SCR-610

Conf-621106--1

Sandia Corporation

REPRINT

HUMAN ERROR QUANTIFICATION

by

A D Swain, J W Altman,

and

L W Rook, Jr

APRIL 1963

DISCLAIMER

Issued by
Sandia Corporation,
a prime contractor to the
United States Atomic Energy Commission

SCR-610

CONF-138-1

SANDIA CORPORATION REPRINT

HUMAN ERROR QUANTIFICATION
A Symposium

A. D. Swain, Sandia Corporation (Chairman)
J. W. Altman, American Institute for Research
L. W. Rook, Jr., Sandia Corporation

Held at 6th Annual Meeting of the Human Factors Society, New York, N. Y., November 28, 29, 30, 1962.

April 1963

CONTENTS

PREFACE

The Human Error Quantification Symposium was held at the annual meeting of the Human Factors Society in New York City November 28-30, 1962. The symposium dealt with an heuristic approach to quantification of human error. It presented methods of predicting error rates and estimating man-caused degradation which could be used in the first stages of man-machine development. The substance of the talks and discussion is published here to satisfy reprint requests made to symposium participants and to further the collection of human reliability data which was proposed during the symposium.

Specifically, the establishment of a human reliability data bank was suggested and this idea has since received considerable support. Until a formal data collection program is organized and facilities for storing the data are set up, Sandia Corporation has offered to serve as an information clearing house. Individuals or organizations interested in participating may send human error data to

Human Factors Staff,
Division 1443,
Reliability Department,
Sandia Corporation
Albuquerque, New Mexico.

Contributors will be placed on a mailing list to receive copies of all such materials received. Suggestions for implementing the data bank program will be welcome and will be available to any organization or institution interested in sponsoring the program.

To facilitate distribution, contributions should be typed with black ribbon for reproduction. Since electronic computers may be used in categorizing errors and variables, contributers might consider this in preparing data for transmittal.

Among the human factors techniques being developed, two presentations at the April 1962 national symposium on aerospace systems reliability were especially relevant to the prediction of human error rates.

These papers were given by David Meister and John Brady in the human factors session on personnel subsystem reliability. Copies of the complete proceedings are available from the

Institute of Aerospace Sciences
2 E. 645th Street
New York 21, New York.

AN APPROACH TO QUANTIFYING HUMAN PERFORMANCE

James W. Altman*

Quantitative estimates of human performance can help the systems engineer balance human factors against other aspects of system design. Although methods for quantifying human factors evaluation certainly have not been perfected, starts toward such quantification have been made. This paper discusses some quantifying techniques used at the American Institute for Research.

Operability Index

Objectives

The American Institute for Research (A.I.R.) developed an index of electronic equipment operability for the Army Electronic Proving Ground. Three objectives were set for the operability index:

1. The operability index would be task-oriented. That is, the equipment design would be evaluated in terms of its compatibility with specific human operations.

2. Scores from the operability index would be directly meaningful in terms of time and error of performance (speed and reliability).

3. Data on speed and reliability of operator performance would be obtained, insofar as possible, from available experimental and field data.

Behavior Analysis

Since tasks, as we usually define them, are too complex to serve as the basic unit of performance prediction, the "behavior" was selected as the basic unit of evaluation for the operability index. A behavior here is a specific step or action in a given task. Each behavior was broken down into three aspects: (1) inputs or stimuli to the human senses, (2) mediating processes, and (3) outputs or responses (i.e., motor activities of the human).

---

*James Altman is director of the engineering psychology program at the American Institute for Research, Pittsburgh, Pennsylvania.

We identified those types of equipment components, behavior components, and equipment-behavior components likely to affect each aspect of behavior. The components, listed in Table I, represent a compromise between what are desirable as psychologically meaningful categories and categories for which data are available. Note, for example, the paucity of categories under "Mediating Process." There is little usable information about the quantitative effects of mediating process variables on performance. To go to an elaborate scheme for categorizing mediating processes at this time would be like raking straw with a fine tooth comb.

TABLE I

List of Input, Mediating Process, and Output Components

Inputs	Mediating Process	Outputs
Circular scales	Identification/Recognition	Cable connections
Counters	Manipulation	Cranks
Labeling		Disconnecting
Lights		Joysticks
Linear scales		Knobs
Nonspeech		Levers
Scopes		Object positioning
Semicircular scales		Pushbuttons
Speech		Rotary selectors
		Speech
		Toggle switches
		Writing

After breaking down behavioral aspects into components, we broke down each component into parameters. Sample parameters for the input component "lights" are listed in Table II.

TABLE II

Parameters Affecting Performance on Lights

Component	Parameters
Lights	Size Brightness Type/Function Number Presentation

Finally, we broke each parameter into dimensions which are either discrete categories or intervals on a continuum. Table III shows sample dimensions.

TABLE III

Sample Dimensions

Component: Circular Scales

Parameter: Scale Diameter (Interval Dimensions)(in.)	Parameter: Scale Style (Discrete Dimensions)
1 to 1.5	
1.6 to 1.8	Moving pointer
1.9 to 2.8	Moving scale

Let us here review the behavioral levels relevant to the "Operability Index." Each level is identified and exemplified below.

Behavioral Levels

Mission: operate fuse jammer

Phase: prepare for operation

Task: activate amplifier

Behavior (or step): throw S11 to ON position

Aspects of Behavior: (inputs, mediating processes, outputs)

Components: (specific categories of an aspect) toggle switch as a component of output

Parameters: (relevant characteristics of components) angle of throw from position, as a parameter of the component toggle switch

Dimensions: (specific values or characteristics of parameters) 40°

Human Performance Data

We searched several thousand research reports for data we could use to estimate the effect of design dimensions on performance. We found usable data in 164 reports. Experimental literature was not consistent with respect to identification, description, or treatment of independent, control, or dependent variables. Although human engineering studies yielded the most relevant information, it was difficult to establish performance standards from these studies since they were aimed at answering specific questions and were not related to any theoretical framework.

We summarized experimental results relevant to each component in tables so we could identify discrepancies and areas of missing data. Most discrepancies between the results of different experiments varying the same component came from gross differences in the mediating processes required of the subjects. By analyzing the nature of these differences, it was possible to reconcile the discrepancies and arrive at comparable error rates for the component in question. To find the missing data, we searched the literature; where this search was fruitless, we interpolated or extrapolated from related studies. Finally, where no experimental data could be obtained, judgment was used to establish estimated effects--sometimes supplemented by "quick-and-dirty" empirical studies using a few members of the A.I.R. staff.

To get standard time estimates, we identified minimum time (in seconds) required for an operation when all dimensions within a given component are optimum. This we call base time. Then we established average time increases for the operation of a component with other (nonoptimum) dimensions. Thus to get the total time for an operation, we add the time increase to the base time.

Time measures taken directly from experimental studies seemed to give reasonable estimates of performance time under actual operational conditions. Error rates in experimental studies, however, seemed to be grossly high for a direct estimate of operational error rates. Apparently these experiments had been designed to get high error rates to obtain more confidence, in statistical comparisons, without requiring an excessive number of trials.

Although no field studies had been conducted which related errors to tasks, behaviors, aspects, components, parameters, or dimensions, earlier field studies suggested that average probability of one or more human errors in a man-machine military mission was about 0.13. Missions most relevant to the operability index averaged 50 behaviors per operator. Therefore, the average probability of operator error on a behavior was about 0.0026 as contrasted with an average probability of error on a laboratory behavior of about 0.31935. Thus, a scale factor of 0.008145 (0.0026 divided by 0.31935) was used for each laboratory estimate of error to obtain an estimate of operational error. Reliability (the probability of correct performance of a behavior) was taken as unity minus the probability of an error.

Experimental studies did not attribute operating time on errors to individual aspects of behavior as we have defined them, but for a total behavior. However, since experimenters usually vary only one aspect at a time in these studies, we assumed we could attribute to the aspect being varied the experimental time increases and error rates for the entire behavior.

We reviewed our time and reliability estimates for consistency and reasonableness and then put the figure for each component in a central data store. A sample data store card for the component joystick is shown in Figure 1.

Joystick

(May move in many planes)

BASE TIME = 1.93 Seconds

Time Added (Seconds)	Reliability	
		1. Stick length
1.50	.9963	a. 6-9 inch
0	.9967	b. 12-18 inch
1.50	.9963	c. 21-27 inch
		2. Extent of stick movement (Extent of movement from one extreme to the other in a single plane)
0	.9981	a. 5-20°
.20	.9975	b. 30-40°
.50	.9960	c. 40-60°
		3. Control resistance
0	.9999	a. 5-10 lb
.50	.9992	b. 10-30 lb
		4. Support of operating member
0	.9990	a. Present
1.00	.9950	b. Absent
		5. Time delay (time lap between movement of control and movement of display)
0	.9967	a. 0.3 sec
.50	.9963	b. 0.6-1.5 sec
3.00	.9957	c. 3.0 sec

Figure 1. Sample data store card.

Method of Evaluation

The evaluation process is diagrammed in Figure 2. The major steps in evaluation are to

1. organize task description and equipment design information for ease of cross comparison,

2. break each operation down to its dimensions,

3. obtain estimates of speed and reliability for each dimension from the data store, and

4. use these estimates to derive scores for units of behavior such as steps, tasks, phases and entire missions.

In general, time scores are the sum of individual time estimates. Reliability scores are the product of individual reliability estimates. Reliability scores are thus estimates of the probability that serious operator error will not occur within the span of performance encompassed by the score.

Note that the index can indicate problem areas because it provides scores for almost any set of design factors or tasks. Also, scores can be obtained for suggested redesign configurations as well as for the actual design at any given time.

Tryout

Three evaluators applied the operability index to the operation and maintenance of several items of electronic equipment. Percent agreement in selecting behavioral dimensions for each of these items was determined by dividing the total number of dimensions used by all three evaluators into the total number of times each one chose the same dimension. Results are summarized in Table IV.

TABLE IV

Agreement Among Three Evaluators in Choosing the Same Dimensions When Applying the Index

Equipment	Total Entries	Total Number of Agreements	Percentage Agreement (%)
AN/APS-94	427	379	89
M-33	1023	894	87
AN/MLQ-8	425	271	64
All	1875	1544	82

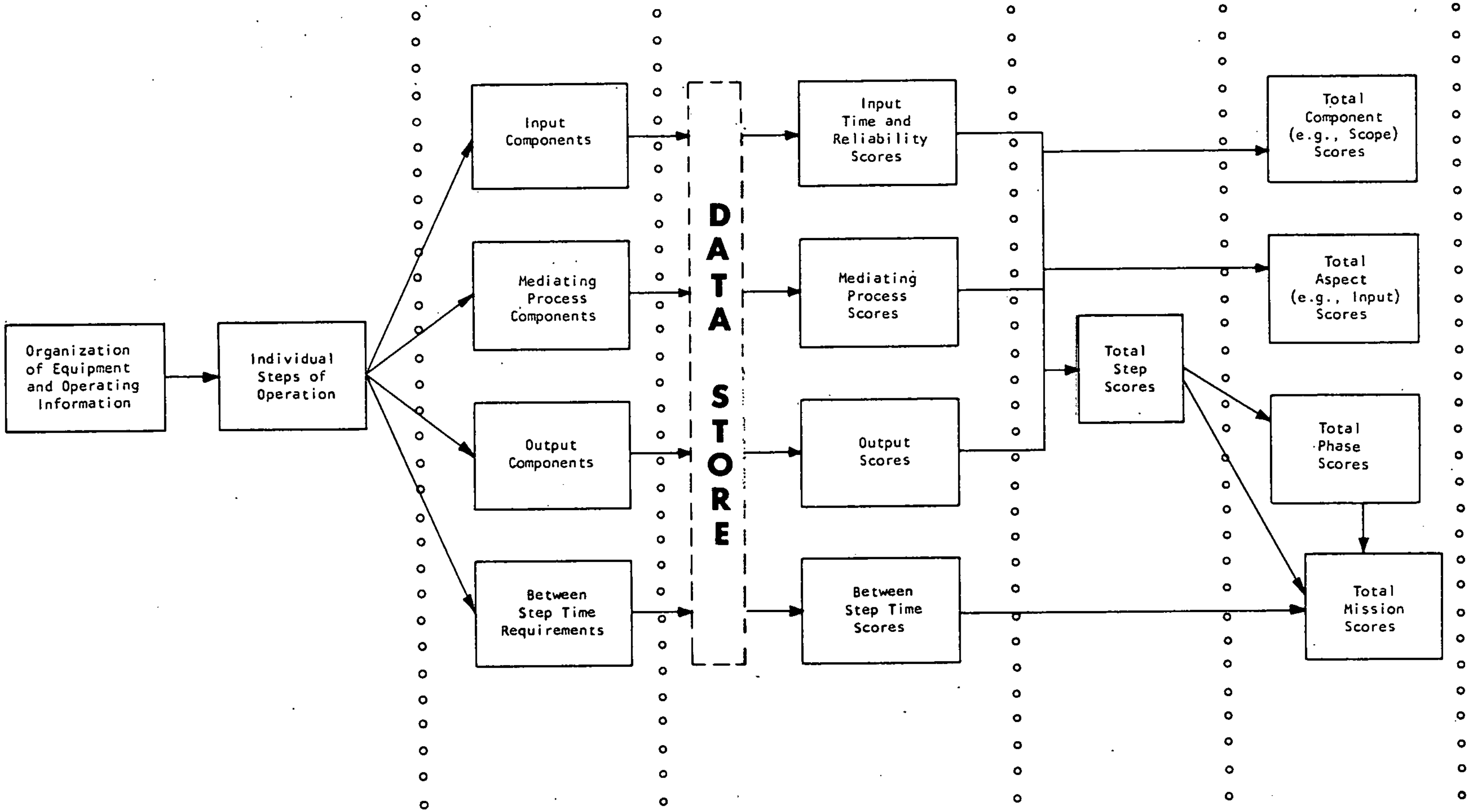

Figure 2. Steps in the basic evaluation process.

Although the agreement of these three evaluators (human factors specialists and engineers) indicates that the index does describe behavioral dimensions with reasonable accuracy, we have not checked our time and reliability estimates in the field. In a preliminary check of index validity, we asked three men experienced in using this electronic equipment to rank order four items according to the complexity of operator tasks. Then we asked them to apply the operability index to these four items to get, first, a total operator performance time and, second, a total operator performance reliability score for each item. The results (Table V) show agreement between their judgments of operation complexity and the time and reliability estimates derived from our operability index.

TABLE V

Agreement Between Index Score and Operator-Judged Complexity

Equipment--Ranked in Order of Judged Complexity of Operator Tasks	Mean Index Scores	
	Operator Performance Time (Seconds)	Reliability of Operator Performance
AN/GRC-50 (Judged most complex)	1098.51	.46
M-33	288.48	.81
AN/MLQ-8	181.04	.96
AN/APS-94	104.14	.96

Conclusion

The Operability Index has successfully estimated human time and accuracy in the operation of electronic equipment. Although the practicality of the index for predicting operational behavior still has to be tested, the techniques used represent what we hope is a long step toward quantification of human performance.

List of References

An Index of Maintainability. Instruction Manual. Pittsburgh: American Institute for Research, 1959.

Evaluation Booklet. An Index of Maintainability. Pittsburgh: American Institute for Research, 1959.

Munger, M. R., & Willis, M. P. Development of an Index of Electronic Maintainability. Pittsburgh: American Institute for Research, 1959.

Munger, M. R., Willis, M. P., & Altman, J. W. Quantification of Expert Judgment in Maintenance Design Decisions. Reprint from IRE Transactions of the Professional Group on Human Factors in Electronics, September 1961, HFE-2, (2).

Munger, Sara J., Smith, R. W., & Payne, D. An Index of Electronic Equipment Operability: Data Store. Pittsburgh: American Institute for Research, 31 January 1962.

Munger, Sara J. An Index of Electronic Equipment Operability: Evaluation Booklet. Pittsburgh: American Institute for Research, 31 January 1962.

Payne, D., & Altman, J. W. An Index of Electronic Equipment Operability: Report of Development. Pittsburgh: American Institute for Research, 31 January 1962.

Payne, D., Altman, J. W., & Smith, R. W. An Index of Electronic Equipment Operability: Instruction Manual. Pittsburgh: American Institute for Research, 31 January 1962.

Smith, R. W., & Payne, D. An Index of Electronic Equipment Operability: Sample Equipment Evaluations. Pittsburgh: American Institute for Research, January 1962.

A METHOD FOR EVALUATING THE HUMAN ERROR CONTRIBUTION* TO SYSTEM DEGRADATION

L. W. Rook, Jr.**

Getting adequate and reliable raw data is a major problem in evaluating human error quantitatively. Many reports of equipment failure and unsatisfactory operation are suspect because human error tends to be either attributed to other causes or to be let go unreported. Industrial production error data is hard to get because it inevitably affects production cost and schedules.

Evaluating human error in the field of nuclear weapons design and production is less troublesome than in most industrial situations because in this field reliability is emphasized while field testing is deemphasized or not done at all. Quality is important, but it has to be determined by unusual methods.

We set up acceptance specifications which require that highly detailed defect data be recorded. Step-by-step records are kept as components progress through manufacturing, assembly, and inspection stages. All records are keyed to specific units so that each part may be traced through all work and rework stages. Finally, because specifications call out assembly operations in great detail, it is possible to start from the description of a defect in one stage of component assembly and reconstruct any human error which produced the defect.

I have spoken of human error in the production and not in the use of nuclear weapons for two reasons. First, a large, if not major part of the human contribution to nuclear weapon failure probability is made

---

*A more detailed description of this method can be found in <u>Reduction of Human Error In Industrial Production</u>, L. W. Rook, Jr., Sandia Corporation SCTM 93-62(14), Albuquerque, New Mexico.

**L. W. Rook, an assistant professor of psychology at the University of New Mexico, is a psychologist in the Advanced Systems Study organization at Sandia Corporation.

in the production process. The no-field-test or wooden-bomb design philosophy operates here. And then the detail with which defect data are recorded in the production of nuclear weapons makes this process easier to deal with than field use of weapons. However, we hope that the system of treating human error which we are devising will be applicable both to production and field errors.

We collected descriptions of 18,000 human-caused product defects, largely from electrical and final assembly processes. Because we wanted to know both what caused the errors and how frequently they occurred, we analyzed the data qualitatively and quantatively. Our efforts, of course, were based on the assumption that human behavior is not so variable that it cannot be predicted. Predictability of any phenomena is a matter of degree; if predictions can be made which are sufficiently accurate to be useful in practice, these phenomena we hold to be predictable.

For these human-caused defects we tried to establish categories of error which would have both psychological relevance and operational utility. First we used the classifications described by Altman, that is, input, mediating, and output behaviors.* But when the data were classified according to this system, classes proved too general; the act of classifying did not tell one enough about the behavior classified. So we superimposed a second system of three categories which cut across the classifications of the first system. The second system defined the intent of the human who made the error. If the act in which the error occurred was performed intentionally, the error was Type A; if the act was unintentional, the error was Type B; and Type C covered errors which occurred because of omitted acts. Although Type C is logically a subclass of B, it seemed to derive from different causes and therefore was given a separate designation.

The system of nine error categories (our three types for each of Altman's three classifications) proved a big help in diagnosing errors; by inspecting the physical results of a human error we could, usually, place the error in one or perhaps two of the categories. Then, if we referred to the defining characteristics of the category, most often it was obvious what ought to be done to reduce the probability of making that error. Also, this classification system pointed out which kinds of error might yield to conventional methods of improving "workmanship", such as training and motivating workers. In many cases these

*<u>An Index of Electronics Equipment Operability</u>, AIR, 1962.

methods seemed likely to produce little if any reduction in error rates because the factors known to modify human behavior within some of the categories usually are not influenced by training and motivation.

This system for classifying human error allowed numerous individual errors to be treated in the same way; fixes could be devised for classes of error rather than for particular instances of error. But after the system of classifying errors was set up, we had to devise some quantitative measure of the effects of error and of the improvement which might be expected from proposed fixes.

In devising a mathematical model for measuring error effects quantitatively, we followed methods used in reliability analysis. Probabilistic models using measured error rates predicted the probability of failure from specific errors.

The form which had the widest applicability was

$$Q_i = 1 - (1 - F_i P_i)^{n_i} \qquad (1)$$

where P_i is the probability that an operation will be performed so as to produce error i, F_i is the probability that failure will result if error i occurs, n_i is the number of similar operations in which error i can occur, and Q_i is the probability of failure as a result of error i.

If two errors must combine to produce failure

$$P_i = P_1 \cdot P_2 \qquad (2)$$

where P_1 and P_2 are the probabilities of the two errors, respectively.

In applying this model one may choose not to use the error categories discussed. We found that while the categories were sufficiently specific for suggesting corrective action, they were too general for incorporation into a mathematical model. The kinds of error description most useful in quantitative treatment are the same kinds which are usually described in reports of product inspection such as "omitted solder joint," "wrong valued component" or "connector not tightened." Errors recur frequently; the variety of errors in any process, however, is usually small. In the assembly of a complex radar fuzing unit, for example, fewer than thirty kinds of human error showed up, and only eight of these contributed materially to failure.

With this model if the values of Q_i for each kind of error are rank-ordered from largest to smallest, the list of errors obtained orders the errors in terms of their contribution to failure probability.

The total probability of failure is given by

$$Q_T = 1 - \left[\prod_{k=1}^{n} (1 - Q_k) \right] \qquad (3)$$

where Q_T is the probability that one or more failure conditions will result from one or more human errors in at least one of n classes or errors.

To apply this model one needs estimates of P_i for various kinds of error and F_i for various kinds of hardware. The determination of P_i for the most frequently observed errors in the assembly of radar-like electronic equipment is a start in this direction. We are also analyzing the variables which influence error-rates and the effects of errors on product (system) performance. Finally, we hope to be able to accurately estimate P_i and F_i for most of the errors in nuclear weapons production.

To date we have determined rates for about 30 of the most frequent production errors in electrical assembly. Over two million opportunities for error were considered. Using this data we have both confirmed and predicted errors successfully.

Certain kinds of errors tend to recur in many unrelated tasks. For example, in all repetitive tasks examined, from soldering printed circuit boards to cementing components together in final assemblies, routine operations were omitted about five times out of 100,000, regardless of environment. But there was also wide variability between operators.

I will speculate that we are dealing with two classes of phenomena. In one case behavior-producing stimuli are internally generated; in the other case environmental stimuli are dominant. I suspect that error rates in the former are primarily functions of internal variables while those in the latter are functions of both internal and environmental variables. How to manipulate the appropriate variables has yet to be thoroughly investigated, although Altman and his colleagues have made a good start at understanding the external ones.

You can see that we have been using the production plant as a laboratory to develop and sharpen our tools and those developed by ATR. We also have begun to apply the methodology to field problems. Swain will say a few words about this.

SUMMARY

A. D. Swain*

The methods discussed by Altman and Rook are empirical methods. However, although they involve a large collection of human-error rates, the psychologist's thumb is still required on the scales in order to properly weigh the behavioral elements involved. No doubt purists among us will be appalled. We who quantify human error predictions do generalize from the basic empirical data. And in some applications, especially those involving behavior under highly stressful conditions, this generalization is stretched a long way.

But what is the alternative? Should we answer the demands for human reliability estimates by saying that further research is needed? We prefer to break down the human aspects of a man-machine system into its behavioral elements, assign the best reliability estimates we can to these elements, and then combine these estimates into an estimate of system degradation due to the human element. To do this we use variations of methods used for component and systems reliability estimation and operations research. Where possible our estimates of human reliability and man-caused system degradation should be checked against error rate data from man-machine system operation. The original error rate estimates or the method or both should be corrected as necessary. Thus, the prediction method is iterative.

---

*A. D. Swain is a human factors specialist in the Sandia Corporation Reliability Department.

At Sandia Corporation, we have given this approach the acronym THERP, or Technique for Human Error Rate Prediction. (As human factors specialists you will recognize that, to succeed, a scientific endeavor must have a pronounceable short title.)

In two recent system studies at Sandia Corporation, we showed quantitatively that proposed designs and procedures would permit an unacceptable human error rate and in both cases the design engineers reacted favorably to the use of human reliability figures. Relatively ambiguous verbal judgements such as "highly error-likely," "moderately error-likely," "sort of error-likely," and so on, are not nearly so likely to influence an engineer's design as are quantitative statements. Yet it is not necessary (or desirable) to assume more preciseness in human reliability estimates than the data warrant. Sandia engineers accept the fact that our predictions of human error rates are not perfect just as they recognize that reliability estimates made for hardware are not perfect. Actually the sources of error in our estimates of human reliability often seem no greater than the sources of error in estimates of hardware reliability.

The attempt to be quantitative forces one to avoid vagueness and to be concrete in thinking about human performance. After we have analyzed man-machine system reliability, often we have been forced to decide, to our surprise, that although certain design features deviate from accepted human engineering practices, their effect upon system reliability would not be important. Thus, our quantitative approach has forced us to pay more than mere lip service to systems considerations. I think all of us in the human factors field can benefit from this kind of quantitative approach.

But though we will provide human reliability estimates now, we insist that further research is needed. We need a larger and more accurate data store of human reliability figures for human behavioral elements, especially from field conditions. Rook collected and analyzed assembly, inspection, and testing, data for human error at one large industrial plant. He found actual error rates in certain assembly tasks to be within a factor of 2, 3, or 4 of the error rates estimated from the AIR data store described by Altman. When one is dealing with error rates varying from 10^{-3} to 10^{-5}, this degree of precision is ordinarily quite adequate.

A similar data collection and analysis effort is needed for military systems. Although a considerable amount of human behavior data is routinely collected by various military agencies, especially in conducting system proficiency tests, most of this data is in the form of error frequencies only. That is, the additional information necessary to determine error rates is not often collected. Part of the problem has been a gross over-estimation by human factors specialists, as well as others, of the preciseness of human reliability predictions which is required to be of practical use. Often all that is needed are estimates which are correct to within an order of magnitude. But even this degree of precision requires that the error data being collected include the frequency of opportunity for each error, that is, the denominator necessary for calculation of an error rate.

In 1957 the Advisory Panel on Psychology and the Social Sciences of the Office of the Director of Defense Research and Engineering began some human behavior research. Subsequently, this study was taken over under contract by the Smithsonian Institute. Reporting on this study in a recent issue of the American Psychologist,* Charles W. Bray of the Smithsonian proposed that the Department of Defense set up an intellectual skills laboratory to establish standards of human performance by measuring the variables in military tasks.

We hope that the Department of Defense will make this long-range program possible. But it takes a lot to get such an effort rolling. To help, the Human Factors Society could sponsor a group to design a system to collect, analyze, and categorize human error rate data. If such a group existed, all of us who collect error rate data could contribute to a data store and, ultimately, to a systematic taxonomy of human tasks. With the nucleus of a workable data store system established, we should find it easier to convince the DOD (or some other well-funded organization) to take it over and make it grow.

Without a massive effort in this area, quantification of human error predictions will continue to be piecemeal and years behind hardware system reliability analysis. Studies like the Stanford Research Institute survey** for the Air Force Systems Command will continue to rediscover, to the surprise of system planners, that from 20 to 50 percent of system malfunctions are due to the human element. And system planners and designers will continue to attempt to design man out, often at a serious cost to system reliability and effectiveness, because they feel we cannot determine man-machine reliability as accurately as they can determine hardware reliability.

---

*Bray, C. W., "Toward a Technology of Human Behavior for Defense Use," American Psychologist, 1962, 17, 527-541.

**Shapero, A., 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, Feb. 1960.