is a predictive structure and if you start off with the assumption that we're just going to evaluate the data bank methods, I would say then to forget it, because that is not what you want. You are talking about a predictive structure.

Harris - There is some question about that, I tried to say this before, but obviously didn't say it well enough, I can't conceive of starting collecting data unless you've got a considerable amount of structure. You've got to know several things, no matter what kind of data. You've got to know what forms, the terms in which it's going to be stated. Certainly the use of this is for prediction of performance and for the very practical use by some to learn if the system can be used to make some judgement about the characteristics of a system.

There has to be a purpose and there has to be considerable analysis before you collect any data.

Meister - I'm glad you said that.

Harris - I thought I said it before.

Mills - I don't think the majority of this analysis should be cerebral either, I think that it should be real investigation, to try to make sure that the structure is relevant.

Harris - I couldn't agree more with you.

Meister - Well, where are we?

Jenkins - You told me that the objectives for the total program have to be pinpointed much more exactly than they are right now. That the concurrent evaluation of predicted structures must be started this year, and that's your recommendation, thus far. I'd like to draw out one thing, I would like to keep the group together as a working group, for several years through some sort of arrangement, if this is agreeable with you all, by having appropriate meetings and discussing the program.

Meister - It would certainly be a novel concept.

Jenkins - The specific arrangements would have to be worked out. The next thing I would like to ask is that since we've defined the overall objective of the human reliability program, do you see at the moment other specific objectives which should be attended to, within the next year to year and a half that we can start out? For example, within the Navy structure, we spend almost as much money on software systems as we do on hardware systems. We use that data quite a bit.

It's a software system. The areas of decision making at a higher command level or even decision making within the CIC, but basically, reliability of human decision making.

Meister - We have enough to go on right now. I wouldn't write off any more.

Swain - I wouldn't either. I'd be afraid of diluting the effort too much.

Meister - You'll get indigestion.

Mills - Not only that, but the kind of things you will find out in terms of directions from your evaluation lists, any software you develop has got to be highly specific. You've got to know before hand how you will develop the software. I don't see how you can possibly develop software like that now, except for data collection purposes only.

Meister - I think we all agree with him. We've got enough on our plates to keep us occupied for a considerable period of time. If we try to do too much in the end you will have nothing at all.

Jenkins - Then we will come up within the next couple of months, with an RFP and we'll go out to the people from companies represented here for a comparative data and prediction systems analysis if you want to participate. This satisfies the needs of the meeting at the moment. Again, I do express the appreciation of Ship Systems Command, the Office of Naval Research and the Naval Air Systems Command people. Thank you very much.

Dr. Altman was unable to attend due to a sudden emergency, but his paper is presented for the reader's information.

THE PROGRESSIVE INFERENCE APPROACH TO DEVELOPMENT OF DATA RESOURCES FOR PREDICTING HUMAN RELIABILITY

James W. Altman

Datagraphics, Incorporated

THE PROGRESSIVE INFERENCE APPROACH TO DEVELOPMENT OF DATA RESOURCES FOR PREDICTING HUMAN RELIABILITY[1]

INTRODUCTION

For present purposes, let a prediction of human reliability be any statement of probability that an individual, functional team, or aggregate of individuals will complete a defined unit of performance within established limits of time and/or quality. Let a data resource be any repository of information having potential to aid the prediction of human reliability.

The term progressive is meant to imply that both empirical and logical activities involved in the approach are (though contingent and iterative) inherently sequential. The term inference is used to imply that the approach requires an explicit statement of belief about causal factors -- beliefs subject to disproof. Certainly the approach should use "strong inference" insofar as this is possible in dealing with performance phenomena. According to Platt (1964).[2] "Strong inference consists of applying the following steps to every problem in science, formally and explicitly and regularly:

1. Devising alternative hypotheses;

2. Devising a crucial experiment (or several of them), with alternative possible outcomes, each of which will, as nearly as possible exclude one or more of the hypotheses;

[1] Prepared for the Navy Human Reliability Workshop, Washington, D.C., 22 and 23 July 1970.

[2] Platt, J.R., Strong inference. Science, 1964, 146(3642), 347-353.

3. Carrying out the experiment so as to get a clean result;

4. Recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain and so on (p. 347)."

The "stick point" in applying the method of strong inference to the development of human performance data resources is likely to be in getting "clean" results, because of difficulties in experimental control.

To the best of my knowledge, the general approach described herein has never before been suggested in the context of generating human performance data resources. Neither am I aware of its being inadvertently applied in the context except in the most fragmentary ways. Thus, what follows is entirely speculative.

I will first discuss some of the general characteristics of the approach which cut across its steps. Then, I will describe these steps.

GENERAL CHARACTERISTICS

The following general characteristics of this speculative approach to the development of human performance data resources are described more fully below:

1. Emphasis on functional relationships between rational and performance analyses.

2. Oriented toward anomalies, exclusions, and disproofs.

3. More suited to broad than to narrow applications.

4. Accepting of both field and laboratory data.

5. Amenability of partial data.

6. Responsive to multiple levels of performance.

7. Freedom from linearity assumptions.

8. Applies established techniques for dealing with complexity.

9. Permits optimum allocation of resources to resolution of depth versus breadth problems.

Functional Relationships

The essential objective of the approach is to establish functional relationships between task requirements and performance characteristics involved in meeting these requirements. The determination of task requirements involves two important aspects. One is to define the kinds of performance information required. The other is to provide information which will help to predict performance.

The assumption is that properly structured performance data can be accumulated and aggregated in some centralized resource and drawn upon selectively to help in estimating the performance that will be achieved in meeting task requirements. Generalized parameters can facilitate task description. We can look forward, though, to these rational analyses always being original and idiosyncratic to some degree.

Exclusions

The approach outlined here is aimed at the definition of relatively precise relationships between task requirements and performance. Anomalous data, disproof of expected relationships, and exclusion of untenable hypotheses will play an important part in the establishment of precise relationships. The emphasis is on exclusion of reasonable expectations that cannot be supported more than on short-run accomplishment of breadth of generalization.

Broad Application

Although there is no obvious reason why the approach suggested here is inherently incompatible with application to a single system or class of systems, there are two major reasons why the approach is more appropriate to broad than to relatively narrow applications. First, the approach has as its focal objective the establishment of functional relationships which will almost certainly generalize beyond the bounds of any particular system or class of systems. In this sense it would be wasteful to limit the data resource development to a single system or narrow class of systems.

Second, there is a problem of critical mass. The proposed approach will probably not be fruitfully mounted with small resources, small either in magnitude employed at a particular time or in being constrained to delivery of results in too short a time. The payoff issue is intensified by the fact that the proposed approach can be expected to yield disproportionately narrow and unreliable results in its early stages as compared with an increasing richness of return beyond a critical point of investment.

Field and Laboratory Data

Although laboratory situations permit a kind of control that can make for much more efficient testing of a hypotheses than can field situations, the suggested approach will accept either type of data. Eventually, of course, validation of projections made from human performance data resources must be validated against actual performance in operating systems.

Partial Data

Efficiencies of data analysis and interpretation will result, of course, from given individuals performing many tasks, from randomization of assignments to conditions, and having complete data across all combinations and permutations of relevant conditions. But the suggested approach is not especially sensitive to such strictness. Rather, it emphasizes utilization of such data as can feasibly be obtained. Such acceptance of obtainable data comes at a cost. Either the creative burden on derivation of alternative testable hypotheses or the basis of predictions must be less precisely defined than desirable.

Performance Levels

The suggested approach is sensitive to the levels at which task requirements are specified, performance estimates made, and data stored. However, it does not begin with predilection for any particular levels. Rather, it has an affinity for clear functional relationships at whatever levels such clarity may be possible.

Linearity

Multivariate procedures seem to have special promise for the analysis of data to be used in human performance information resources. Any of the more facile of these procedures, however, makes stringent assumptions of linearity. Although the proposed approach can, where appropriate, make full use of multivariate and other procedures involving linearity assumptions, it is in no sense limited to such procedures.

Complexity

The suggested approach is not only free of restrictive assumptions of linearity, it is quite flexible with respect to the use of any analytic tools. The mathematical-statistical procedures it will accept are essentially unlimited. Full use can be made of the computer as a processing and simulation aid, but no particular kind of use is prescribed.

Depth Versus Breadth

The suggested approach involves no *a priori* commitment to breadth of performance covered versus depth of analysis for any particular domain of performance. Rather, synthesis of hypotheses and analysis of data must take whatever course may be required to establish sufficient functional relationships to be useful in supporting estimates of human performance across a domain of interest.

STEPS

Major steps involved in the speculative approach suggested here are as follows:

1. Organize and analyse background data.
2. Establish a strategy for selecting tasks.
3. Analyze a first task.
4. Continue to analyze tasks incrementally.
5. Validate.

Background Data

As a preliminary step in the development of a human performance data resource, it will be well to review and organize existing data and beliefs concerning the relevant region of performance. This can include existing data resources, field data, laboratory data, and accepted functional relationships. This, of course, does not have to be a one-shot proposition since one can re-interpret, dig in greater depth, and explore data newly found to be relevant as the subsequent steps are carried out.

Task Selection

One has to be careful not to end up in a "chicken-and-egg" situation in trying to establish a basis for subsequent selection of tasks. If there were a generally accepted taxonomic framework for tasks, then the problem would be relatively straightforward. But no comprehensive and generally accepted task taxonomy exists, nor do I expect to see one any time in the near future. Consequently, one must be rather arbitrary and tentative about choosing the dimensions and categories that will be used to establish a basis for selecting tasks for study.

Fortunately, inherent to the approach suggested here is an evolutionary clarification of functional similarities and differences among tasks. That is, the approach should continually yield additional insight into the factors that make for similarities and differences among tasks. This insight should help one to continue throughout application of the approach to improve his selection of tasks for analysis.

First Task

The pivotal assumption for the entire approach is that ability to predict performance accurately for one task will facilitate ability to make accurate predictions for other tasks. Consequently, considerable emphasis should be given to definitive study of the initial tasks. This includes, of course, both rational-descriptive analysis to establish task requirements rigorously. It also involves empirical study of performance.

The initial task(s) should study performance by a variety of personnel across a range of situations. Emphasis should be given to definition of performance variables, including description of all of the different kinds of errors possible. Empirical performance distributions should be predicted, "explained" after the fact, and deviations from predicted values eliminated through re-analysis and replications. Initial "sufficient" models for predicting performance should be refined and made as parsimonious as possible.

Incremental Tasks

Once an effective basis has been established for predicting performance on a single task or initial cluster of tasks, a second task or cluster of similar tasks should be similarly analyzed. This process should be continued until the entire region of interest has been covered. Predictions should be sufficiently fine-grained to suffice for the purposes of the human reliability technique to be supported by the data resource.

An important difference between initial and subsequent tasks to be analyzed for purposes of developing a data resource is that the subsequent analyses have the benefit of information and insight gained from all of the previous analyses. This is important in terms of identifying general performance predictions which can be considered in the subsequent analyses. Also, the more tasks with established performance characteristics surrounding a new task, the narrower the region of uncertainty initially surrounding performance characteristics of the new task.

Validate

In a sense, validation is an inherent part of each task analysis as defined here. However, a series of validations using predictions from the data resource are called for. These validations should involve careful performance observations under actual system operation.