IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-12, NO. 3, MAY/JUNE 1982

# Correspondence

## Analysis and Classification of Human Errors in Troubleshooting Live Aircraft Power Plants

WILLIAM B. JOHNSON AND WILLIAM B. ROUSE,
SENIOR MEMBER, IEEE

*Abstract*—**Two experimental studies were employed to develop and evaluate a scheme for classifying human errors in troubleshooting tasks. Both experiments involved trainees in a Federal Aviation Administration (FAA) certificate program in power plant maintenance. Each trainee received one of three training methods, two of which were computer-based, and then transferred to troubleshooting live aircraft power plants. The transfer data from the first experiment ($N = 36$) were used to develop the error analysis and classification procedures. In addition results of this analysis led to modifications of the training methods. The transfer data from the second experiment ($N = 22$) were employed to evaluate the effects of the changes in training methods. Results indicated a decrease in errors due to these changes. The development of the methodology is discussed and the resulting behavioral interpretations are presented.**

### INTRODUCTION

When evaluating instructional methods, the mistakes made by trainees are often of more interest than those aspects of performance that are correct. This is due to the fact that mistakes often provide considerable insight into fundamental human limitations while correct behavior is, to a great extent, solely a reflection of task constraints. Insights such as these can be quite valuable from both a theoretical perspective and a practical point of view regarding the design of instructional methods.

Most researchers who have studied human errors have reached the conclusion that all errors are not equally important or interesting. As a result, a variety of error classification schemes have been proposed. Some of these schemes are quite general. Norman's [1] classification of "slips of the mind" is motivated by a desire to describe the underlying behavioral mechanisms responsible for producing errors. Halpin *et al.* [2] discuss the distinction and interaction between structural and attitudinal factors and their effect on information processing errors. Rasmussen [3] suggests that human errors must be evaluated with respect to an appropriate mode of behavior rather than a particular desirable action sequence.

More task-specific approaches to classifying human error have been suggested by Lees [4], Ruffell Smith [5], Rasmussen *et al.* [6], and van Eekhout and Rouse [7]. While at first glance the general and task-specific schemes may seem somewhat incompatible, it actually is more typical for the task-specific schemes to be very special cases of the general schemes. From this perspective, the more useful of the general schemes are those that can easily be specialized so as to be employed for particular tasks.

In this correspondence a task-specific error classification system is presented for troubleshooting or fault diagnosis tasks. This scheme is a modification of that used by van Eekhout and Rouse [7], which itself was a modification of a scheme developed by Rasmussen and his colleagues [3], [6]. The scheme proposed here is a direct outgrowth of an attempt to explain the effects of alternative methods for training aircraft power plant mechanics. The next section will focus on this data base.

### EXPERIMENTS

The human error data to be discussed in this correspondence were compiled from the results of two experimental studies of troubleshooting. The primary objective of the two experiments was evaluation of the learning transfer from computer simulations to live aircraft power plant troubleshooting. The first experiment was conducted in the spring of 1980 and is discussed in detail elsewhere [8], [9]. The second, which took place in the spring of 1981, was modified with respect to the number and variety of computer simulations and, to a slight extent, in the manner in which the live troubleshooting was conducted.

#### *Subjects*

The participants for both experiments were advanced aviation maintenance trainees from the University of Illinois Institute of Aviation. The trainees were enrolled in the twelfth, and final, course leading to the Federal Aviation Administration (FAA) certification as an aircraft power plant mechanic. The trainees had completed, or were concurrently enrolled in, all of the power plant systems courses. They were ready to integrate all prerequisite courses into one course centered on engine overhaul, testing, and fault diagnosis. It was during the start of this course that the experiment was conducted.

#### *Training Methods*

The training methods consisted of either computer simulations or a control treatment of instructional television. The computer simulations, which have been discussed extensively in the literature, consisted of a context-free simulation, called TASK [10], [11], and/or a context-specific simulation, called FAULT [12]. The context-free simulations varied in problem size and levels of computer aiding. The context-specific simulation used problems on one automobile and two aircraft engines. The level of feedback for the context-specific simulation varied between the experiments and will be discussed later.

In both experiments the control group was trained with instructional television termed VIDEO. The programs included the type of information normally presented in traditional troubleshooting training. A combination of fault diagnosis situations including actual demonstrations with live engines were presented. The VIDEO programs were supplemented with troubleshooting reading assignments and computerized quizzes.

VIDEO was basically a surrogate for highly context-specific traditional instruction. In order to more fully determine the value of the type of context-specific information provided by VIDEO, procedures were presented, within the VIDEO programs, that provided an expert solution to three of the five problems to be encountered by subjects when troubleshooting live equipment. This enabled an evaluation of the effects of the level of specificity of procedural information presented during training.

#### *Transfer Tasks*

In order to make a determination of the value of the different training methods, the trainees' abilities to troubleshoot live equipment were assessed using a four- and a six-cylinder engine used in modern general aviation aircraft. The five real problems cho-

---

Manuscript received June 26, 1981; revised November 9, 1981. This work was supported by the Army Research Institute for the Behavioral and Social Sciences under Contract MDA 903-79-C-0421.

W. B. Johnson is with the University of Illinois, Urbana, IL 61801.

W. B. Rouse was with the University of Illinois, Urbana, IL. He is now with the Center for Man-Machine Systems Research, Georgia Institute of Technology, Atlanta, GA 30332.

0018-9472/82/0500-0389$00.75

sen represented four engine subsystems: electrical, ignition, lubrication, and fuel. The following is a brief discussion of each failure, its resultant symptom, and a reasonable course of fault diagnosis action.

1) *Starter Lead* (*STARTER*): An open wire was placed between the solenoid and the starter motor. When the ignition switch was activated it would result in a single click from the solenoid. A competent troubleshooter would have used a voltmeter to measure battery voltage followed by a check of voltage up to the solenoid and starter motor.
2) *Spark Plug Lead* (*LEAD*): A defective spark plug wire was routed between the magneto and one plug, resulting in an excessive revolutions per minute drop during the magneto run-up check. The troubleshooter should have noted the excessive r/min drop and noted the drop as being on the right or left side. By allowing the engine to operate with the excessive drop the troubleshooter could zoom in on the failure by feeling for the cold cylinder.
3) *Zero Oil Pressure* (*OIL*): An obstructed fitting resulted in a zero reading on the gauge. The competent troubleshooter had to notice the problem within 30 s of engine start-up. The engine had to be shut down immediately, the oil level checked, and the accuracy of the oil pressure gauge and line established with the use of an auxiliary gauge.
4) *Failed Spark Plug* (*PLUG*): A defective spark plug was installed to cause an excessive revolutions per minute drop during the magneto test. A test to identify the cold cylinder was the best sequence for problem identification.
5) *Fuel Exhaustion* (*FUEL*): The fuel line was obstructed allowing the engine to operate for 35–50 s before stopping completely. A competent troubleshooter should have recognized fuel exhaustion by the symptom. The fuel quantity level should have been checked followed by an inspection of the line for possible obstructions.

### *Measures of Performance*

The accurate and objective measurement of live system troubleshooting performance required extensive planning and the development of a standardized checklist [8]. The evaluation form was designed to allow the recording of each significant step in the process of reaching the solution. In addition, the real time to solution was recorded along with an evaluator's rating which was determined immediately upon completion of each problem.

After the data were collected the checklist information was used to calculate the following measures: performance index—a measure reflecting the mean quality of performance for each subject on each problem; and adjusted time—real time to solution adjusted to the manufacturer's labor time schedule. The performance index, as well as the evaluator's rating, used five-point scales, with five indicating superior performance and one indicating poor performance.

The evaluators were blind as to the training received by each subject. In the first experiment there was only one evaluator. In the second experiment there were two evaluators. All evaluators were experienced power plant mechanics. The evaluator used for the first experiment trained the evaluators for the second experiment in an attempt to standardize the assessment of the performance index and adjusted time to solution. However the evaluator's rating remained completely dependent on the professional judgment of the individual evaluator.

### *Experiment One*

Thirty-six subjects were randomly divided into three equal groups for TASK, FAULT, and VIDEO. All groups received three training sessions which required about 6 h. The live system performance evaluation was accomplished soon after the training was completed, using a single evaluator to collect all data on the aforementioned checklists.

Those subjects trained with context-free simulations (i.e., TASK) solved a total of 60 problems of varying size and level of computer aiding. The context-specific (i.e., FAULT) trainees solved 35 problems including automobile, turboprop, and turbojet simulations. The VIDEO group began each training treatment with a reading assignment on troubleshooting. Then they watched a VIDEO program, each of which was approximately 25 min in duration. Finally, the VIDEO trainees were given a computerized quiz covering materials from the VIDEO programs and the reading assignments.

### *Experiment Two*

Based on the results of the first experiment, the second experiment was modified to utilize a combination of both computer simulation training methods. The control treatment remained essentially unchanged. The second experiment, therefore, had the primary goal of testing an improved combination of context-free and context-specific computer simulations. Twenty-two subjects were randomly divided into two groups for this experiment.

The group receiving the computer simulation training used both TASK and FAULT. Forty TASK problems and 45 FAULT problems were presented to each subject. The TASK problems were identical to the TASK problems in the first experiment. However the FAULT simulation underwent a significant change from the first experiment. The changes were related to the amount of context-specific information provided to the troubleshooter. For example, unlike during the first experiment, FAULT told the trainee how a test was generally performed in the real world (i.e., necessary test equipment and how to interpret results). In addition the new version of FAULT also provided feedback related to inferential errors.

There were other changes between the two experiments. Training time was increased from 6 to 10 h for both groups. This time permitted the computerized training group to solve additional problems. In addition the control group viewed the VIDEO programs twice rather than the single viewing of the first experiment. While in the first experiment only the VIDEO group had the three reading assignments, all subjects in the second experiment were given the reading materials on troubleshooting.

There were also two modifications to the manner in which the real system failures were induced and presented. The first change was the manner in which the fuel starvation problem was induced. In the first experiment a small fuel shut-off valve was used to stop fuel flow. As a result it was possible to use visual inspection to find this failure. The modification for the second experiment involved the installation of an obstructed fuel line fitting which could not be detected with external visual inspection. The second change involved the order in which the problems were presented. For the second experiment, the spark plug lead failure always preceded the spark plug failure. In the first experiment the order of these problems was not held constant.

### *Summary of Results*

In the first experiment, the VIDEO group had a significantly higher mean performance index for the three of five problems whose solutions were explicitly demonstrated in the VIDEO programs. Similarly, the VIDEO group in the first experiment had a significantly higher average evaluator's rating. Both of these differences disappeared in the second experiment. The adjusted time to solution was not statistically different between the training groups for either of the experiments, with the exception of one problem in the first experiment where VIDEO training resulted in faster problem solution.

Why were TASK and FAULT inferior to VIDEO in the first experiment, for the three problems explicitly demonstrated in VIDEO, while the TASK/FAULT combination was similar to VIDEO for the second experiment? The answer lies in a more fine-grained analysis of the errors subjects made during the live equipment evaluation. As will be shown, the error analysis for the data from the first experiment led to the aforementioned changes of FAULT for the second experiment. These changes resulted in FAULT's inadequacies being eliminated.

## ANALYSIS AND CLASSIFICATION OF ERRORS

The troubleshooting checklists noted earlier were very useful for identifying errors. Quite simply, any action that the evaluator checked as "inappropriate" or "most inappropriate" was deemed an error. While this approach to identifying errors does suffer from being subjective, it nevertheless benefits from being able to incorporate the opinions of expert troubleshooters (in this case, three different individuals) rather than the opinions of the experimenters. Further, as noted earlier, all three evaluators were blind with respect to the training method received by each trainee.

For all errors noted during the real equipment performance assessment, the evaluators wrote a one paragraph description of what the trainee specifically did to merit an error designation. The purpose of this paragraph was to provide enough information to allow someone who was not totally familiar with aircraft power plant maintenance to be able to classify the errors.

During the next phase of the analysis, each of the authors independently studied these descriptions and attempted to classify each error. For the first experiment, the authors first tried to use the classification scheme suggested by van Eekhout and Rouse [7]. However this scheme proved to be unsatisfactory.

The main difficulty is that van Eekhout and Rouse studied operators in the process of detecting, diagnosing, and compensating for failures while the studies noted in this paper dealt purely with diagnosis. As a consequence, the classification system of van Eekhout and Rouse aggregates diagnostic errors to a much greater extent than is desirable when the errors are all diagnostic in nature. Further, their scheme includes a category of "choice of goal" which is clearly unnecessary when only one goal (i.e., diagnosis) is possible.

As a result of these difficulties, the classification system of van Eekhout and Rouse was modified and is shown in Table I. The general categories in this scheme are defined as follows.

1) *Observation of State*: Occurred when a mechanic failed to collect sufficient information, misinterpreted the information collected, or collected the same information more than once before forming initial hypotheses about the cause of the symptoms.
2) *Choice of Hypotheses*: Occurred when a mechanic chose a hypothesis that may have been functionally related to the symptoms but was a poor choice because of the nature of the specific symptoms, a very low probability of being true, or a very high cost of testing. Also occurred when a hypothesis was functionally unrelated to the symptoms.
3) *Choice of Procedures*: Occurred when a mechanic's choice of procedure, including informal procedures, was incomplete or inappropriate with respect to the hypothesis being tested. Also occurred when a systematic procedure was not adopted.
4) *Execution of Procedures*: Occurred when a mechanic omitted procedural steps, performed steps out of sequence, etc., or committed apparently inadvertent actions.
5) *Consequence of Previous Error*: Occurred when an error was a logical consequence of a previous error.

Using the scheme shown in Table I, each author again independently attempted to classify each error in the first experiment. Upon completion of their independent classifications, the authors compared results and found almost complete agreement. Of course this may not be surprising since the classification system was developed to fit the data, at least the data from the first experiment. However a partial test of the adequacy of the classification system is reflected in the results of classifying the data from the second experiment where no changes in the classification were allowed. In this case the independent classifications of the authors were identical for almost 95 percent of the errors, with the disagreement mostly attributable to misunderstandings which were easily resolved.

TABLE I
ERROR CLASSIFICATION SCHEME

| GENERAL CATEGORY | SPECIFIC CATEGORY |
|---|---|
| 1. Observation of System State | a. incomplete<br>b. misinterpreted<br>c. repeated |
| 2. Choice of Hypotheses | a. inconsistent with symptoms<br>b. consistent but unlikely<br>c. consistent but costly<br>d. functionally irrelevant |
| 3. Choice of Procedures | a. incomplete<br>b. inappropriate<br>c. lack |
| 4. Execution of Procedures | a. omission of steps<br>b. other<br>c. inadvertent action |
| 5. Consequence of Previous Error | |

## RESULTS

The results are shown in Tables II and III for the first and second experiments, respectively. In order to determine whether or not any of the apparent differences in these tables were other than due to chance, several analyses of variance were conducted. The independent variables in each analysis were training methods (three and two levels for the first and second experiments, respectively) and failures (five levels). The dependent variable in each analysis was number of errors within a general category (i.e., totaled across specific categories).

The ANOVA for the first experiment indicated that the only significant effect was the interaction between training methods and failures for the general category of Execution of Procedures ($F_{8,132} = 2.12, p < 0.05$). Looking at the Execution of Procedures category in Table II, it can be seen that the total number of errors was 5, 12, and 3 for TASK, FAULT, and VIDEO, respectively. FAULT is clearly the significantly different training method.

Again referring to Table II, it can be seen that the STARTER and FUEL failures accounted for 11 of the 12 errors with FAULT. Looking at the previously noted one-paragraph descriptions of these 11 errors, it was found that all of them could be attributed to inappropriate use of test equipment, namely, forgetting to turn the equipment on before using it (i.e., specific error category = omission of steps). This phenomenon, as well as other insights gained by a general review of the error descriptions, led to the conclusion that subjects trained with FAULT knew *what* tests to make but did not necessarily know *how* to make them. The fact that training with the context-free simulator TASK did not result in these errors suggested that the moderate level of fidelity in FAULT might have actually mislead subjects into thinking that they knew how to make tests.

Based on these conclusions, FAULT was redesigned to provide information about how tests were performed and how to interpret test results. As discussed earlier, this new version of FAULT was used in the second experiment. The results of the error analysis for this experiment are shown in Table III. The ANOVA for this data indicated that there were significant differences between problems for two general categories: 1) Observation of System

TABLE II
RESULTS FOR FIRST EXPERIMENT

| ERROR CATEGORIES | | FAILURE | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GENERAL CATEGORY | SPECIFIC CATEGORY | STARTER | | | LEAD | | | OIL | | | PLUG | | | FUEL | | |
| | | T | F | V | T | F | V | T | F | V | T | F | V | T | F | V |
| 1. Observation of System State | a. incomplete | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| | b. misinterpreted | 0 | 3 | 0 | 1 | 4 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| | c. repeated | 1 | 4 | 2 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 0 |
| 2. Choice of Hypotheses | a. inconsistent | 0 | 0 | 0 | 7 | 2 | 2 | 3 | 5 | 5 | 8 | 4 | 4 | 3 | 9 | 2 |
| | b. unlikely | 3 | 1 | 1 | 2 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| | c. costly | 2 | 1 | 3 | 0 | 1 | 0 | 6 | 5 | 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| | d. irrelevant | 4 | 2 | 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 | 2 |
| 3. Choice of Procedures | a. incomplete | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | b. inappropriate | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| | c. lack | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4. Execution of Procedures | a. omission | 3 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| | b. other | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| | c. inadvertent | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5. Consequence of Previous Error | | 4 | 5 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 4 | 0 |

TABLE III
RESULTS FOR SECOND EXPERIMENT

| ERROR CATEGORIES | | FAILURE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GENERAL CATEGORY | SPECIFIC CATEGORY | STARTER | | LEAD | | OIL | | PLUG | | FUEL | |
| | | T/F | V | T/F | V | T/F | V | T/F | V | T/F | V |
| 1. Observation of System State | a. incomplete | 1 | 1 | 1 | 2 | 3 | 4 | 0 | 0 | 0 | 2 |
| | b. misinterpreted | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| | c. repeated | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| 2. Choice of Hypotheses | a. inconsistent | 1 | 0 | 12 | 3 | 0 | 0 | 2 | 1 | 0 | 1 |
| | b. unlikely | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| | c. costly | 2 | 3 | 0 | 0 | 8 | 6 | 2 | 1 | 2 | 2 |
| | d. irrelevant | 5 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 |
| 3. Choice of Procedures | a. incomplete | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | b. inappropriate | 1 | 2 | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 3 |
| | c. lack | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4. Execution of Procedures | a. omission | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | b. other | 4 | 0 | 0 | 3 | 0 | 1 | 0 | 1 | 0 | 1 |
| | c. inadvertent | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 |
| 5. Consequence of Previous Error | | 9 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 6 | 1 |

State ($F_{4,80} = 3.88$, $p < 0.01$) and 2) Consequence of Previous Error ($F_{4,80} = 2.62$, $p < 0.05$). Before considering these differences in more detail, it is worth noting that there were no significant differences between training methods for the Execution of Procedures category. In fact, omission of steps decreased from 14 (all for TASK or FAULT) in the first experiment to 4 (1 for TASK/FAULT) in the second experiment. Apparently, the redesign of FAULT, at least when combined with TASK, had the desired effect.

Considering the significant differences that did emerge in the second experiment, the difference in the general category of Observation of System State is clearly due to the OIL failure. Comparing Tables II and III, it can be seen that TASK/FAULT training improved tremendously from the first to the second experiment for failures other than OIL (i.e., 23 versus 4). The difficulty that the OIL failure presented for both TASK/FAULT and VIDEO groups can be explained by the nature of the failure. For the other four failures, the symptoms were obvious. Either the engine would not start, started then died, or had an excessive revolutions per minute drop. However, for the OIL failure, the engine would seem to be running fine and, if one did not notice the zero oil pressure, it would not be clear that anything was wrong. Thus the OIL failure resulted in more Observation of System State errors because the failure was not obvious without careful observation.

The other significant difference in the second experiment was for Consequence of Previous Error. From Table III it can be seen that the STARTER and FUEL failures caused particular difficulty. An examination of the descriptions of these particular errors showed that almost two-thirds of them could be attributed to a single subject. Thus it would be difficult to draw any general conclusions.

## CONCLUSION

The results of the research reported in this correspondence are of interest in two different ways. First, it has been shown how a fine-grained analysis and classification of human errors led to the redesign of a training program and, subsequently, a substantial decrease in the frequency with which particular types of human error occurred. It is doubtful that such an improvement would have resulted if analysis had been limited to overall performance measures. This is due to the fact that overall measures only indicate how well the trainee is performing but do not necessarily provide evidence about the process of performing the task or specific misunderstandings on the part of the trainee.

The second way in which the results reported here are of interest concerns the general problem of analyzing and classifying human errors. The error classification scheme presented here is for the task of fault diagnosis. As noted earlier, it is a modification of the classification system developed by van Eekhout and Rouse [7] for tasks involving detection, diagnosis, and compensation for system failures. Their scheme is a modification of a very broad scheme developed by Rasmussen and his colleagues [3], [6] for process plant operators in a wide variety of tasks. From this series of modifications, it can be seen that for error classification schemes to be useful they must be adapted to the particular task of interest (i.e., detection, diagnosis, control, etc.), although a special classification system may not be needed for each type of technical system. Alternatively, a general scheme can be developed but it must have numerous fine-grained levels, many of which will only be useful for particular types of tasks.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the help of Ruston Hunt, Richard Henneman, Charles Ziegler, and John Marshall in this work.

## REFERENCES

[1] D. A. Norman, "Categorization of action slips," *Psych. Rev.*, vol. 88, no. 1, pp. 1–15, 1981.

[2] S. M. Halpin, E. M. Johnson, and J. A. Thornberry, "Cognitive reliability in manned systems," *IEEE Trans. Rel.*, vol. R-22, no. 3, pp. 165–170, 1973.

[3] J. Rasmussen, "What can be learned from human error reports," in *Proc. NATO Conf. on Changes in the Nature and Quality of Working Life*, Thessaloniki, Greece, 1979.

[4] F. P. Lees, "Quantification of man-machine system reliability in process control," *IEEE Trans. Rel.*, vol. R-22, no. 3, pp. 124–131, 1973.

[5] H. P. Ruffel Smith, "A simulator study of the interaction of pilot workload with errors, vigilance, and decisions," Moffett Field, CA: NASA Ames Research Center, NASA Tech. Memorandum 78482, 1972.

[6] J. Rasmussen, O. M. Pederson, A. Carnino, M. Griffon, G. Mancini, and P. Gagnolet, "Classification system for reporting events involving human malfunction," Roskilde, Denmark: Riso National Lab., Rep. M-2240, Mar. 1981.

[7] J. M. van Eekhout and W. B. Rouse, "Human errors in detection, diagnosis, and compensation for failures in the engine control room of a supertanker," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-11, no. 12, pp. 813–816, 1981.

[8] W. B. Johnson, "Computer simulations in fault diagnosis training: an empirical study of learning transfer from simulation to live system performance," Ph.D. dissertation, Univ. Illinois, Urbana, IL, 1980.

[9] W. B. Johnson and W. B. Rouse, "Computer simulations for fault diagnosis training: from simulation to live system performance," in *Proc. Ann. Meeting of the Human Factors Soc.*, Los Angeles, Oct. 1980.

[10] W. B. Rouse, "Problem solving performance of maintenance trainees in a fault diagnosis task," *Human Factors*, vol. 21, no. 2, pp. 195–203, 1979.

[11] W. B. Rouse, "Problem solving performance of first semester maintenance trainees in two fault diagnosis tasks," *Human Factors*, vol. 21, no. 5, pp. 611–618, 1979.

[12] R. M. Hunt and W. B. Rouse, "Problem solving skills of maintenance trainees in diagnosing faults in simulated power plants," *Human Factors*, vol. 23, no. 3, pp. 317–328, 1981.

---

# A Study on Human Tracking Performance in a Complex G Field Experiment

DANIEL W. REPPERGER, MEMBER, IEEE, DANA B. ROGERS, MEMBER, IEEE, JOHN W. FRAZIER, MEMBER, IEEE, AND ROBERT E. VAN PATTEN

*Abstract*—**Various aspects of analyzing closed-loop tracking error within a phase plane context for a man–machine experiment are investigated. A study is conducted on candidate density functions that could characterize such boundaries. Using a cumulative distribution function (CDF), a definition of such a boundary is obtained. Kolmogorov–Smirnov (K-S) tests are performed on empirical data from an acceleration stress experiment to investigate certain assumptions concerning normality of the distributions of these boundaries.**

## I. INTRODUCTION

The term performance can have a multitude of meanings when considered within the context of man–machine systems. For the study of humans exposed to acceleration forces, much work has been done both on the physiological effects on humans and performance measurements. The classical works of Fraser [1] and Brown [2] together list almost 300 references describing work done in this area. In the evaluation of human performance under stress fields, the term performance can be used to describe tolerance limits, changes in reaction time, proprioceptive changes, visual perception loss, degradation in scores obtained from different motor related tasks, and a variety of physiological indicators or measurements under the stress condition on humans.

This correspondence will investigate performance measurements in human tracking in a different context. An experiment designed to study G effects on manual tracking will be used to test this procedure [3] and to examine certain properties of human tracking error under this analysis approach. The experiment in [3] provided an excellent data base by which tracking performance showed degradation as the human was subjected to the environmental stressor. This experiment provided modeling answers to a study involving optimal control modeling [8], [9]. Use will be made of this empirical data to examine some assumptions made in [4] concerning performance measures in tracking.

## II. PERFORMANCE MEASURES FOR HUMAN TRACKING

Fig. 1 illustrates the man–machine system of interest in this study. $f(t)$ represents the target input forcing function and $e(t)$ represents the closed-loop tracking error signal which appears on a compensatory display. The human's stick commands may effect the cab roll dynamics of a three-degree of freedom human centrifuge (Fig. 2) to produce G forces on a human subject. The stick commands also drive the "machine" block dynamics to produce an output variable $x(t)$ which is compared to the reference input trajectory $f(t)$. The display error $e(t)$ is the difference between $f(t)$ and $x(t)$.

The purpose of this correspondence is to examine phase plane boundaries of the closed-loop error signal (Fig. 3) and properties of the density functions $f_i(x)$ $i = 1, \cdots, 8$ as displayed in Fig. 3. Several comments, however, should be made concerning this type of analysis. In the man–machine area the phase plane type of analysis is a somewhat archaic method to investigate tracking behavior as compared to more sophisticated methods that occur in the modern control theory applications. This method, however, still provides an interesting approach to obtain information regarding characteristics of human tracking behavior. In previous works, e.g., Young and Meiry [5] used phase plane methods as a

---

Manuscript received June 2, 1980; revised July 27, 1981.

The authors are with the Air Force Aerospace Medical Research Laboratory, Wright Patterson Air Force Base, Dayton, OH 45433.

0018-9472/82/0500-0393$00.75