

Technical Report, NAVAIR 01-1449-1

STUDY OF TRAINING PERFORMANCE EVALUATION TECHNIQUES

David Angell
James W. Spearer
David C. Berlimer

American Institute for Research
Palo Alto, California

10 October 1964

COPIES
NAVAL COPY
MILITARY

U.S. NAVAL TRAINING DEVICE CENTER

PORT WASHINGTON, N.Y.

NTC

Contract No. 33-1339-1449

PROCESSING COPY

ABSTRACT

STUDY OF TRAINING PERFORMANCE EVALUATION TECHNIQUES

The report discusses performance evaluation in the training environment, specifically in training situations involving the use of simulators and other complex training equipment. The important variables involved in developing a system of performance evaluation are seen as (1) types of behaviors, (2) types of measures or mensural indices, and (3) types of instruments for recording performance. Factors relating to these variables are discussed, and some of their interrelationships are delineated. Matrices which facilitate the consideration of interrelationships among the three variables are presented. An illustrative application of an automatic training/evaluation system is given.

O.T.S. AVAILABILITY NOTICE

Stock quantities of this publication are available at Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C., for sale to the general public.

Reproduction of this publication in whole or in part is permitted for any purpose of the United States Government.

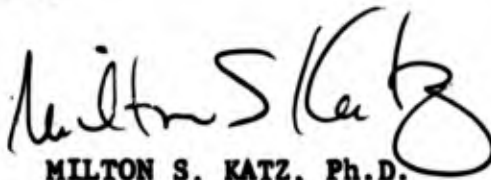
FOREWORD

Meaningful scoring is absolutely necessary to enable objective evaluation of the effectiveness of training devices and curricula. It is a virtual prerequisite to ensure that trainees are in fact learning what they are expected to learn at a reasonable rate, and to be certain that training devices, often extremely expensive, are supplying satisfactory conditions for learning.

The Naval Training Device Center is daily faced with the complex problem of deciding what device is best suited to a given training goal, and how best to equip instructors with a means for increasing student progress for purposes of evaluation and feedback. A derivative, but crucial, use for reliable objective scoring techniques is that they can tell us how successful existing devices are in accomplishing their intended purpose, and how improvements may be achieved in future devices.

Accordingly, the present study represents an orderly look at the many and easily underestimated problems and considerations to be faced in setting up performance evaluation systems. Some aids to thinking about the subject are tendered, while no pat, cookbook solutions can be offered.

A particularly encouraging aspect of this report is its demonstration that analytic techniques can be applied successfully in dealing with what often seems the hopelessly complex and disorderly business of training performance evaluation. In addition, new directions are pointed which hinge on the development and application of computers to training tasks, and open up new vistas for better and better training devices.



MILTON S. KATZ, Ph.D.

Project Psychologist

U. S. Naval Training Device Center

ACKNOWLEDGMENTS

The principal investigator of the project reported here was Dr. Leslie J. Briggs, who is director of the Palo Alto Office of the American Institute for Research, and of the Institute's Instructional Methods Program. Both Dr. Briggs and the members of the project staff who authored this report wish to acknowledge the assistance and advice which was provided by the following persons during the course of the study.

Marshall J. Farr, now assistant head of the Human Engineering Division of the Office of Naval Research, initiated the study and served as the project's Scientific Officer during three-fourths of its duration. When Mr. Farr transferred from the Training Device Center to ONR, monitorship of the project was assumed by Dr. Milton S. Katz, who is chief of the Communications Psychology Division of NTDC. Mr. Farr and Dr. Katz provided timely suggestions and advice, and supported the work in a number of highly contributory ways.

Mr. Ed Fennessey of the San Francisco Regional Office of NTDC provided helpful liaison between members of the project staff and personnel at several Naval installations in Northern California. Mr. Jerry Nelson, Regional Representative of NTDC in San Diego, and several members of his staff--Mr. Jim Cook and Mr. Fred Rothenberg--shared their considerable knowledge of training and training devices with project personnel, and gave valuable assistance to the staff in arranging visits to Naval facilities in the San Diego area.

TABLE OF CONTENTS

| <u>Section</u> | | <u>Page</u> |
|----------------|---|-------------|
| I | INTRODUCTION | 1 |
| | Purposes of Proficiency Measurement | 1 |
| | Times at which Performance is Evaluated | 2 |
| | Simulation and Proficiency Measurement | 3 |
| II | STATEMENT OF THE PROBLEM | 6 |
| | The Military Problem | 6 |
| | The Research Problem | 6 |
| III | METHODS OF OBTAINING INFORMATION | 7 |
| | Literature Survey | 7 |
| | Examination of Devices and Interviews with Users | 8 |
| IV | SUMMARY OF CURRENT NAVY PROFICIENCY-EVALUATION METHODS | 11 |
| | Types and Purposes of Proficiency Evaluation | 11 |
| | Current Uses | 11 |
| | Some Evaluative Instruments | 13 |
| | Some Difficulties in Performance Evaluation | 15 |
| V | DEVELOPMENT OF A BEHAVIORAL CLASSIFICATION | 16 |
| | Choice of a Behavioral Classification System | 16 |
| | Tryout of Existing Systems | 16 |
| | Reliability of Classification | 17 |
| | A System to Enhance Reliability | 19 |
| VI | MEASURES AND MENSURAL OPERATIONS | 22 |
| | Time Measures | 22 |
| | Accuracy Measures | 23 |
| | Frequency Measures | 24 |
| VII | INSTRUMENTS AND DEVICES FOR RECORDING PERFORMANCE . . | 25 |
| | Timers and Counters | 25 |
| | Graphic Recorders | 25 |
| | X-Y Plotters | 26 |

NAVTRADEVCEEN 1449-1

| <u>Section</u> | | <u>Page</u> |
|----------------|---|-------------|
| | Picture and Sound Recorders | 28 |
| | Computer Uses in Performance Evaluation | 28 |
| | Human Observers | 31 |
| VIII | INTERRELATIONSHIPS AMONG PERFORMANCE-MEASUREMENT FACTORS | 32 |
| | Example A: Timing Detection Behavior | 36 |
| | Example B: Timing and Counting Scanning and Surveying Behavior | 36 |
| | Example C: Assessing Error-Amplitude with an X-Y Plotter in a Search Mission | 39 |
| | Example D: Evaluating Sequential Responding in Simple Motor Behavior | 44 |
| IX | IMPLEMENTING AN AUTOMATIC SCORING SYSTEM | 47 |
| | Computerization of Flight Simulators | 47 |
| | Computers Used with Other Simulation Devices | 59 |
| X | SUMMARY AND RECOMMENDATIONS | 61 |
| XI | SELECTED LIST OF REFERENCES | 63 |

LIST OF TABLES

| <u>Table</u> | <u>Page</u> |
|---|-------------|
| 1. Training Devices Examined During the Project | 9 |
| 2. Classification of Behaviors | 20 |

LIST OF ILLUSTRATIONS

| <u>Figure</u> | <u>Page</u> |
|--|-------------|
| 1. Checklist evaluation form for artillery-spotter trainee | 14 |
| 2. Probable relationships between number of task categories and (A) ease of location, and (B) completeness of coverage | 18 |
| 3. Probable relationship between number of task categories and reliability of classifying specific behaviors | 18 |
| 4. Event-recorder transcription (hypothetical) of certain components of a simulated submarine attack mission | 27 |
| 5. Continuous-recorder transcription (hypothetical) of certain components of a simulated aircraft mission | 29 |
| 6. Matrix of behaviors, measures, and instruments relevant to performance evaluation | 35 |
| 7. Measures and instruments appropriate to the assessment of information-seeking activities | 35 |
| 8. Measures and instruments appropriate to the assessment of identification activities | 37 |
| 9. Measures and instruments appropriate to the assessment of information-processing activities | 38 |
| 10. Measures and instruments appropriate to the assessment of problem-solving activities | 40 |

| <u>Figure</u> | <u>Page</u> |
|--|-------------|
| 11. Measures and instruments appropriate to the assessment of communication-type activities | 41 |
| 12. Measures and instruments appropriate to the assessment of simple motor activities | 42 |
| 13. Measures and instruments appropriate to the assessment of complex motor activities | 43 |
| 14. Event-recorder transcriptions (hypothetical) of the performance of a sequential checkout task | 45 |
| 15. Sample of information protocol to be used with computer evaluation of simulator training | 57 |

SECTION I

INTRODUCTION

This report is addressed primarily to the problem of evaluating performance in the training environment. More specifically, it is concerned with automating the recording and assessment of proficiency in simulators, flight trainers, system trainers and other complex and sophisticated training apparatus. The report contains some discussion of considerations relating to measurement in general, because the problem of assessing proficiency during a training program is obviously but a segment of the broader topic--the measurement of human behavior. All aspects of the broader topic have some relevance for the narrower matter of direct concern. For some aspects, the relevance is direct and specific; for others, it is indirect and diffuse. In this report, discussion of topics relating mainly to the broader subject is not intended to be complete nor definitive. Its intention, rather, is to provide a general background against which the more specific topics may be seen in somewhat sharper focus.

Purposes of Performance Measurement

The common purposes served by the measurement and assessment of human performance are useful and important ones. A recapitulation of those major purposes may provide a part of the background for considering performance evaluation in the context of training and simulation. Following are what are seen as the three foremost reasons for assessing performance by the application of standard and objective measurement operations:

1. To determine the adequacy with which an activity can be performed at the present time, without regard, necessarily, for antecedent events or circumstances. These are measures of achievement.
2. To predict the level of proficiency at which a person might perform some activity in the future, if, say, he were to be given instruction concerning the activity. These are measures of aptitude.
3. To observe the effects upon performance of variation in some independent circumstances such as (a) instructional techniques, (b) curriculum content, (c) selection standards, (d) equipment configurations, or the like. These are measures of treatment efficacy.

This list of measurement purposes is shorter than many other such lists. Additional reasons for measuring performance are frequently offered, such as diagnosis of strong and weak areas of proficiency, selection of persons for promotion or advancement or placement, and plotting the rate at which learning is taking place. The view here is that these are not separate and distinct "purposes," but rather that they are special application of measures which are made with the more basic and more general aim either of evaluating present adequacy, or of predicting future proficiency, or of observing the effect of manipulating some independent variable.

Times at which Performance is Evaluated

The purposes which measures are intended to serve determine, or at any rate they interact with, the time when measurement operations are applied. Just as there are three major purposes of performance evaluation, so are there three general time periods during which it is most sensible to obtain measures of performance proficiency. These are (a) before any instruction or training is given, (b) after the completion of training, and (c) while training is going on. It will be convenient to refer to these as initial measures, terminal measures, and interim measures.

Initial Measures. Measures resulting from testing done prior to instruction or practice in an activity are intended mainly to serve the purpose of selection. These measures are obtained with the hope that they will provide information permitting reasonably accurate predictions to be made regarding performance proficiency at some future time. Often, the kind of performance measured to predict future proficiency does not bear a close external resemblance to the terminal behavior whose proficiency is being forecast. It doesn't have to, so long as there is evidence of a close correlation, between the two sets of measures. If individuals who score high on one set of measures also score high on another set, and those who score low on one set score low on the other, then it matters little, for purely predictive purposes, whether the tasks involved appear similar to each other. Accuracy of prediction is not always a function of external similarity of the activities. More important than external similarity is the degree to which the initial and the terminal measures are indices of the same basic behaviors--skills and knowledges--disregarding superficial environmental and contextual features. Accuracy of prediction is also influenced, of course, by the precision and comprehensiveness with which both the initial and the terminal measures are made. For a discussion of personnel selection in the context of modern man-machine systems, see Horst (1962).

Terminal Measures. Post-training measures are ordinarily made when the testees are "on-the-job." These measures are intended to provide evidence of the adequacy with which a task or a mission is currently being performed. Assessing criterion performance will almost always involve sampling from the probably large number of specific behaviors and activities involved in the performance of a job. It will also involve, more likely than not, the use of actual operational equipment. A familiar example of terminal proficiency measurement is the test which an experienced automobile driver takes when he seeks to be licensed in a new jurisdiction. He is asked to perform certain activities which are routine parts of the job, and to demonstrate that he is familiar with regulations which influence, to some extent, his effectiveness as a vehicle operator. And he uses the equipment he will be using on-the-job, in an environment very much like that he will routinely encounter on-the-job. In the military situation, terminal measurements are often made in order to determine the "combat readiness" of a crew or a unit. The reader should see Glaser & Klaus (1962) for a general discussion of the measurement of terminal proficiency.

Interim Measures. There are a number of specific reasons for evaluating performance during training. The measures may be used to assess the effectiveness of the training program, or of some component of the total training system; they may be used to identify trainees who require additional instruction, or different instructions; they may also be used to maintain a high level of motivation. Basically, though, interim performance measures are made because they are more accurately predictive of terminal proficiency than are measures made earlier, before many job-relevant activities had been experienced. Interim measures should be more accurately predictive, because learners, during the training process, will have been exposed to circumstances and events and equipment which resemble those they will encounter on the job, and measures which take such experience into account ought to provide better information about trainees' chances of performing a job satisfactorily than measures which do not include such experiences.

The application of interim performance measuring procedures during a training program inevitably results in some attrition of the student population. This is a reflection of the fact that assessments made at different times (that is, after testees have had different amounts of job-relevant experience) may lead to different results. The measuring tools used for initial selection cannot easily take into account many of the things which may be importantly related to job success, but interim measures take more and more of these into account as the training goes forward.

More than anything else, improvement in the accuracy with which measures predict terminal performance levels is a function of the similarity of the testing conditions to the operational conditions. This fact leads us to a consideration of the role which training equipment, such as simulators and system trainers, may play in the development of a program of performance evaluation during training.

Simulation and Proficiency Measurement

Training in the operation of modern weapon systems and other complex equipment has come more and more to depend upon the use of devices which simulate, with a high degree of fidelity, the operational circumstances of the actual job or actual mission. The use of simulators has increased as operational equipment has become more complicated and more costly, and as the missions or jobs for which training is to be provided have become infeasible or extravagantly expensive to perform in real-world circumstances.

The assumption behind the use of simulators is that transfer from the training situation to the operational situation increases as the two situations become more similar. The assumption is by no means unreasonable, although reliable evidence on this point is scarce. And of course, there are almost always points at which the return in increased proficiency for the dollar spent in faithful simulation diminishes and finally disappears. For some missions--those which are complicated and hazardous and which require highly developed skills--it is essential that the transfer from the training environment to the operational environment be perfect.

First-of-a-kind space flights are dramatic examples of this situation. It is in this pioneering area that simulators achieve the highest sophistication, in terms of their ability to produce realistic circumstances for trainees to practice in. Some remarkable successes in mission performance have been achieved, for which much of the credit must surely accrue to the simulation devices used in training.

Use of Simulators for Training and for Evaluating. Devices as sophisticated as modern simulators, possessing many components of actual operational systems and often controlled by versatile computing machinery, have the capability of serving the purposes both of training and of evaluation. This dual employment of the devices, with each use achieving its maximum potential, is not, in most cases, immediately possible. One factor which makes this so is that the characteristics and features which are required of apparatus whose main function is to provide optimal conditions for learning a task are not necessarily those required of equipment whose primary purpose is to afford the basis for evaluation of proficiency. And vice versa. For example, machines which are intended to provide conditions for improving performance (for teaching, in other words), should provide feedback to the learner which will indicate the adequacy of his behavior, and they ought, ideally, to provide guidance for the correction of imperfect response patterns. (Incidentally, simulators do not ordinarily have these capabilities, and hence they are less than ideal training devices.) Machines whose purpose is to provide conditions for testing performance do not require the feedback and guidance capabilities, but they ought to be able to maintain a record of the testee's behavior in the examination situation. Another problem, of a somewhat different order, is that few designers and users of simulators seem aware of the performance-measurement capabilities of the devices (even though, in a sense, their common use is more testing-like than teaching-like), and consequently the collection of performance data is not made easy nor routine, as ideally it should be. Neither is the interpretation of those data which are collected made a simple and routine matter.

Advantages of Automatic Measurement of Performance in Simulators. Advances in simulator technology, and the development of fast and versatile electronic devices, have opened the possibility of automating many evaluative functions which in the past have been performed less completely, less objectively, and less rapidly, by human observers. One obvious advantage of employing instrument-aided observation and transcription of behavior is the increased precision and increased reliability which come with the decreased dependence upon fallible humans. The instrumentation which is already a part of many simulators provides a solid groundwork upon which to base a program of proficiency evaluation designed to yield accurate indications of the level of competence with which jobs and missions are being performed. Simulators are already being used quite widely to provide proficiency evaluations of a sort, as some years ago Gagné (1954) suggested they were most clearly useful in doing. That the techniques of evaluation are very often subjective, that the "testing" situations are unstandardized, and the results thus quite unreliable, does not diminish the generally solid feeling among persons responsible for training that the simulators, used as they now are, do a highly satisfactory, if not outstanding job of training.

This fact itself suggests one reason for utilizing the performance-measurement capabilities of simulators. Namely, to assist in deciding whether the simulators are in fact effective devices for building skills. That decision cannot be made in a definitive way if there are no reliable indices of performance in simulators. To be sure, the question of whether simulators are effective may be examined by an experimental design in which ultimate criterion performance for trainees who did and who did not have simulator experience (other experiences being similar) is compared. For this comparison, measures of simulator performance are not essential. However, more complete and more precise indications of the degree of transfer from simulator performance to on-the-job performance are obtainable when degrees of simulator proficiency may be correlated with degrees of on-the-job proficiency. To do this, reliable measures of both simulator and on-the-job proficiency are required.

A number of other advantages are to be realized from instrumenting simulators to provide more or less continuous assessment of learning progress. Among the benefits are that (a) timely correction and guidance of learners' behaviors are made possible, (b) instructional procedures may be modified as results indicate their effectiveness, or their lack of it, (c) there may be early awareness of the attainment of desired achievement levels, (d) rates of acquisition of skills may be determined. Records and transcriptions of performance of individuals and crews in practice missions may be analyzed to identify particularly effective or ineffective behaviors, and thus lead to a fuller and sharper description of a job in terms of its essential behavioral components. The records would provide sound data for the development of norms, which would permit more precise evaluations of trainees. Records of simulator performance may also serve as the basis for evaluating selection and assignment procedures, and as dependent variables when alternate training procedures or different equipment configurations are being evaluated.

Of great practical importance are the considerable savings in time, money, equipment depreciation, and sometimes in human life, which are effected by using simulator performance for proficiency evaluation rather than using operational equipment. In the present generation of simulators, missions are simulated with such a high degree of fidelity that simulators may provide conditions which differ only insignificantly from on-the-job conditions, from a point of view of eliciting job-relevant behaviors in real-world circumstances. There is, however, one important difference between missions run in a simulator and those run in operational equipment. That difference is that the simulator permits a firmer exercise of control over measurement conditions and closer standardization of the testing environment. This is a matter of considerable significance, since little value accrues to measures which are unreliable, and since the reliability of performance measures is most importantly a function of controlled conditions of measurement.

All in all, simulators, flight trainers and other large-scale training devices seem to provide promising situations for the meaningful assessment of individual or system proficiency. This report considers behavioral, mensural, and instrumental factors important in planning and implementing performance-measurement systems.

SECTION II

STATEMENT OF THE PROBLEM

The problem, seen broadly, is one of describing methods of assessing the proficiency with which individuals or groups can perform some activity, when the circumstances under which the performance is observed are not entirely like those obtaining when the activity is performed in "real life." From this initial general statement, the problem may be sharpened by restating it, first, in the somewhat narrower context of military training, and second, as a specific research problem which may be approached by the application of techniques of system analysis and behavioral investigation.

The Military Problem

From the operational military point of view, the problem is to develop a set of guidelines and recommendations which will assist design engineers in providing training equipments with a performance-measurement capability, and which will promote the effective utilization of performance-measurement features by training personnel. These aims will be reached by making more objective and more automatic the procedures by which proficiency is evaluated in training systems.

The Research Problem

Viewed as a research activity, the problem is construed as having several components. First, it seems that a relationship needs to be stated between certain kinds of behavioral acts, on the one hand, and certain types of observations of behavior, on the other hand. The systematic statement of such relationships is seen as requiring the categorization of (a) behaviors involved in the performance of military tasks, and (b) measures available for ordering the effectiveness with which such behaviors are performed. Second, it appears essential to indicate some specific methods of accomplishing the measurement operations which are considered to be appropriate to the evaluation of specified behaviors. This is seen as involving (a) the description and categorization of instruments and devices which perform recording functions; (b) the illustration of the kinds of records to be obtained from various classes of instruments as they might transcribe important features of the performance of typical jobs or missions; (c) discussion of ways in which behavioral records may be examined and evaluated; and (d) the exemplification of a system of automated proficiency evaluation.

SECTION III

METHODS OF OBTAINING INFORMATION

The rationale underlying the approach to the problem was that performance-evaluation guidelines would be developed most systematically by examining the relationships existing among, first, the behaviors which are involved in the performance of typical military activities and, second, the types of behavioral records and indices which may best characterize the quality of performance of those behaviors. This rationale provided a basis for selecting areas of literature to review, for defining the analyses to be made and the empirical procedures to be employed, and for suggesting useful ways of presenting information on factors involved in performance evaluation. The steps of the research approach are described in this section.

Literature Survey

The literature pertaining to performance evaluation is extensive. The sizable portion of that literature which was examined during this project is represented in the list of selected references at the end of this report. Annotations are provided for those documents--just less than one-half of the total--which bear most closely upon the topic of the present study.

Three recent reports are concerned with aspects of the same general problem as is considered here. The main concern of a paper by Smode and others (1962) is with proficiency measurement in simulated space-flight missions. That report contains a good discussion of general concepts of measurement, it presents a thorough description of procedures for designing measurement systems, and it gives the results (for the specific mission which is its concern) of a joint consideration of task types and measurement operations, similar in structure, though not in detail, to the consideration of those variables in the present project. An Air Force report which is without formal authorship but which was largely the work of R. Buckhout & T. E. Cotterman (AMRL Memorandum P-40, June 1963) discusses current Air Force proficiency evaluation techniques and considers factors involved in the development of automatic scoring equipment for assessing proficiency in flight simulators. Another Air Force technical report (Benenati and others, 1962) describes a design study of an automatic monitoring system for flight simulators, the main functions of which were, first, recording and playback, and second, evaluation and scoring.

Of these reports, the one by Smode, et al is the more general and the least directly translatable into an operational system. The report by Benenati, et al is very specific, and was intended, in fact, to constitute the first phase of a program which would result in the construction of a model evaluation system. The model was not developed, but essential details for its development are contained in the report. For engineering considerations relating to the implementation of an automatic evaluation system, the report by Benenati and his colleagues is the most informative.

Additional references which have specific relevance for points considered here are cited in their appropriate contexts. Some non-cited but generally pertinent sources are also given later, in the list of selected references.

Examination of Devices and Interviews with Users

To become acquainted with classes of devices currently used to promote training, and with the methods and techniques for employing these devices, a number of simulators and system trainers were observed in operation, and interviews were held with responsible training personnel. A list of the devices examined is given in Table 1 on the following page.

On the basis of thorough examination of the Navy Stock Lists of Training Devices, the particular devices examined seemed to represent quite fairly the class of complex trainers which could provide most valuable proficiency-related information, if they possessed performance measurement capabilities of a sort which could be routinely used by training personnel.

In the attempt to find out what provisions exist for objectively measuring performance in complex simulators, and to discover how proficiency is presently being assessed during training, and to become familiar with the general and special problems that are involved in evaluating performance in simulated environments, a list of topics was drawn up which lent structure to, and provided standardization for, the interviews which were held with training personnel. The following questions formed the basis for discussions with individuals responsible for training. The questions, as they are presented here, serve mainly to indicate the topics discussed; obviously, they were not posed verbatim as they are here recorded. Neither were the topics necessarily discussed in the order in which the questions are here listed. Very often, discussion which ensued from one question would lead into other related matters, where valuable information was also obtained. Here then, succinctly, are the questions which structured the discussions with training personnel:

1. What kinds of records are made of the trainees' actual performance in the simulator?
2. What kinds of scores or proficiency indices are assigned?
 - a. What kinds of performance are the scores based upon, and
 - b. At what points in training are various scores given?
3. Are standardized problems and simulator conditions used, especially when comparative indices of performance proficiency are given?

NAVTRADEVCEEN 1449-1

Table 1

Training Devices Examined During the Project

| <u>Device Number</u> | <u>Name</u> | <u>Location</u> |
|--------------------------|--|--------------------------------------|
| 21A3 | Submarine Attack Teacher, Mk7 | Mare Island Naval Base |
| 21C4 | Submarine Diving Trainer, Askania | Mare Island Naval Base |
| 2F76 | Weapon System Trainer, Aircraft A4D5 | Lemoore Naval Air Station |
| 2F62 | Weapon System Trainer, Aircraft A4D-2N | Lemoore Naval Air Station |
| 2F34A | Weapon System Trainer, Aircraft P5M-15, Modified | North Island Naval Air Station |
| 2F70 | Weapon System Trainer, Aircraft P2V-5FS ASW | North Island Naval Air Station |
| 16C54 | Amphibious Supporting Arms Evaluator | Naval Amphibious School, Coronado |
| 14H3 | ASW Coordination Trainer, Interim | FAETUPAC |
| RS14 | CIC Trainer; Simulation Equipment | Fleet AAW Training Center |
| X14E3 | Sonar Operator's Target Classification Trainer | Fleet ASW School |
| 21B52 | Radar Simulation System Emer- gency Ship Handling Trainer | Fleet ASW School |
| 14A1 | Action Speed Tactical Trainer | Fleet ASW School |
| ---- | Ames Five-Degree-of-Freedom Manned Flight Simulator | NASA Ames Research Center |
| X-14A | VTOL Test Vehicle | NASA Ames Research Center |
| ---- | Transport Landing Simulator | NASA Ames Research Center |
| ---- | Ames Mid-Course Navigation and Guidance Simulator | NASA Ames Research Center |
| ---- | Unlimited Angular-Motion Simulator | NASA Ames Research Center |

4. Are the instructors familiar with the objectives and techniques of performance measurement? Do they know, or do they have any feeling for, the requisite characteristics of accurate measurement; objectivity, reliability, validity?
5. How objective are the measures which are used? (The question, considered conversely, has to do with the extent to which evaluations may vary as a function of observer differences.)
6. To what extent do the observations made and the measures used assess performance that is critical, in terms of the effect of the performance upon mission success or failure.
7. How are cut-off or wash-out standards determined?
8. Do the operating manuals or handbooks for the training devices specify or suggest performance-evaluation procedures?
9. Are evaluations ever made of the validity with which performance measures during training indicate the level of job proficiency?
10. What capabilities (if any) would users like for the device to have, both for training and evaluation, that it does not presently possess?
11. What training is provided for the users of the training equipment, i.e., for the instructors directly responsible for giving training with the devices?
12. What are the main characteristics of the population of trainees using the devices?

In the section which follows immediately, there are summarized the conclusions and impressions pertaining to the current state of proficiency evaluation in the Navy drawn from the recent literature and from interviews with personnel responsible for conducting training. Formal survey methods were not used, and there are no "hard" data such as percentages or central-tendency measures to describe current methods and attitudes concerning evaluation. (Other recent studies--e.g., Harris & Mackie, 1962--have employed the more formal approach.) While the observations and conclusions given below are not supportable by empirical documentation, a measure of confidence in their soundness arises from the frequency with which similar points were made by persons in quite different situations, and from the compelling obviousness of certain observations which could reasonably lead only to certain conclusions.

SECTION IV

SUMMARY OF CURRENT NAVY PROFICIENCY-EVALUATION
METHODS

Types and Purposes of Proficiency Evaluation

The proficiency-measurement methods used by the Navy can be grouped into three main classes; written tests of knowledge, performance tests, and judgments of trained experts.

Written tests are typically tests of facts and relationships which are designed to measure how much a trainee knows about a particular job or type of equipment. Performance tests are actual or simulated work situations in which a trainee demonstrates his mastery of skills required for a particular job or type of job. The third type of evaluational method consists of making use of personnel who are experienced and highly proficient in the task or series of tasks on which performance is being evaluated. The qualified "expert" observes trainee performance and makes judgments regarding the effectiveness of that performance, on the basis of his experience. One variation of knowledge-type tests which is quite often used in conjunction with complex simulator training is the oral test, during which an instructor asks a trainee questions pertinent to the tasks which the trainee is in the process of performing or has already completed. This type of question/answer session is a common occurrence when a simulator mission is "frozen" in order that various aspects of trainee performance can be discussed and trainee errors pointed out during the mission, and also in debriefings which follow training missions.

All of these evaluation methods are used independently and in combination to perform evaluative functions in the Navy. Such methods are used for a number of purposes, including:

- (a) to help in assigning men to billets;
- (b) to evaluate whether personnel qualify for advancement in ratings;
- (c) to measure qualifications of trainees entering school;
- (d) to measure achievement of trainees at various stages within training sequences;
- (e) to measure achievement of trainees at course termination; and
- (f) to assess the effectiveness of training programs.

Current Uses

A report by Harris & Mackie (1962) describes their study of the extent to which various evaluation methods are used in the Navy, and the factors which influence the use of performance tests. For the sample

examined, it was found that 87 percent of instructors in the school environment used written tests to evaluate trainee performance, 72 percent used performance tests, and 74 percent used expert judgments. For supervisors in the operating environment, the corresponding percentages obtained were 18 percent, 17 percent, and 97 percent, respectively.

Although no formal survey of proficiency-measurement methods used by training instructors was conducted as a part of the present investigation, information collected on an informal basis through discussions with administrative personnel, instructors, trainees, maintenance personnel and technical representatives would seem to support the conclusion that, although some form of evaluation is almost always performed in training situations involving use of complex simulators, the proportion using performance tests effectively for evaluative purposes is rather low. Among the reasons voiced by training personnel and/or observed by project staff members as to why performance tests are not made better use of, are:

1. Equipment used for recording and evaluating performance does not work or is always in need of adjustment, alignment or some other type of maintenance (which takes too much time).
2. There are not enough qualified maintenance personnel available to keep equipment running.
3. Replacement parts cannot be gotten in reasonable time.
4. Instructors are sometimes maintenance technicians who have no experience in operating the equipment being simulated, and they are therefore not qualified to evaluate many aspects of trainee performance.
5. Instructors do not know how to evaluate performance objectively.
6. No good performance tests are available.
7. Training schedules are tight and instructors are so concerned with adhering to schedules that there is no time for administering performance tests.
8. Simulators are used for familiarization and practice, much like cockpit procedure trainers, rather than for more formal training involving guidance of responses, feedback, etc.
9. The task is such that objective evaluation of proficiency is not considered to be realistic or feasible.

10. Training personnel develop custom-made evaluation instruments which are often characterized by face validity but are of little practical value.

The above are similar to questionnaire data collected by Harris and Mackie (op. cit.) regarding the least favorable features of performance tests, and the problems involved in their use.

Some Evaluative Instruments

Written tests of different varieties that are used by the Navy, such as multiple-choice, matching, completion (short answer and essay), to name a few, are familiar to most readers since such tests are widely used in academic environments.

The rating scale is one type of evaluative tool which is often used in assessing an individual's effectiveness on the job. Ordinarily, a rating scale will describe some trait or activity which is being rated, and will give qualitative descriptions of different amounts of the trait, or of different degrees of excellence of performance of the activity. The intent is to make the rating procedures more objective and less subject to observer bias or ignorance, by enabling the rater simply to match the behavior of the ratee with some external description of behavior. Good rating scales provide raters with some common and familiar frame of reference to establish an objective basis for estimating whether an attribute is possessed in high or medium or low degree, or whether a behavior is performed with a high or medium or low degree of skill.

The behavioral checklist is another common evaluation instrument which is essentially a sequential listing of subtasks or tasks to be performed. To use a checklist an instructor merely checks each item on the sheet as it is performed and thereby obtains a record of which tasks were performed, and whether or not they were performed in correct sequence. A refinement of the checklist, aimed at making it a more discriminating instrument, is to have the rater indicate whether each task performed was done satisfactorily or unsatisfactorily. Hopefully, there would be some objective basis for making such decisions. Performance scores are developed by summing points of weights which have been assigned to the tasks observed. A variation of this procedure involves subtracting points from some total assigned at the start of an evaluation session, for tasks which the trainee omitted, or performed incorrectly or imperfectly. Figure 1 on the following page illustrates such a checklist. At the time of this study, that checklist was being used as one aid in evaluating the performance of Marine artillery-spotter trainees.

Expert judgments on the part of experienced personnel are used extensively to evaluate performance. One way this method is used can be illustrated by an instance in which an instructor observes and listens while a trainee or group of trainees runs through a simulated mission, and, upon completion of the tasks involved, rates the performance along some dimension which is described by terms such as pass/fail; good/fair/poor; or acceptable/unacceptable. Occasionally, trainees are assigned

| | | |
|---|------------------------------------|-----------------------------|
| Ammo expended: | <u>5</u> | |
| Error in voice radio procedure | | - 3 |
| Use of wrong terminology | | - 3 |
| Incorrect sequence of commands | | - 3 |
| Wasted time | | - 5 |
| Omission of any element | | - 5 |
| Target location error: | | |
| 200-600 yards | | - 5 |
| over 600 yards | | - 10 |
| Height error in excess of 100 ft. | | - 5 |
| Poor target description | | - 3 |
| Improper classification | | - 5 |
| Wrong number of guns or armament | | - 3 |
| Poor choice of ammunition | | - 5 |
| Poor choice of fuze | | - 5 |
| Omission of any element of a subsequent command | | - 5 |
| Wrong sensing | | - 5 |
| Failure to establish bracket, or improper handling of bracket (per wasted round) | | - 10 |
| Failure to hit target in FFE due to improper handling of bracket | | - 10 |
| Failure to give initial height spot | | - 10 |
| Name <u>Pfc Nichols</u> | Instructor <u>OP Durney, Capt.</u> | |
| Class <u>64-17B</u> | Date <u>4/4/64</u> | Grade: <u>100 - 23 = 77</u> |

Fig. 1. Checklist evaluation form for artillery-spotter trainee.

numerical scores on the basis of subjective judgments of performance without any trappings of objective evaluation. Another way in which the judgmental method is used involves basing an evaluation on a combination of objective data such as errors collected during a simulated exercise and subjective judgments of general proficiency.

Some Difficulties in Performance Evaluation

An evaluation instrument such as that shown in Figure 1 is often deceptively precise. The nice indisputable numbers tend to obscure the fact that some aspects of the performance may be so loosely defined as to be subject to varying interpretations by different raters. In the illustration, for example, it is not hard to imagine that not all raters will have the same ideas about how much time constitutes "Wasted time." Carelessly used, by observers who are untrained in evaluation methods and unaware of the requirements for valid measurement, devices like these can lead to a sort of pseudo-quantification which is scarcely better than no measurement at all. Numerical scores are assigned which lack real meaning and which bear a largely accidental relation to true proficiency, and they may then be treated as though they constituted incontrovertible evidence of actual level of performance effectiveness. The authors watched an artillery-spotter trainee run through an exercise observed by an instructor who had in hand the checklist illustrated above, but who did not once consult nor mark that list. At the end of the exercise, the instructor wrote 72.5(!) as the trainee's grade.

Quite probably, the effective use of performance tests which depend on humans for observation and evaluation is inversely related to the complexity of simulator systems. One reason for this is that a mission in which a simulator is a part, can, and very often does, involve evaluation of more than one trainee. It may be a crew, a team, or a multi-team operation which is being trained (and which needs to be evaluated), rather than a single individual. Also, as the scope and complexity of simulated missions increases in terms of length of time required and numbers of men and equipment units involved, the importance of the part played by communications and coordination increases tremendously. As a result, the already difficult job of evaluating performance becomes even more difficult for the instructor, and it becomes more and more practical to turn to automatic (computerized) evaluation techniques.

The discussions with training personnel and the observations of devices in use did not suggest any new approaches to a fruitful solution of the problem of simplifying performance evaluation in training; neither did they indicate anything calling for modification of the rationale, tentatively adopted at the start of the project, which suggested that behaviors and activities needed to be systematically enumerated and categorized, in order to be related in meaningful and useful ways to different measurement operations. Thus the next step taken was to develop a behavioral classification system, as a step towards specifying behavioral/mensural/instrumental relationships.

SECTION V

DEVELOPMENT OF A BEHAVIORAL CLASSIFICATION

Choice of a Behavioral Classification System

The choice of language with which behaviors would be described is one of the first problems faced in the development of a behavioral classification system. Any decision about this is influenced by a consideration of the audience for whom the guidelines and recommendations on performance evaluation were being formulated. The audience was presumed to consist mainly of training personnel and design engineers, and it was assumed that not many of them would be conversant with the rather specialized descriptive language with which psychologists often describe behavioral processes. What was sought was a terminology whose units would be interpreted identically by the majority of persons, and which would be familiar and usable.

The behavioral activities which are capable of being identified most reliably are those simple acts which may be labeled by common active verbs such as "rotates," "guides," "holds," "tracks," and the like. All activities are capable, eventually, of being expressed in such elementary terms as these. The processes of task description and task analysis, integral to the design and development of modern man-machine systems, aim at describing in these basic and widely understood terms, tasks and processes which in their totality are highly complex and involved. It was felt that if measurement methods could be related specifically to behaviors which are (a) capable of being reliably identified, and which are (b) quite simple acts, the performance of which may often be evaluated by relatively simple means, and which are (c) general in their occurrence, hence involved in a great many military jobs and missions, then the evaluation techniques proposed would have wide utility and application.

Tentatively, the basic behavioral units to be employed were identified as an as-yet-undetermined number of action verbs, describing fairly simple and easily observable activities of the sort involved in the performance of military tasks. The verbs would constitute an enumeration of assessable behaviors. Taken by itself, such a list would seem not as useful, and not as potentially instructive, as it would if the behavioral units were grouped according to certain characteristics possessed in common. In other words, classification was seen as organizing and providing a necessary structure for the system.

Tryout of Existing Systems

Behavioral classification systems have been proposed by a number of writers. Generally, these have been attempts to establish relations between types of behavior and conditions of training. The comparable but not identical problem of the present project has been to establish relations between types of behavior and conditions

of measurement. Because conditions which are optimal for training are not necessarily optimal for measurement, there was an awareness that the categories of behavior which had been suggested by others as meaningful ones for consideration in arranging instructional conditions might not be appropriate for consideration in arranging conditions of measurement. Still, a number of previously-suggested systems of behavioral classification were examined, both with the specific purpose of evaluating their appropriateness to project aims, and with the general purpose of gaining familiarity with rationales and procedures for developing behavior taxonomies.

As one step in the exploratory approach toward determining feasible behavioral categories, a number of tasks were "invented" by pairing action verbs with referents relating to equipment, or to systems, or to some other part of the environment; some examples of these tasks are: "aligns scope crosshairs," "estimates distance," "interpolates map coordinates," and "identifies target." The staff members (and others) then attempted to pigeonhole some 40 tasks such as these into the categories of various suggested classifications. The categories were those used in the classification systems proposed by Miller (1962), Gagné (1963), Willis (1961), Smode, Gruber, & Ely (1962), and Berliner (unpublished). The number of categories in a system varied from six to twenty.

There was not very good agreement between different persons with respect to the categories into which tasks were sorted. The longer and more complex classifications tended to yield quite unreliable sortings. On the basis of this casual finding, some other relationships between the number of task categories and various dependent variables were examined.

Reliability of Classification

It became apparent that using a behavioral classification system whose basic elements were specific, clearly-described activities would require that the user locate a particular behavior via an indexing system of some sort. Ease of location then, was probably an important factor to be taken into account in developing a behavioral classification. It is apparent that it becomes less and less easy to locate a particular task in a system as the number of categories in the system increases. A negative relationship exists between ease of location and the number of categories which are used. The opposite relationship holds, however, when completeness of coverage is considered as a function of number of task categories. For the classification systems which were being used (but not for all systems), a greater number of task categories tended to provide more extensive coverage of the range of behavior tasks being assigned to those categories. These two opposite relationships--ease of location and completeness of coverage as functions of the number of task categories--are illustrated quite simply as follows (Figure 2):

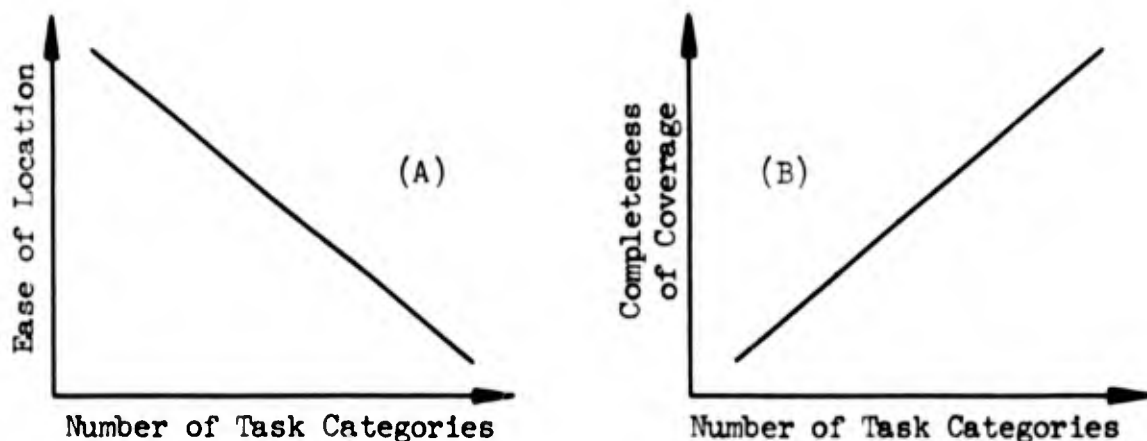


Fig. 2. Probable relationships between number of task categories and (A) ease of location, and (B) completeness of coverage.

The relationships are shown as linear, but of course they are illustrative only. The directions of the trends are probably indicative, but the precise shapes and slopes of the curves are not known.

Finally, when the reliability with which tasks are pigeonholed into categories was examined more closely, as a function of number of categories used, it appeared that reliability would be higher with either a small number or a large number of categories, and lower with an intermediate number. Obviously, reliability is perfect when there is but one category, and it is also perfect when the number and kind of categories exactly matches the number and kind of tasks being sorted. But between these two extremes, reliability is less than perfect. The form of this relationship is probably something like that shown below.

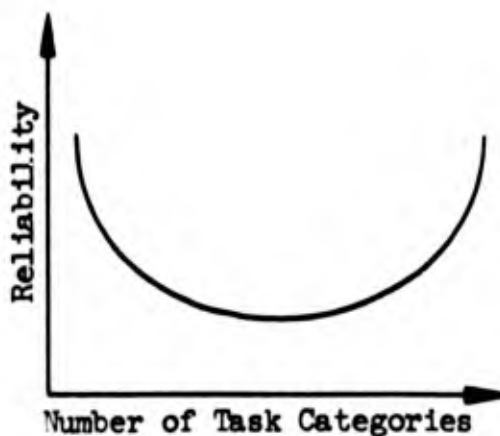


Fig. 3. Probable relationship between number of task categories and reliability of classifying specific behaviors.

Taking the three relationships into consideration, it seemed that a classification system which employed few broad categories would lead to the reliable and relatively easy location of behaviors in the taxonomy, perhaps at the expense of complete coverage. Using many narrow categories would give good coverage and good reliability, but would make location difficult. The middle course, involving the use of some intermediate number of task categories, might give fair coverage and provide for a fairly easy (but less than perfectly reliable) location of behaviors.

A System to Enhance Reliability

There is an alternative to the middle course, however. A multi-level classification system could be employed, containing several broad general categories, under which there would probably be organized some sub-categories, and finally, a number of quite specific tasks or behaviors. A classification system of this kind is, in fact, just about what would result if the system were developed beginning with an enumeration of the smallest and most narrowly defined activities and working from there to the more molar and inclusive levels of description. Since the tentative decision had been made, on other grounds, to use specific behaviors as important units of the behavioral classification, the findings and considerations discussed above reinforced that decision.

A list of some 100 active verbs, representing activities involved in the performance of military tasks and missions was compiled from various sources. The behaviors represented by the words (no referents were associated with them) were analyzed to identify commonalities which would afford a basis for categorizing them into a smaller number of more general behavioral processes. As trial classification systems were developed, their usefulness was examined by having persons sort the activities (eventually reduced to a list of 78, by eliminating some synonyms and other words) into the tentative categories. To assess the degree to which the categories were mutually exclusive, the sorters were shown only one category designation at a time, and they decided, for each activity, whether it fit that category or not. The words were sorted by each person as many times as there were categories being examined. By making changes in the category designations, and by combining some of the categories and fractionizing some others, small improvements continued to be effected in the amount of agreement between different observers as to which specific behaviors fit which general categories. The best results, in terms of interobserver agreement, were obtained finally with a system in which four major behavioral processes encompassed six broad types of activities, under which there were subsumed, in turn, some 50 specific behaviors. Table 2 on the following page shows the processes, the activities, and some of the specific behaviors.

A considerable amount of overlap remains in the system. The categories are not mutually exclusive. That is, not all the behaviors are seen by observers as belonging in single categories, without fitting any others. While mutual exclusiveness of the categories is certainly a desirable characteristic of a classification system, it is not an essential one. R. B. Miller has pointed out (Miller, 1962) that mutual exclu-

Table 2

Classification of Behaviors

| <u>Processes</u> | <u>Activities</u> | <u>Specific Behaviors</u> |
|-------------------------|---|---|
| Perceptual Processes | Searching for and Receiving Information | Detects Inspects Observes Reads Receives Scans Surveys |
| | Identifying Objects, Actions, Events | Discriminates Identifies Locates |
| Mediational Processes | Information Processing | Categorizes Calculates Codes Computes Interpolates Itemizes Tabulates Translates |
| | Problem Solving and Decision Making | Analyzes Calculates Chooses Compares Computes Estimates Plans |
| Communication Processes | Communicating | Advises Answers Communicates Directs Indicates Informs Instructs Requests Transmits |
| Motor Processes | Simple/Discrete | Activates Closes Connects Disconnects Joins Moves Presses Sets |
| | Complex/Continuous | Adjusts Aligns Regulates Synchronizes Tracks |

siveness of terms in a taxonomy may in fact be an unattainable objective, and the findings of the present study provide no evidence to disconfirm this. Nevertheless the system as it has been developed to this point does show that observers with rather diverse backgrounds and interests can agree quite well on whether or not a specific activity possesses characteristics which put it in a class of behaviors whose general nature is described by some broad behavioral-process designation.

The enumeration of specific behaviors in the above classification system clearly falls short of exhaustiveness. The behaviors appearing in Table 2 are those which passed an informal screening to select words more appropriately descriptive of activities involved in the performance of military missions, and to eliminate those with most obviously ambiguous or equivocal meanings. Further, the words are, from an initial population of about 100 verbs, those whose classification into the activity categories shown was agreed upon by at least six of the eight persons who participated in the categorizing exercise. (Other trial categories, remember, had led to agreement on even fewer words.) There is not much reason to think that the 47 words in Table 2 are adequate to describe all jobs and missions. One of the early assessments of the usefulness of this system ought to be the determination of whether adequate task analyses may be made using such a limited vocabulary of action verbs.

A classification system something like that shown here should prove a useful aid in performance-evaluation decisions, to the extent that jobs or missions are describable in terms of such action verbs as constitute the lowest level of the system, and to the extent that the higher-order category headings represent processes or activities unitary enough to possess some common implications for their measurement.

Regardless of the ultimate adequacy or inadequacy of the present classification scheme, the use of a system such as this requires that task-analysis procedures be employed to produce a description of the job or mission in terms of action verbs of the class shown in Table 2.

SECTION VI

MEASURES AND MENSURAL OPERATIONS

The activities subsumed in the behavioral classification system may be viewed as constituting one dimension of a three-way matrix. A second dimension is represented by a classification of measures, or mensural operations. When these are considered together with types of instruments for recording behavior, a basis is provided for personnel or system evaluation.

For purposes of exposition, we wish to distinguish between measures and scores. Measures refer, in this paper, to raw records of performance, in terms of one of several measurement dimensions. Scores are more refined indices because they are reduced records or abstractions of measures and because they usually imply a comparison of a raw record, or measure, with some sort of standard performance.

The objectively describable dimensions along which performance of most tasks may vary are few in number. Variations in no more than three basically different mensural indices suffice to provide virtually all the information required for proficiency evaluations having high precision and good discriminability to be made. Those indices are (a) time, (b) accuracy, and (c) frequency. Other authors, it should be noted, have considered a far greater number of "measures" to be appropriate to the assessment of performance in a simulated environment. Smode, Gruber, & Ely (1962) catalogued no fewer than 83 separate measures which they have grouped under 21 class headings. The reader should consult that report for a different perspective on measures, as well as for a more thorough coverage of basic factors involved in performance evaluation, than is given here. For activities of the relatively narrow compass of those constituting the lowest level of the behavioral classification system adopted here, however, combinations of values of the variables time, accuracy, and frequency seem adequate for obtaining performance records with evaluative capabilities. Scores derived from any of these measures ordinarily would reflect weighting on the basis of the criticality of the behavior.

Time Measures

Three time measures appropriate to performance-measurement purposes are:

- (a) time to respond to a signal (reaction time),
- (b) time to complete an activity after first responding to a signal, and
- (c) total time, from signal onset to completion of the activity.

The first two of these measures are obviously components of the third; summed, the components equal total time. Total time is usually the most relevant and representative index of time for performance eval-

uation. The decision whether to use one of the components rather than the total might hinge upon such a matter as ease of instrumenting the recording of the event. The decision as to which component to use could depend upon the fact that one component contains a major portion of the variance while the other contains an insignificant part. For example, the time to respond to a signal could be approximately the same for each respondent but the time to complete the activity could vary among these individuals. Other things equal, the general rule is that measures should be made on those aspects of behavior which show the wider range of individual differences.

Many military tasks require the sequential performance of a number of distinct operations. Deviations from a certain serial order might "abort" the entire task or might have a depressing effect on performance effectiveness by increasing the time required to complete the task. The time measure here is obviously confounded with an accuracy measure, and the interaction could be complex and difficult to quantify and assess. The extent to which completion time of the task would be influenced by an error in sequence of performance depends upon a number of factors: (a) how long it takes for the error to be observed; (b) how much "retracing" has to be done (itself a function both of (a) above and also of the particular nature of the task); (c) how long the total task normally takes, and the number and size of the discrete sequential steps which make up the total task. In a procedural task of any considerable length, the number of error possibilities and error/time combinations becomes very large. Therefore, if time measures are to be used for sequential tasks, computing machinery will probably be required to provide meaningful evaluations (see Section VII, p. 30, and Section IX, p. 47).

Accuracy Measures

Accuracy--the major indices of which have to do with the occurrence or non-occurrence of errors--is the second major category of measure relevant to the types of behaviors thus far described. Aspects of accuracy which may relate quite directly and importantly to proficiency are

- (a) whether a correct choice is made when several alternative actions are possible,
- (b) whether responses are performed or actions taken in the correct order when sequential responding is required,
- (c) error frequency, and
- (d) error amplitude.

By error frequency, is meant, simply, the number of times errors are made during a given period. Error amplitude refers to the size of the discrepancy between an individual's (or crew's) performance of some act and an ideal standard of performance of that act. An example of the latter might be the difference, in engine revolutions, between a pilot's power setting for take-off and the optimal power setting for that particular aircraft, its present load, the runway length, the wind conditions, etc. Another could be the difference, in degrees, between a sonarman's estimate of the bearing of a target from his submarine and the target's true bearing from the submarine. If some of the attributes of measures

can be thought of as applying to performance, then error frequency would seem to indicate something about the reliability of the responding (i.e., is the same response or pattern of responses--correct or incorrect--made repeatedly on different occasions?), and error amplitude would seem to indicate something about the validity of the responding (i.e., how closely does the form of a response correspond to what the form should be?). Amplitude and frequency are the aspects of error occurrence which are most amenable to automatic recording, a consideration which is of some importance and which will enter into the decisions made regarding appropriateness of measures to behaviors.

Error amplitude, incidentally, is here conceived broadly enough to include such things as averages of deviations which have been sampled over time, and integrated absolute error, often used to measure accuracy of performance in a tracking task.

Frequency Measures

Frequency of occurrence is a measure which obviously applies to events other than errors. Conceivably, there are behaviors whose performance proficiency may be assessed simply by observing how often the act occurs in a specified time period. As one example, an individual might be required to make at least so many verbal reports, content unimportant, in every so many minutes (or hours), to provide evidence of his alertness, or at least of his consciousness. As another, quality of performance could be positively related to the frequency with which a person scanned a scope, or observed an instrument or sought information of some kind from some source. Frequency of occurrence of some event when considered for a given time period, yields an index of rate of behaving which is a common performance measure.

SECTION VII

INSTRUMENTS AND DEVICES FOR RECORDING PERFORMANCE

Decisions on how to instrument a performance-measurement system depend, of course, upon the behaviors chosen to be observed and the measures used to discriminate differences in the performance of those behaviors. Devices appropriate for providing measures of the behaviors involved in military-type tasks appear to fall into four general categories: (a) timers and counters; (b) graphic recorders; (c) plotters; and (d) sound and picture (transcription) recorders. Electronic computers are viewed as constituting a class of devices somewhat separate from those identified above, and their use is considered in a separate part of this section. Human observers are also considered separately.

Timers and Counters

Time is an important dimension of many different behaviors of many military tasks, and consequently devices which measure time are among the most appropriate and useful of any in the instrument inventory. Timers may range, in accuracy and versatility, from common wall clocks, through pocket stop watches and electric chronometers, to complex vacuum-tube or solid-state electronic instruments. The capabilities of timers, in terms of their precision and accuracy, exceed that which is required for most purposes. Differences smaller than, say, tenths of a second are not likely to have great significance in the performance of most military tasks, and even simple and relatively inexpensive stop watches and chronometers are capable of providing measures accurate to hundredths of a second. In some cases, considerations relating to storage capacity, sequence handling, and integration with other functions might be important.

Counters also vary in their complexity. They range from simple hand-operated devices to complex, extremely fast electronic counter-printers. Whatever their form, counters provide information about the numerical characteristics of some variable. A number of potentially meaningful performance indices could result from the joint use of timers and counters. In simulated submarine missions, for example, a timer could record total length of time a sub's periscope was above water, while a counter kept track of the number of separate times the periscope was raised and lowered during a mission. Combined, the two measures yield an index of average scanning time, an aspect of performance which might discriminate between experienced and inexperienced submarine personnel. Wherever measures of the frequency of occurrence of some event per unit of time are desired, counters and timers used in combination will provide such rate indices. Other devices will yield this same sort of information in somewhat different form.

Graphic Recorders

Devices in this class are electromechanical in operation, and they provide continuous records of the states or the magnitudes of events or variables. The records are plotted, as a function of time, on moving

graphic charts. Devices are made which will accept either analog or digital inputs, which have varying numbers of channels (up to hundreds), which use different means of writing or scribing, and which operate at various speeds. Because of their versatility, they are useful for capturing records of many different components of an activity. There are two major groups of graphic recorders: event recorders and continuous recorders. An event recorder monitors status situations such as "yes-no" or "on-off." A continuous recorder displays a continuous record of the magnitude of some variable. The two types differ very little in terms of engineering implementation.

Event Recorders. These instruments can record such things as time of occurrence, sequence, and duration of a considerable variety of events which may be represented by the opening and closing of relays, switches, valves, circuit breakers, and telephone circuits. They are appropriate devices for preserving performance records of sequential behavioral actions, such as those which take place during the pre-flight checkout of an aircraft. With its time base, an event recorder could also be used to record the time of occurrence of a simulated malfunction inserted into a system, and the time and nature of the corrective action taken by the trainee. Figure 4 illustrates the type of record which could be obtained from an event recorder.

Continuous Recorders. These devices can provide a continuous analog representation of many different kinds of variables. One could be used, for example, in conjunction with a flight simulator to yield records of power application, of angle-of-climb and angle-of-descent, of speed, of heading deviations (from preset headings), of yaw and roll deviations (from zero), and of deviations in gravitational force (from any preset "G" value). In a submarine simulator, such variables as diving angle, keel depth, bow angle, and stern angle may be recorded, as well as power consumption, heading, and speed. An illustrative example of a continuous recorder's transcription of an aircraft flight is shown in Figure 5.

Some recording devices have channels for both analog (continuous) and digital (event) representation. Such an instrument could record flight characteristics such as those above, and in addition could show the times of occurrence of simulated malfunctions and corrective actions. Many of the performance data required to yield precise evaluations of simulated missions are capable of being recorded by this class of instruments.

X-Y Plotters

X-Y plotting recorders form a special class of the continuous graphic recorders described above. These devices are designed to display information about variables which can be plotted using Cartesian coordinates. The plotter's ability to accept and record on a single display, input information from independent sources makes it an appropriate instrument for recording movements of several objects, such as a friendly vessel and an enemy vessel. For some kinds of tracking tasks, for graphing the course of simulated torpedos, missiles, or artillery projectiles, and for simulated docking of space capsules, X-Y plotters can provide maximum information.

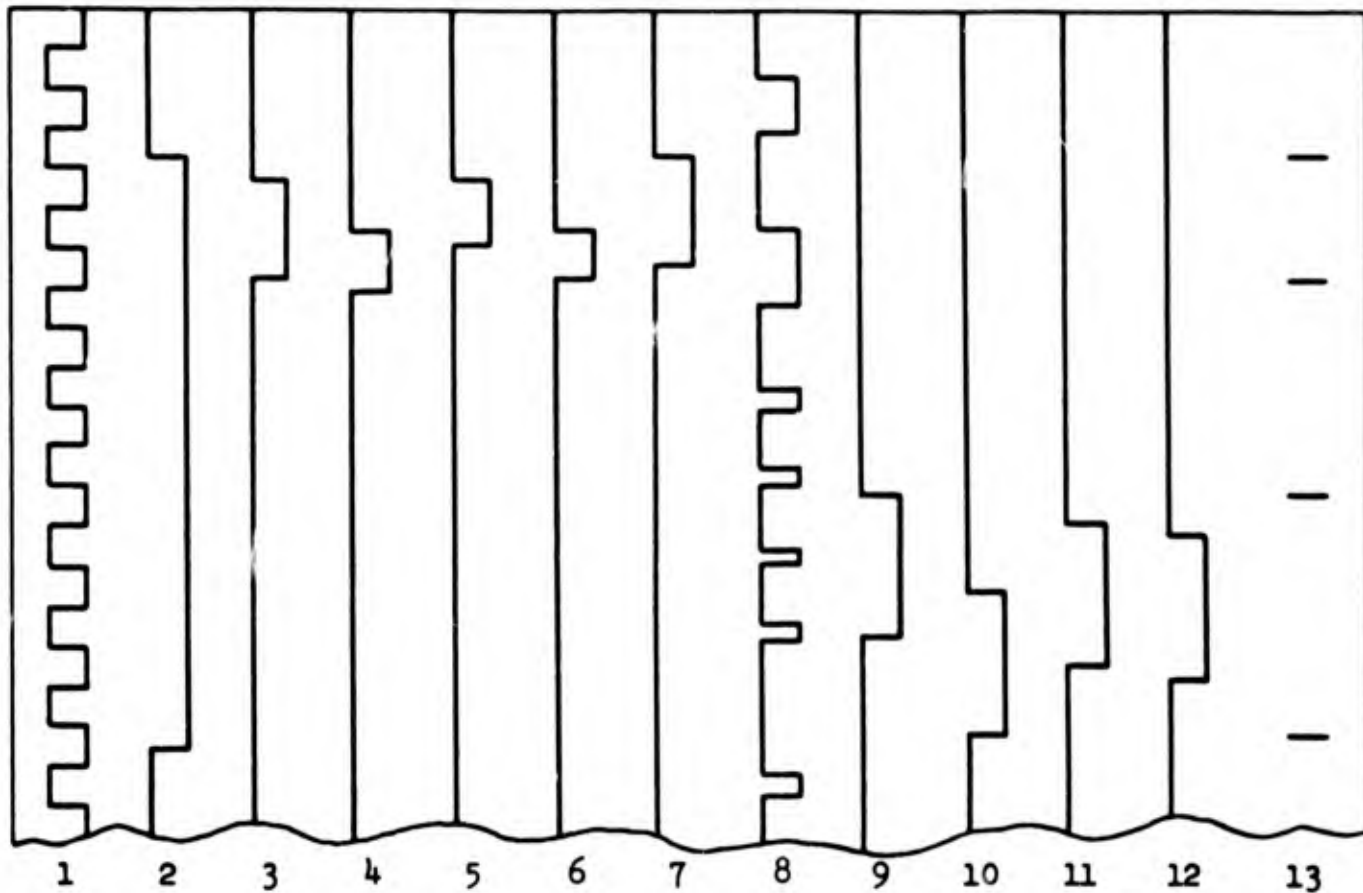


Fig. 4. An event-recorder transcription (hypothetical) of certain components of a simulated submarine attack mission. Time of occurrence and duration of the following events could be recorded during such a training mission:

| Channel(s) | Event |
|------------|---|
| 1 | Time marked in 2-minute intervals. |
| 2 | Call to battle stations. The marker for this channel indicates that battle station status is in effect. |
| 3 - 7 | Changeover to silent running: engines (3); radio communications (4); blow tanks (5); generators (6); sonar (7). |
| 8 | Periscope up. The mark indicates the time that the periscope is above water. |
| 9 - 12 | These channels record the status of forward torpedo tubes: tube #1 armed and ready (9); tube #2 armed and ready (10); etc. The termination of a line could indicate that the torpedo had been fired or that the tube had been disarmed. |
| 13 | Special event marker. |

The point plotter, an X-Y plotter which shows path and trend information by timed point or alphanumeric printout, is adaptable for specific purposes such as simulated war games or large force exercises where many channels are required, but continuous monitoring is not required, to obtain the information needed for performance evaluation.

Picture and Sound Recorders

These devices have been in use for a long time and are pre-eminently capable of providing records of performance. But for use as evaluative devices, the records must be examined by experts who can determine the level of proficiency displayed, or the behaviors under investigation must be matched against known standards previously recorded on another tape or film clip. While the capability is available to record the entire visual and auditory events of a problem for complete real time playback, it is more sensible to attach timers, relays, voice keys, etc. to the recorders and obtain records of only those pertinent segments of standard problems. Since the records obtained are quite literal ones, additional measurements may be taken on these records to achieve more information on the performance under surveillance, without witnessing the actual performance. Costello & Stephen (1963) describe the use of a tape recorder to measure reaction times and also various indices of tracking performance: percentage time-on-target, number of hits, and length of time-on-target during each hit.

Computer Uses in Performance Evaluation

New applications of electronic computers seem to be made almost daily. As our comprehension of their capability goes up, and as their size goes down and the unit costs of using them diminish, computers may be expected to have at least the impact in the field of education and training that they have had in science, industry, and government. Uses of computers in a variety of educational situations are discussed by a number of authors in a book edited by Coulson (1961). Application of computers to military training where simulation is indicated is the specific topic of an Air Force report by Benenati and others (1962). Another Air Force paper (AMRL Memorandum P-40, June 1963) discusses the use of computers in evaluation of pilot and air crew proficiency. The evaluative functions which a computer can perform must probably be considered secondary in importance to the control functions it may also exercise. That is, the automatic recording and assessment of performance in training may not be, in themselves, critical enough objectives to merit computerization of the processes. But where the use of computers as control devices for complex training equipment is indicated, then advantage ought certainly to be taken of the machines' obvious capabilities for performance measurement.

Computer Control. The advantages of computer control can be seen well in the simulation of flight missions, where each different mission phase (pre-flight, take-off, cruise, etc.) has its own characteristics. These segment characteristics, and the performance of each event within the segment, can be described by mathematical and statistical functions, defining the parameters of each mission phase. A number of alternative

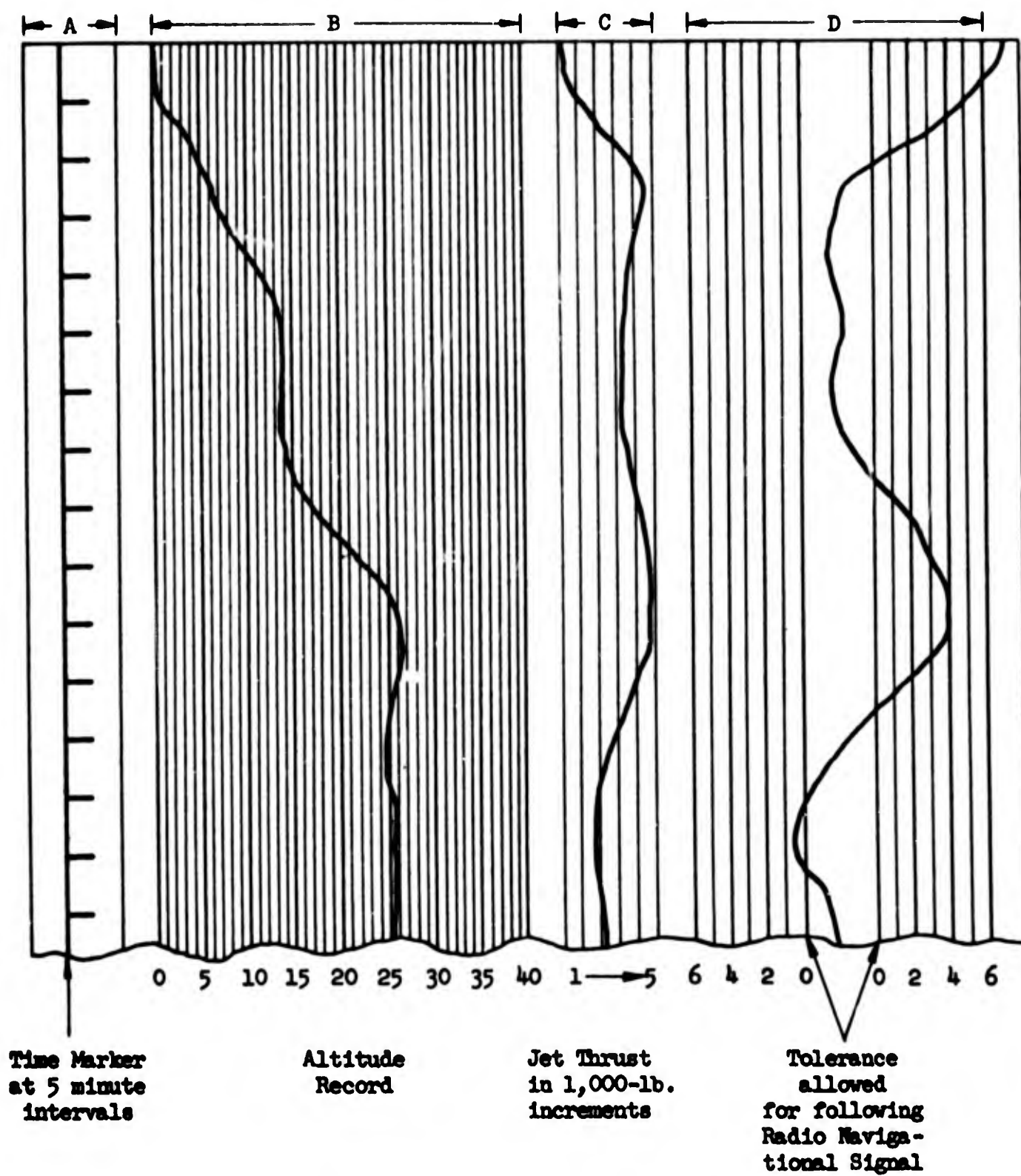


Fig. 5. Hypothetical record of simulated aircraft mission using a continuous ink recorder.

programs for each phase may be written, calling for different behaviors, which the computer can present on different occasions, and in diverse combinations, or for practice in one poorly performed phase, such as landing. The control of a simulated enemy vessel or aircraft can also be programmed and its movements determined by the trainee's behavior. Careful specification of events and detailed analysis in initial programming can produce an ideal control device. Many other control advantages such as instantaneous reaction to changing stimuli, minimum possibility of overloading, large storage capacity of contingent behaviors for special events, proper knowledge and use of all associated equipment, and help in locating equipment malfunctions are all part of the computer's control advantages.

Computer Evaluation. There are also numerous ways to make use of the computer's evaluative capability. Given the type of standardized problems noted above, the computer can place in storage the behaviors elicited by trainees, building a set of normative values which is always current. The computer can also store the "expert" performances of known high performers as another basis for evaluation. Simple percentile or stanine evaluation and output can be taught to the computer and used where applicable. Computer computations are made so rapidly that tasks like the calculation of integrated error rates (in tracking) are completed in a fraction of the time required by a human evaluator. The computer has available for interrogation its own highly accurate internal clock, thus the tasks whose performance is best described in terms of time can easily be measured. The measurement of status situations, characteristic of event recorders, can also be easily accomplished, due to the binary nature of such events. The functions of counting, for frequency information, are likewise simply instrumented, and classification of a more qualitative nature may also be provided. The analog characteristics of continuous recorders and X-Y plotters may either be processed and displayed by the computer, or these display devices can be computer controlled. When very complex motor skills are involved, evaluation of segmented motor performance can be made, and feedback provided at points where the approximation to the standard skill level is poorest. The ability of the computer to weight segments of a mission, provide scores on these segments or on total mission performance, and to provide feedback to the trainee and instructor through numerous output devices, all combine to make the computer a highly valued simulator consideration.

Additional Functions and Considerations. The additional functions of keeping records of trainee behavior, such as providing information on how far the trainee has progressed, what problems he has mastered, what problems must be practiced, and even predictions about when he will be ready for operational assignment, all help to increase the computer's utility.

The consideration of whether to use digital or analog or hybrid equipment should be based primarily on the control functions the computer will be asked to perform. The speed of the analog computer and the accuracy of the digital techniques are not important considerations for performance evaluation since the "slowness" and "inaccuracies" of these devices is considered in terms of physical, not behavioral, stand-

ards. There is a more detailed discussion of the utilization of computers in training/evaluation systems in Section IX of this report.

Human Observers

Although the emphasis in this report has been upon automatic performance evaluation, it has not been intended that the human observer be excluded as an instrument for recording and measuring. At our present level of technological development, there are some areas in which the human observer can make more subtle judgments and more sophisticated evaluations than can any electromechanical instruments. One such area is the interpretation of communications between crew members, leading to an evaluation of some such vaguely-defined quality as "crew cohesiveness." With some systems, for some kinds of missions, psycho-social variables like this seem to have an important influence on system effectiveness. Even though no very precise definition can be given to the variable, and even though its components remain rather mysterious, there may be better-than-chance agreement among human observers concerning degrees to which various crews demonstrate such a quality. Not enough is known yet to program any machine to make the discriminations and judgments involved. Using a human as a recording instrument necessitates no lesser amount of detailed planning than is required with any other recording technique, and most probably more is needed. The human observer/teacher should be not an adjunct, but rather an integral part of the total measurement system.

SECTION VIII

INTERRELATIONSHIPS AMONG PERFORMANCE-MEASUREMENT FACTORS

Each of the factors which is seen as being importantly involved in performance evaluation--behaviors, measures, recording instruments--has been treated separately in order to examine individually the particularly significant features and details of that class of variables. In making decisions about the design and implementation of a program of performance evaluation, interrelationships among the factors must of course be considered. The interrelationships among the variables of concern here (or, at least, their broad compass) might be more clearly apprehended if the factors were represented as forming a matrix comprising three dimensions. They are so shown in Figure 6. This form of presentation makes apparent to the viewer the possible combinations of:

- (a) classes of behavior which encompass specific activities involved in the performance of military jobs and missions,
- (b) classes of measures appropriate to the assessment of individual differences in the performance of those behaviors, and
- (c) types of equipment which can be used to record performances of individuals and crews.

Details have had to be omitted from the matrix as it is shown in Figure 6, partly because of limited space, and partly because of the obvious difficulty of displaying things in more dimensions than are in fact available. Omitted for the first of these reasons, are, then, (1) the action verbs which constitute the lowest level of the behavioral classification scheme (see Table 2, page 20), and (2) the subcategories of measures which fall under the Time and the Accuracy headings of the MEASURES dimension, (see page 22) and (3) several subcategories of devices within the INSTRUMENTS dimension. With the inclusion of the major subcategories (but not considering each of the 50-odd action verbs as a category), the matrix has the dimensions 7 x 8 x 6 (behaviors, measures, instruments, respectively), or a total of 336 cells.

What is omitted from Figure 6 for the second reason above, are entries within the cells of the matrix which might provide information concerning the relationships described by intersections of the three factors. Some of these details are able to be provided by partialing out one dimension of the matrix, and then serving the matrix up, as it were, in separate pieces. This is what has been done in the seven figures which follow.

Figures 7-13 are, in effect, slices of the three-dimensional matrix on Figure 6. Each is a two-way matrix, with classes of mensural indices and types of recording devices as its dimensions. Each has as a parameter one of the classes of behavioral activities in the behavioral classification system. Since each matrix is to be considered separately in terms

[illegible]

Fig. 6. Matrix of behaviors, measures, and instruments relevant to performance evaluation.

of a specified class of activities, each cell actually represents a three-way intersection, though but two dimensions are shown for the matrix. Notations which appear in the cells of the matrices indicate that performance of certain of the specific activities within the behavioral class being considered may properly be assessed by the class of measures described in the heading of the row in which the notation occurs, and that such a measure is best recorded by the type of instrument indicated by the heading of the column in which the notation occurs.

Most often, the notation is a general one (X) which means that the measure appears appropriate to most, if not all, of the specific behaviors within that behavioral category, and that the device shown will provide a record yielding that measure for all those behaviors. In some instances, particular behaviors are noted within the cells formed by the measurement/device intersections. And in some other cases, references are given to more detailed examples of how the measures and the devices may be applied to the assessment of performance of concrete tasks and missions.

To a major extent, the appropriateness of measures and devices to behaviors depends upon a precise and a detailed specification of the behavior. The referent for a behavior--what it is that is being acted on, or with--may easily affect the measure best applied to it. When a notation occurs in more than one column for a given row, a sort of interaction may be indicated. That is, one class of devices may appear best for providing a given type of measure with one (or several) specific behaviors, while another class seems more appropriate for yielding the same type of measure, with some other specific behavior within the same behavioral classification. The system does not prescribe formulae nor recipes for performing evaluations. Given the inherent imprecision of our language, no system should be expected to. But the system does provide guidelines and suggestions for assessment of behavior. The matrices are presented simply as guides to the consideration of possible means of measurement. The notations within them represent the consensus of the authors as to what are appropriate measures and recording media. The possibilities are by no means exhausted, and the relationships shown are in no way definitive.

PERCEPTUAL PROCESSES

Searching For and Receiving Information

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|--|---------------------|-------------------|-----------------|----------|-------------------------|-------|
| | | Event-Type | Continuous-Type | | Audio | Video |
| Time to React to a Signal | Example A X | Detecting | | | | |
| Time to Complete an Activity | Example B X | Inspecting | | | | |
| Total Time | X | | | | | |
| Correct Choice from among Alternatives | | | | | | |
| Correct Sequence of Responding | | Inspecting | | | | |
| Error Frequency | | | | | | |
| Error Amplitude | | | | | | |
| Frequency (or Rate) of Responding | Example B X | | | | | |

Specific Behaviors: Detecting; Inspecting; Observing; Reading; Receiving; Scanning; Surveying

Fig. 7. Measures and instruments appropriate to the assessment of information-seeking activities.

Example A: Timing Detection Behavior

Much of the practice which is given in flight-training devices for one- or two-place aircraft, such as Devices 2F76 and 2F62, which simulate the A4D5 and A4D2N aircraft, emphasizes the pilot's responses to malfunction indications which have been inserted by the instructor. Reaction time to the indication is an important component of the pilot's performance in emergency situations. (Other components are his decision as to appropriate corrective action, and his execution of whatever maneuvers may be required.) It would not be a difficult matter to instrument a flight trainer to record reaction times to malfunction indications. The switch activating the timer would be closed by the instructor at the time a malfunction was inserted, and would be opened either by the pilot's initiating a corrective action, or by a voice key through which he verbally acknowledges the indication. It would probably be a somewhat easier task to instrument the simulator, than it would be to obtain meaningful criterion data for evaluating trainees' performances. The criticalness of reaction time to malfunctions depends upon a number of factors, all of which would have to be taken into account, by weighting or by holding them always constant, in developing normative or absolute standards. In this as in many other instances, obtaining a performance measure is likely more easily accomplished than interpreting that measure in a valid way.

Example B: Timing and Counting Scanning and Surveying Behavior

The Submarine Attack Teacher (Device 21A3) incorporates a periscope which provides trainees with a periscopic view of the surrounding sea. In the simulated problems, the prospective sub commanders must use the periscope just as it would be used in actual combat operations. Minimum scope-up time is obviously to be sought, since it lessens the possibility of detection by aircraft or surface vessels. Timers and counters are appropriate instruments for obtaining data on the trainees' employment of the periscope: the former for measuring total time above water, the latter for counting the number of times the scope was raised and lowered. (The measure may be combined, as mentioned earlier, to yield an index of average scanning time.) The interpretation of these performance data would depend, of course, upon the availability of some standards of comparison. This would mean that normative data would have to have been obtained from other sub commanders and trainees running standard problems, or else that some absolute criterion had been determined.

PERCEPTUAL PROCESSES

Identifying Objects, Actions, Events

| | Timers and Counters | Graphic Recorder's | | Plotters | Transcription Recorders | |
|--|---------------------|--------------------|-----------------|----------|-------------------------|-------|
| | | Event-Type | Continuous-Type | | Audio | Video |
| Time Measures | | | | | | |
| Time to React to a Signal | | | | | | |
| Time to Complete an Activity | X | | | | | |
| Total Time | X | | | | | |
| Accuracy Measures | | | | | | |
| Correct Choice from among Alternatives | | X | | | X | |
| Correct Sequence of Responding | | | | | | |
| Error Frequency | | | | | | |
| Error Amplitude | | | | | | |
| Frequency (or Rate) of Responding | | | | | | |

Specific Behaviors: Discriminating; Identifying; Locating

Fig. 8. Measures and instruments appropriate to the assessment of identification activities.

MEDIATIONAL PROCESSES

Information Processing

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|-------------------|--|-------------------|-----------------|----------------|-------------------------|-------|
| | | Event-Type | Continuous-Type | | Audio | Video |
| Time Measures | Time to React to a Signal | | | | | |
| | Time to Complete an Activity | X | | | | |
| | Total Time | X | | | | |
| | Correct Choice from among Alternatives | | | | | |
| Accuracy Measures | Correct Sequence of Responding | | | | | |
| | Error Frequency | | | | | |
| | Error Amplitude | | Interpolating | Example C X | | |
| | Frequency (or Rate) of Responding | | | | | |

Specific Behaviors: Categorizing; Calculating; Coding; Computing; Interpolating; Itemizing; Tabulating; Translating

Fig. 9. Measures and instruments appropriate to the assessment of information-processing activities.

Example C: Assessing Error-Amplitude with an X-Y Plotter in a Search Mission

In an aircraft's hunt for an enemy submarine, such as with the P2V-5FS full mission simulator, a two-pen X-Y plotter can record the movements of both these variables, within a specified area, through time. The graphical lines upon which such data are typically recorded are analog to the distances involved in the training problem. Such information-processing acts as calculating and interpolating for distances, headings, and patterns of evasive action, can be evaluated against the known true values (scaled down to workable size) as shown on the X-Y recorder.

Presently in use for simulators of the type mentioned above is an X-Y display system (a projected image of the pattern made by a scribe run over a smoked glass slide) which presents the problem characteristics by charting aircraft and submarine movements. The advantages of a different type of X-Y recorder is that it could not only display, but could allow for measurement of the two moving variables. Furthermore, the X-Y display system is unintelligible near the end of a training problem because of the crisscrossing of the display traces. With the X-Y recorder, the range of measurement can be changed as the situation dictates, and a readable record is thus always available.

MEDIATIONAL PROCESSES

Problem Solving and Decision Making

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|-------------------|---|-------------------|---------------------|----------|----------------------------|-------|
| | | Event- Type | Continuous- Type | | Audio | Video |
| Time Measures | Time to React to a Signal | | | | | |
| | Time to Complete an Activity | X | | | X | |
| | Total Time | | | | | |
| | Correct Choice from among Alternatives | | X | | X | |
| Accuracy Measures | Correct Sequence of Responding | | X | | | |
| | Error Frequency | | | | | |
| | Error Amplitude | | | X | | X |
| | Frequency (or Rate) of Responding | | | | | |

Specific Behaviors: Analyzing; Calculating; Choosing; Comparing; Computing; Estimating;
Planning

Fig. 10. Measures and instruments appropriate to the assessment of problem-solving activities.

COMMUNICATION PROCESSES

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|-------------------|--|-------------------|-----------------|----------|-------------------------|-------|
| | | Event-Type | Continuous-Type | | Audio | Video |
| Time Measures | Time to React to a Signal | | | | | |
| | Time to Complete an Activity | | | | X | |
| | Total Time | | | | X | |
| | Correct Choice from among Alternatives | | | | | |
| Accuracy Measures | Correct Sequence of Responding | | | | X | |
| | Error Frequency | | | | | |
| | Error Amplitude | | | | | |
| | Frequency (or Rate) of Responding | | | | X | |

Specific Behaviors: Advising; Answering; Communicating; Directing; Indicating; Informing; Instructing; Requesting; Transmitting

Fig. 11. Measures and instruments appropriate to the assessment of communication-type activities.

MOTOR PROCESSES

Simple/Discrete

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|-------------------|---|-------------------|---------------------|----------|----------------------------|-------|
| | | Event- Type | Continuous- Type | | Audio | Video |
| Time Measures | Time to React to a Signal | X | | | | |
| | Time to Complete an Activity | | | | | |
| | Total Time | | | | | |
| | Correct Choice from among Alternatives | | X | | | X |
| Accuracy Measures | Correct Sequence of Responding | | Example D X | | | |
| | Error Frequency | | | | | |
| | Error Amplitude | | | | | |
| | Frequency (or Rate) of Responding | X | | | | |

Specific Behaviors: Activating; Closing; Connecting; Disconnecting;
Joining; Moving; Pressing; Setting

Fig. 12. Measures and instruments appropriate to the assessment of simple motor activities.

MOTOR PROCESSES

Complex/Continuous

| | Timers and Counters | Graphic Recorders | | Plotters | Transcription Recorders | |
|--|---------------------|-------------------|-----------------|----------|-------------------------|-------|
| | | Event-Type | Continuous-Type | | Audio | Video |
| Time Measures | | | | | | |
| Time to React to a Signal | | | | | | |
| Time to Complete an Activity | X | | | | | |
| Total Time | | | | | | |
| Accuracy Measures | | | | | | |
| Correct Choice from among Alternatives | | | | | | |
| Correct Sequence of Responding | | X | | | | |
| Error Frequency | | X | | | | |
| Error Amplitude | | | X | X | | |
| Frequency (or Rate) of Responding | | | | | | |

Specific Behaviors: Adjusting; Aligning; Regulating; Synchronizing; Tracking

Fig. 13. Measures and instruments appropriate to the assessment of complex motor activities.

Example D: Evaluating Sequential Responding in Simple Motor Behavior

Many military tasks require that responses be made in a certain sequence, i.e., that activities be undertaken in a certain prescribed order. Sequential accuracy is the measure appropriate to evaluation of performance of these jobs; event recorders provide a means of capturing a record of the behavior for assessment. As an example, suppose that some checkout task consists of the seven subtasks listed here, which must be performed in proper sequence, all within a certain time limit:

1. Set POWER switch to "ON"
2. Press PANEL TEST button
3. Set MODE SELECT switch to "TEST"
4. Set VOLTAGE switch to "PRIMARY"
5. Press BATTERY #1 button
6. Press BATTERY #2 button
7. Press RESET button

Figure 14 (top) presents an optimum "model" of the performance of this task, including tolerances, and (bottom) a trainee's performance record. Channel numbers at the left in the figure correspond to the numbering of the subtasks given above. Time is scribed at five-second intervals. On the model record, various graphic devices which would not be produced by the event recorder have been employed, to facilitate a comparison of the trainee's record with the standard. Thus, the bars marking an event's time of occurrence and of cessation have been darkened in the model. The shaded area in channel 2 indicates the time limits within which that subtask must be performed. The start of the shaded areas in channels 5 and 6, indicate the latest times at which those subtasks could be completed. The start of the lined area in channel 7 represents the average time taken to finish that subtask, which ordinarily marks the completion of the total task. The cross-hatched segment of channel 7 shows that the task must be completed before that amount of time has elapsed.

Assessment of such a record by manual means is not easy, even for a simple task of only a minute's duration, as in the example. Job aids such as templates and transparent overlays (see, e.g., Weislogel & Jacobs, 1956) would facilitate making comparisons, but if the records were complicated, or long, or numerous, the comparison task would be tedious and easily subject to errors of observation. In the illustrated record, for example, an error so gross as the trainee's having omitted subtask 4 is not vividly apparent, and any variances in time between obtained and standard performance are very difficult to specify with precision. Once again, it is obvious that obtaining a record of performance is a simpler part of the total evaluation process than is interpreting the obtained record.

In the case of records such as those produced by graphic recorders, the solution seems to lie in the use of computers. Even a small computer could perform, with all the speed and accuracy desired, the kinds of operations involved in comparing times of occurrence of various events (leading to measures of duration and of sequential order), and in com-

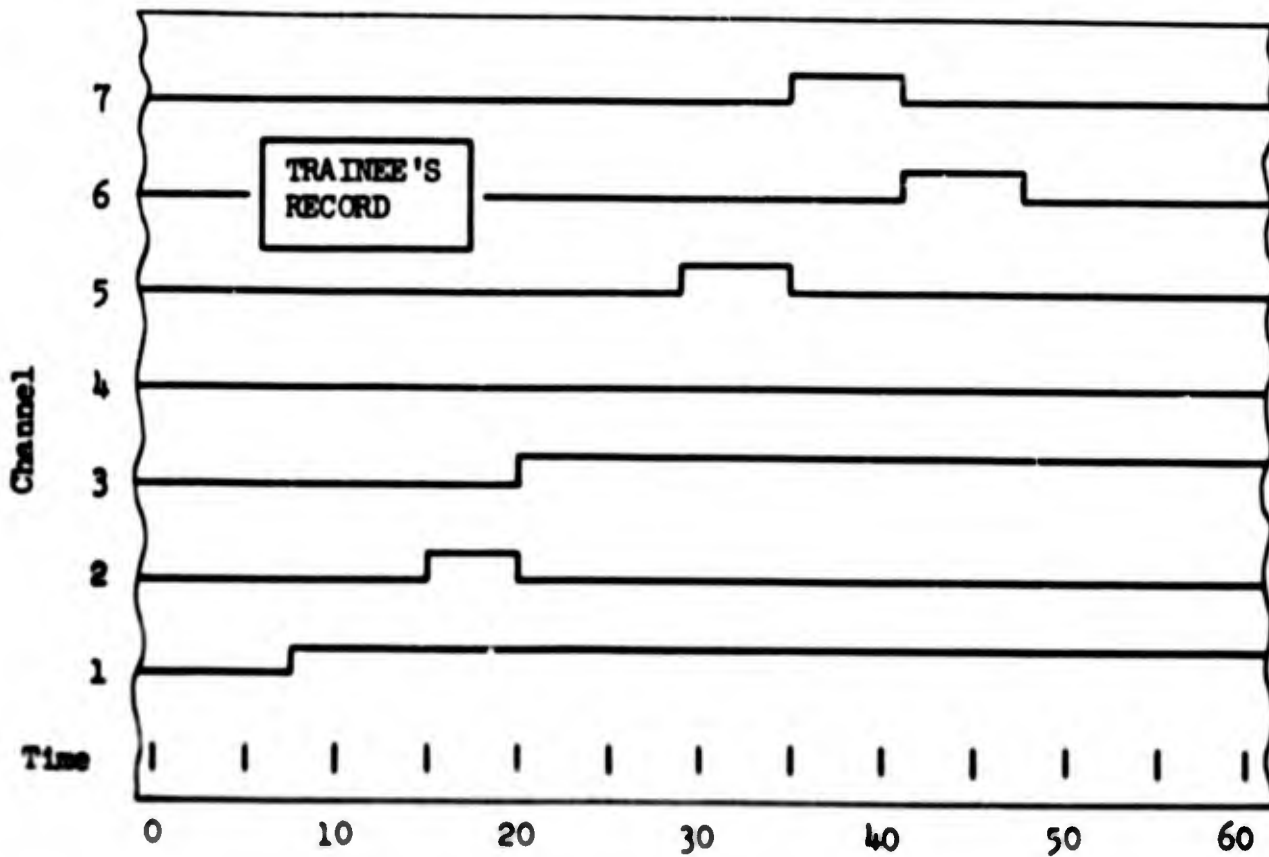
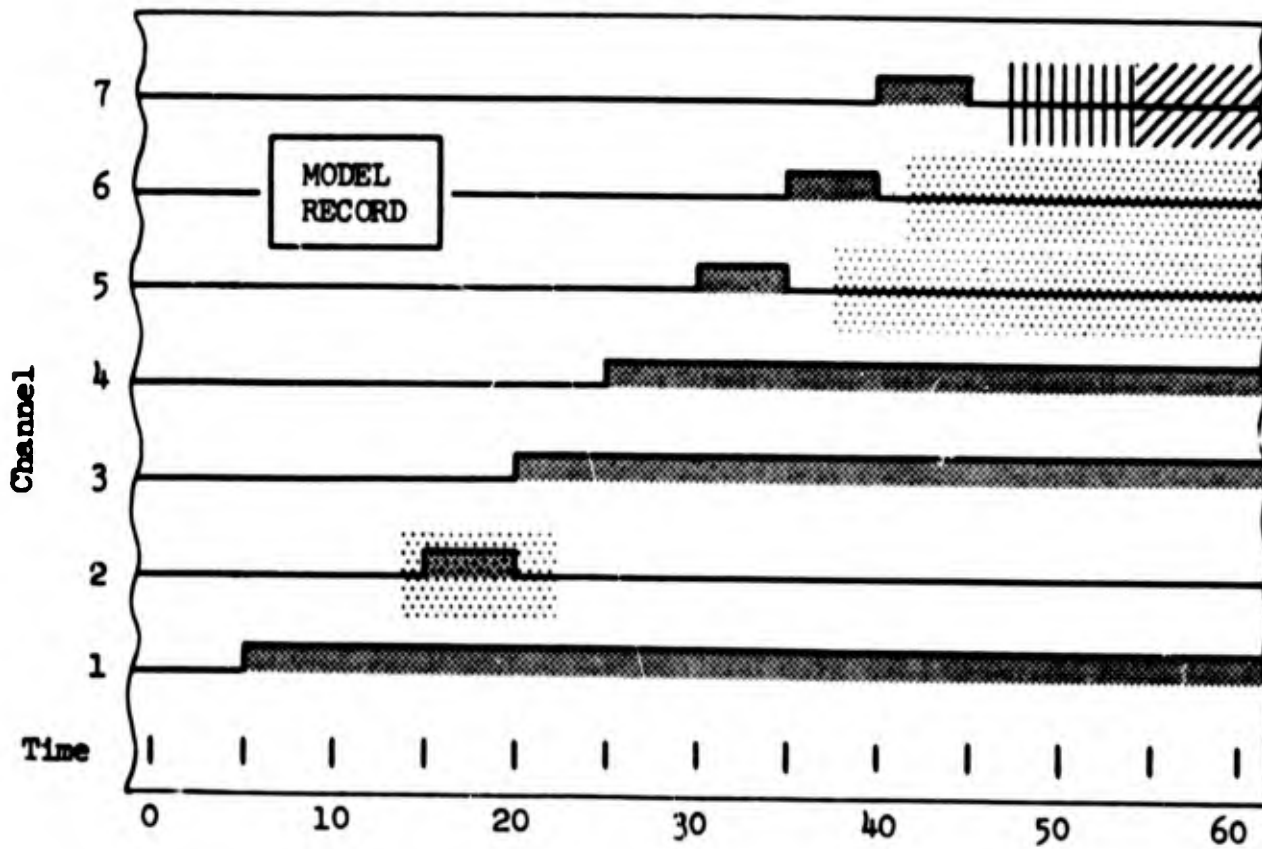


Fig. 14. Event-recorder transcriptions (hypothetical) of the performance of a sequential checkout task.

paring these measures with stored information about other comparable performances. A device would be required which would read the printed record into the computer. It would sense optically graphic lines on the chart and produce electrical signals for the computer. Chart readers are available which will do this. Some modifications might have to be made on them to accommodate records using varying numbers of channels or having some nonstandard feature, but adopting such devices as are presently on the shelf ought not to be a difficult matter. In fact, one-of-a-kind devices of this sort should not be too difficult nor expensive to construct.

SECTION IX

IMPLEMENTING AN AUTOMATIC SCORING SYSTEM

This section is divided into two parts, (a) computers used with flight simulators and (b) computers used with other simulation devices. In both sections, illustrations and the applications of an automatic scoring system are presented. The versatility of the computer as an evaluative instrument is indicated.

Some of the problems which are used for illustrative purposes are fictional in nature, but they are realistic in that they typify some of the military behaviors found in the systems under discussion. This section is intended not as a guide for utilizing automatic equipment, but instead is meant to be an illustration of the accomplishments and changes that are brought about by using this equipment.

A. Computerization of Flight Simulators

This section will first describe the roles of the crew, instructor, and digital control computer in a simulated multi-engine aircraft emergency training flight. Emergency training was chosen for this fictional mission because it is a common simulator mission in which the pilot and crew are called upon to exercise practices and procedures that are infrequently employed in their regular flights, but which must be performed with good speed and perfect accuracy when they are required. The judgments, decisions, and responses involved in the mission, and indeed almost all of the pertinent flight activities which take place, may be computer-rated for effectiveness and acceptability in terms of time, accuracy, and frequency. To do this, the computer needs to have in storage the reference data against which to compare the trainee behavior. These reference data (the criteria) might be based upon the performance of other trainees or the performance of skilled operators (that is, they might consist of a set of norms), or they might represent some absolute standards of performance. These reference data, combined with the computer's capability of calculating percentile or stanine scores when comparing trainee-generated behavior with reference behavior, can provide scoring at every step of the mission.

The device to be discussed requires digital-computer simulation of all the necessary instrumentation for flying operational missions in multi-engine aircraft. The design would incorporate full crew facilities and equipment. High engineering fidelity between the simulator and operational equipment is assumed. The digital control computer would be an advanced version of the experimental and research digital control computer called UDOFT (Universal Digital Operational Flight Trainer, AMRL-TDR-63-133). Some of the advances envisioned would be the use of solid state electronics, an increased programing versatility and an increased immediate storage capacity. The biggest difference, and the one to be presented in the following pages, is the use of the computer not only to control the mission, in terms of presenting stimuli and providing feedback, but to evaluate performance during the simulated flight as well. The examples used to display computer functions are composites of actual multi-engine aircraft emergencies (Ronan, 1954) and the computer

activities discussed are based in part upon systems descriptions of the IBM 1710 Control System.

Input to the computer could be punched cards providing specific commands, or a more sophisticated general program could be written if certain parameters of the mission can be described. The programs would not need to cover entire missions. Greater flexibility could be obtained if separate programs were written for different phases of the mission. For example, flight training might be broken down into mission segments such as these:

1. Pre-flight checkout and start-up
2. Taxi (for take-off and after landing)
3. Take-off
4. Climb/Descent/Maneuvers
5. Cruise
6. Land
7. Post-flight checkout

Each mission phase ought to have several different programs written for it, each calling for somewhat different behaviors, but utilizing standard pre-programmed subroutines. Entire mission simulation would be provided by combining a number of mission segments.

The tabular format that follows attempts to describe the activities of the crew, the instructor and the computer for selected problems.

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|--|--|---|---|
| 1 Pre-flight checkout and start-up. | Pilot and co-pilot proceed through a 40-step sequential pre-flight checkout. This includes switch positioning, visual examinations, and try-out of controls. | Monitors crew inter-communication. Where applicable, instructor provides reference data for checkout (e.g., no oil spots on ground). Instructor may also play the role of a crew member, e.g., co-pilot or navigator. | An "ideal" sequence of pre-flight checkout is available in ready storage (possibly a random access magnetic core memory system), against which the trainee's behavior is evaluated. Time information is provided by interrogation of an internal clock, and the important activities can be given percentile or stanine scores based on deviation from the standard times. Sequence of operations can also be evaluated against the standard, and |

NAVTRADEV CEN 1449-1

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---|---|--|
| | | | the computer may react in preprogramed ways to any missed items of the checkout. Both the instructor and the computer's own electromechanical connections will supply the information about what is and what is not checked out to the comparator section for evaluation. For example, if a check of the fuel gauge on #2 tank was sensed by the computer as having been missed in checkout, the computer might perform these actions: Enter new programed subroutine; Insert EMPTY indication on pilot's #2 fuel tank gauge; Inform instructor of temporary branching sequence. |
| | Pilot sees empty indication on #2 tank, which is known to have been filled. Correct pilot action is to switch to #2 tank in beginning of flight to prevent ending mission using a tank with faulty gauge. | Instructor receives trainee's preflight checkout score from computer. Instructor records own comments on communication processes and non-computer sensed items of the checkout. | |
| 1 | Pilot begins engine start-up procedure for multi-engine aircraft. | Instructor monitors communications and provides reference data. | Computer records time to respond to a signal (the empty indication) and evaluates response to this particular malfunction indication. Computer matches start procedure against the standard in memory. The computer checks engine settings, fuel mixture, spark, and power run-up against the known optimum standards. The computer evaluates deviations from optimum settings and |

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---|--|--|
| | | | displays this information through the output unit (possibly a typewriter printer) to the instructor. |
| 1 | Power and boost applied by pilot. | | The computer inserts a low torque reading and fluctuating cylinder-head temperature for engine #1 on pilot's instrument panel. Computer records time to respond to malfunction indication. |
| 1 | Pilot checks instruments with co-pilot. | Instructor monitors pilot and co-pilot communications. Instructor calls out smoke indication on #1 engine. | Computer inserts FIRE WARNING light for #1 engine on co-pilots instrument panel. |
| 1 | Pilot pulls POWER OFF and FIRE EXTINGUISH switches for #1 engine. | | The computer measures the time of the actions taken to stop an engine fire and provides scoring on the time and procedures involved. |
| | | Instructor prepares mission phase protocol and provides override to start next phase. | An instructor's override at the end of a phase would probably be a four-address instruction to the computer: (a) return all controls to normal indications, (b) weight important scores that were obtained during the phase evaluation, (c) assign an easy-access storage address to the weighted phase score and (d) provide transfer instruction to jump to a new phase program. |

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---|---|--|
| 2 Taxi | Taxiing procedure, brake tests, and tower communications are initiated. | Instructor performs tower role (Computer might even give instructor "weather" for the flight.) | Computer inserts "ZERO" hydraulic brake pressure readings on pilot instrument panel. |
| 2 | Pilot spots "ZERO" hydraulic brake pressure indication and repeatedly tries pressure override switch. | | The computer records time to notice malfunction and accepts or rejects the override, depending on the time measure. If the reaction time to the malfunction indication was too long, then the override is rejected. |
| 2 | Pilot senses override not working and calls for hand pump operation. | Instructor could take crew role and operate hand pump, or could monitor interactions if crew is on board. | The computer matches the time it takes to sense that the override is not effective against a standard, and if the time taken was too long, then the computer refuses the hand pump operation. |
| 2 | Pilot chooses to reverse engines and abort mission. | | Computer follows and evaluates abort procedure. The events involved in this emergency are scored in terms of time until the abort is initiated, time of each preceding action, frequency of responding with override switch and accuracy of the choices made during the emergency. |

NAVTRADEVCEEN 1449-1

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---------------|---|-----------------|
| | | The instructor records evaluative information as printed out from computer, he notes any comments about pilot-crew communications and continues to play tower role. | |

The computer can work effectively with the instructor in cases where checkouts are made that may be primarily visual. For example:

| | | | |
|-------------|--|---|--|
| 5 Cruise | | | The computer inserts an out of tolerance carburetor air temperature indication and alerts instructor to provide time and sequence data which the computer needs. |
| 5 | Pilot calls for instrument check. | Instructor signals computer that malfunction was noticed. | Computer notes and scores time to recognize a malfunction. |
| | Manifold pressure, cylinder head pressure and fuel flow readings are called out. | The instructor informs computer of check-out. | The computer accepts input from instructor and stores data. |
| 5 | Air temperature, intercooler and air plug position checked. | The instructor informs computer of actions taken. | The computer at this point might make the decision that the most probable causes of malfunctions are being checked out first. It could then "pop" out circuit breaker not yet checked. |

NAVTRADEVCEEN 1449-1

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---|--|--|
| 5 | Fuses and circuit breakers checked--mal-function found and corrected. | The instructor prepares a protocol on the communications and search procedures used during the hunt for the malfunction. | The computer senses that the circuit breaker is returned to operational status and notes and rates the time involved in this emergency. The computer also evaluates the decision procedures used in finding the malfunction. |

Some more specialized and intricate evaluative functions can be seen in the following specific examples.

| | | | |
|-------------|---|---|--|
| 5 Cruise | | | The computer fails all gyro's, compasses, and other navigational devices and simulates radio navigational beacon. |
| 5 | Pilot picks up radio navigational signal and begins to fly by beam to get to destination. | Instructor signals computer when pilot homed on radio beam. | The computer records time to regain navigational aid, and records either literally or by sampling, the time on heading (time on target) and the error amplitude. For this and more complex tracking tasks the computer could perform the calculations to get the integrated absolute error rates, useful in evaluating most tracking performance. The computer control of a graphic recorder could display the integrated performance. |

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|---|---|--|
| | | <p>The instructor fills out the protocol on tracking behavior.</p> <p>The instructor plays role of ground flight control operator and informs pilot of thunderheads, giving instructions to change altitude.</p> | <p>The computer informs instructor of violent thunderstorms ahead.</p> |
| 5 | <p>The pilot informs crew that he is climbing from 30,000 to 45,000 feet.</p> | <p>The instructor notes conditions and communications.</p> <p>The instructor continues to monitor mission and make notes of computer evaluation for debriefing.</p> <p>The instructor informs aircraft that he is to hold exactly 45,000 feet as new trip altitude.</p> | <p>The computer monitors and evaluates climb of 15,000 feet for angle of climb, rate of climb, engine settings and mixture, etc. These data are evaluated on the basis of known optimum standards. Computer feeds results to instructor when altitude is obtained.</p> <p>The computer gives instructor traffic information.</p> |

| Mission Segment | Crew Behavior | Instructor Behavior | Computer Events |
|-----------------|--|---------------------|--|
| 5 | The pilot maintains 45,000 foot altitude and watches closely for deviations. | | The computer monitors deviation in altitude and scores performance by time at altitude and error magnitude. To test pilot efficiency in aircraft control, these same measures can be taken as an engine is failed. |

The emergency training procedures described above do not at all tax the computer's limits for speed and versatility. For a somewhat more demanding example of computer control, assume that this same multi-engine aircraft were on a simulated anti-submarine mission to evaluate the crew performance, especially that of the tactical officer, in the hunting and killing of an enemy submarine. In this instance the computer would have to control the simulated submarine input to all the equipment on board, as well as the other functions already mentioned. Such a mission segment might appear as follows:

| | | | |
|--------|--|--|--|
| Search | <p>The tac officer orders SONABUOY dropped.</p> <p>The JEZEBEL operator picks up submarine.</p> <p>The JEZEBEL operator identifies submarine signature.</p> <p>The tac officer orders course change.</p> | <p>The instructor monitors communication, checks coordination on buoy drop.</p> <p>The instructor informs computer of contact.</p> <p>The instructor informs computer.</p> | <p>The computer provides enemy submarine input to detection devices on aircraft.</p> <p>The computer records time to detect a contact.</p> <p>The computer scores identification in terms of class, size, screws, country of origin, etc.</p> <p>The new heading is rated in terms of accuracy of the inference from information so far available.</p> |
|--------|--|--|--|

The functions required of the computer in a hunt and kill exercise require more computer processing time (especially in the last stages of the mission where the submarine is taking evasive action and the aircraft is closing in), but evaluative computations can usually occur on a time-sharing basis with the computer's control functions. An alternative method would be to store the information in some literal fashion, provided that the storage capacity is big enough, and when time is available the events can be brought out from memory and evaluated. The information would be transmitted to the instructor slightly later than would normally be the case.

Instructor Functions when Utilizing Computers. The above discussion mentions the instructor's behaviors in an automatic control and evaluation computerized simulator. It is to be expected that the instructor's role will change with the changing technology. The most noticeable physical change from today's simulators would be at the instructor's station. The considerable array of input and output devices which connect the instructor with the simulator would be somewhat reduced. The facility for communication with the computer is made up of a few clearcut status buttons and an input/output typewriter with which to converse. The station would probably have a coded running "script" of the ongoing events, so the instructor can always find his place in a standard computer-controlled mission. Access to an accurate timing device and an area in which to fill out mission protocols are also needed. The instructor would still be in charge of sequencing events, but now he would act through the computer. All the simulated missions will have been fractionized into segments or phases and the instructor will determine which trainee will have what phase sequence as well as when it should be presented on the basis of a joint computer-instructor decision. The instructor will provide for the interconnections of equipment and devices necessary for the particular training mission, and in so doing will determine what will be on-line and off-line equipment within the system.

The instructor will act as the linkage to the computer for those behaviors not easily discriminated by the computer. For example, a change in course is readily sensed by the computer, but a change in mission destination may not be apparent for some time. The instructor will converse regularly with the computer through his status switches and high speed typewriter and the instructor can have control of the buffer input. The computer output will be recorded by the instructor on standard forms. These forms, or mission protocols will be the basis for debriefing sessions between the instructor and the trainees at the end of the mission. Fig. 15, drawn without the constraints that actual knowledge of a system would have imposed, presents some of the types of information that it is possible to obtain under a computerized evaluative system.

Possible Protocol Items when Utilizing Computers. The protocol identifies the TRAINEE by some code, so that the computer can obtain and store his records. The trainee's GROUP is also identified so that up-to-the-minute knowledge of a squadron's "readiness" or operational ability can also be processed. The INSTRUCTOR should note his own initials. The type of PROBLEM to be run on the simulator is described and, with knowledge of who the trainee is and what problem is to be run, a PROGRAM can be picked

TRAINEE 1-06 GP. 205 INSTRUCTOR DCB

PROBLEM Eng. Trng. PROGRAM A C B O B C NA
1 2 3 4 5 6 7

NOTES Pilot needs practice in T/O & landing emergencies
Previous scores at stations 3, 4. Skip postflight
if time short.

| <u>PHASE</u> | <u>TNG NORM</u> | <u>TNG STD</u> | <u>COMMENTS</u> |
|----------------|-----------------|----------------|---|
| 1 | 8 | 70 | I Good C/O C cont. hydr. Practise |
| 2 | 9 | — | I Procedure pat C — |
| 3 | 5 | 30 | I steep angle of C climb. Carb. of setting rear |
| 4 | 8 | 60 | I Pilot failed to notify C crew of maneuver C Climb too steep |
| 5 | 9 | — | I — C — |
| 6 | 8 | 70 | I C approach too tight |
| 7 | NA | | I C NA |
| TOTAL SCORE | 8 | 50 | |

GENERAL EVALUATIVE INFORMATION:

- Reaction time to malfunction indications
- Identification of malfunction
- General problem solving ability
- Procedural knowledge
- Aircraft control
- Communication information index
- Crew cooperation
- Pilot leadership ability
- Etc.

| E | G | P | NA |
|---|---|---|----|
| | X | | |
| | X | | |
| | | X | |
| X | | | |
| | | X | |
| | X | | |
| | X | | |
| | X | | |

FOR NEXT TNG SESSION:

Continue practice in
T/O & landing maneuvers.

Climb & descent in turns
substandard.

Tracking practice
needed.

Fig. 15. Sample of information protocol to be used with computer evaluation of simulator training.

out by the computer. With the calculation of a training program, the computer could print out NOTES to inform the instructor of any pertinent information about this trainee that should be known.

It should be pointed out that the use of the computer as a record keeper, in addition to its control and evaluation functions, would help maximize computer usage. It would provide the instructor with all the behavioral knowledge he needs on the trainees, and the computer would be able to pick programs that best suit the needs of the trainee. Additionally, the computer would efficiently keep track of which phase programs were already practiced. The letters seen in the PROGRAM blank on the protocol would represent the particular routine for each phase (the numbers below the line) that the computer has picked.

The PHASE column is used to identify the segments of the simulated mission. As mentioned earlier in this section, a mission could be divided into phases such as preflight checkout, taxiing procedure, take-off, etc. TNG. NORM would be a stanine score for the trainee on a particular mission phase, in which he was evaluated against others of his "type" (service rate, length of service, previous training, etc.). TNG. STD. would be a percentile score in which a trainee was matched against some higher standard for that mission phase, such as the performance of expert operators of the equipment. These data (the external standards) would not be expected to be as complete as the TNG. NORM data, thus there would be some blanks under this heading. COMMENTS for each phase are recorded along the sides of the protocol and could be of two types. "I" would be comments of the instructor for those behaviors he has monitored and "C" would be comments the computer has made (through its output printer) to the instructor on the occurrence of or at the computer's cognizance of certain events. The TOTAL SCORE section would provide stanine and percentile evaluation under the TNG. NORM and TNG. STD. columns, but in this case the scores would not be based upon the raw mission phase scores, but instead would be weighted for those mission characteristics most important to success and failure. Generally, high phase scores and low total mission scores would require explanation in the comment section.

The information found in the GENERAL EVALUATIVE INFORMATION section would be of a type that would cut across the mission phases. Generalized ratings on reaction times, procedure following, control of the aircraft, etc., could be made by the computer in such gross evaluative measures as "E" for excellent, "G" for good, and "P" for poor. "NA" would be a non-applicable or no-judgment-possible category. The instructor would also be expected to provide generalized evaluations for such things as information value of communications, crew cooperation, leadership, etc.

The information in the FOR NEXT TNG. SESSION block would be coded along with other important data on the trainee and placed on punched cards to be made available upon demand to the next instructor or for the next training mission.

The use of a protocol containing this type of information should greatly enhance the learning from a debriefing session. The knowledge of results is almost immediate, and the evaluation against other trainees

will provide motivational interest. The criterion imposed by the training standards score permits the trainee to see the amount of improvement required or progress needed, to be operationally prepared.

The computer-generated and instructor-generated comments used in the debriefing session, with the ratings of general knowledges and skills, should give more information, of a more objective kind than is currently available.

The instructor, in computerized flight simulation where control, evaluation and record-keeping functions are to be performed by the computer, is seen to be the "eyes" and "ears" of the computer for those types of inputs which, for one reason or another, are not the direct result of the trainee's activity in the simulator. He is to record independently those behaviors (allocated to him on the basis of careful man-machine analysis) which are best discriminated by a human, and he will aid in providing data to the computer for its use in evaluating combined records into a meaningful learning experience for the trainees, in debriefing sessions at mission's end.

B. Computers Used With Other Simulation Devices

The above discussion was limited to computers in flight simulation, which typically necessitates the use of large, complex, and costly computerization. Some other uses of the computer as an evaluative tool, usually in a smaller and cheaper form should also be discussed. For example, many devices have available now the type of input and output to which a computer could readily be adapted, and if computerization would be an aid in integrating many factors of the mission, then small scale computers should be considered for utilization.

In the submarine attack group of teaching devices, a computer could monitor periscope-up time, degrees of scan, average scan time, and, with knowledge of such things as bow angle, keel depth, and coordinates of enemy ship positions, the computer can also calculate hits and misses of the torpedoes fired. Most of these inputs are available, and if a central processing/integrating computer is provided with standard problems and pre-programed runs, then evaluation functions can be accomplished objectively and accurately with this equipment.

In the field of sonar training, as another example, the computer can assign probability estimates to the identification of various sounds on the standard training tapes that are usually used in this type of training. Evaluation can be accomplished with measurement of the accuracy of identification, time until detection, angle off the bow, going away or coming towards, etc. Segment scores and/or whole mission scores, and training with most of the library of underwater sound sources can be accomplished.

In artillery simulation, with the use of a terrain board for the artillery spotter's training, the coordinates of the simulated artillery shells are plotted to provide puffs of smoke to the trainee to show him where the shells he directed have hit. This same information could be

assimilated by a computer which, with knowledge of the target coordinates, can provide data such as probability of a hit, burst area, error in elevation, range, azimuth, etc. The evaluation would include the characteristics of the guns used, the type of shells used, the closeness to "own" troops, and like information.

Some of the functions which the computer has been described as assuming may sound familiar, because they were discussed in connection with some of the individual recording devices considered earlier. It must be emphasized, though, that the computer's role is more than that of a recorder; it is also a scorer and an evaluator. As was indicated earlier, a behavioral record is not in itself an assessment, though it is essential to the development of a valid assessment. The difference between a record and an evaluation is that the latter implies a comparison of an obtained behavioral record with some sort of performance standard. If a score is to tell us something meaningful, it must be related to some point of reference. In evaluating performance, the essential operations are those of relating a raw record of behavior to some normative or absolute standard, gauging the magnitude of the deviation between the trainee's behavior and the criterion, and stating that value in terms of the effect it may have upon the outcome of the performance. A computer, because of the tremendous amount of information it can keep in its memory, and because of its astonishing speed in performing arithmetical operations, is excellently qualified to perform such evaluative functions. It needs to have the reference data, and of course it needs inputs about trainee's performance.

The computer may be used in conjunction with other recording devices, and in fact, its most effective utilization may come from such integrated applications. The transcription of behavior obtained on a graphic recorder can be scanned by a computer reading device, and evaluated by comparison with a reference chart in the computer's storage. Another way to do this is to use the identical signals that are used to activate the event pens to activate a punched tape device and both the display for immediate data examination and the necessary computer input data can both be recorded simultaneously.

Many of the simulated missions, even though computerized, would also require the display characteristics of an X-Y plotter or continuous graphic recorder. These devices do not require input directly from the simulator, for the computer can control these display devices directly. Additionally, the computer can point out important events by control of an event marker. Again this would provide the quick-look display data by continuous monitoring as well as the evaluation and measurement processes that would be going on internal to the computer.

It is expected that the digital computerization of many simulators will be accomplished in the future. In the rush to use the computer as an all-purpose control mechanism, its performance-evaluation capabilities should surely be employed to their highest advantage.

SECTION X

CONCLUSIONS AND RECOMMENDATIONS

Automated systems of performance measurement in training situations involving the use of simulators, flight trainers, and other complex equipment appear to offer a number of advantages. Among the most obvious are:

1. More objective measures are provided by automatic equipment than by fallible human observers, and with an increase in objectivity, all of the purposes of performance measurement are more directly and more accurately fulfilled.
2. Automatic equipment has the capability of providing evaluation information of a more detailed nature than is easily available by other observational methods.
3. Evaluations may be made very rapidly by high-speed automatic equipment, and this has implications for training effectiveness and for assignment procedures.
4. Training equipment may be used to its maximum capability with the incorporation of measuring instruments of relatively small cost and size.

Decisions regarding the design and implementation of automatic evaluation systems are thought to be importantly influenced by three kinds of variables: (a) classes of behaviors involved in performing typical military tasks; (b) types of measures or mensural indices; and (c) types of behavior-recording instruments and transcription devices. Consideration of interrelationships among these factors will provide some guidance for personnel concerned with evaluation and with the design of training equipment in determining equipment requirements and uses.

It needs emphasizing that the value of measures provided by automatic equipment depends very largely upon (a) the use of standard problems, run under comparable sets of circumstances, (b) the availability of valid criterion data, either in the form of population norms or in terms of some absolute standards of performance, and (c) system performance. The success of performance evaluation depends not only upon the precision with which behavior is recorded, but more basically upon measurement of relevant and important behaviors, whose identification requires that task and mission analyses be carefully made.

NAVTRADEVCEEN 1449-1

Optimal performance-evaluation systems are seen as being developed through the use of electronic (digital) computer equipment. Computers are capable of making the kinds of behavioral comparisons not feasible by other means, especially in interpretation of error-amplitude data such as is appropriate to the assessment of something like tracking performance. Since computers are coming into much wider usage as control devices for training equipment, the use of performance-evaluation capabilities ought always to be an important consideration in the design of training systems.

SECTION XI

SELECTED LIST OF REFERENCES

An extensive but by no means exhaustive list of references pertaining to performance evaluation is given in this selected bibliography. Annotations are provided for those several dozen publications which bear closely upon aspects of the specific topic of the present report.

Askren, W. B., Jr. Bibliography on maintenance personnel performance measurement. Wright-Patterson Air Force Base, Ohio. June 1963. AMRL Memorandum P-45.

Behavioral Sciences Laboratory. The UDOLT flight simulator system. Wright-Patterson Air Force Base, Ohio: 6570th Aerospace Medical Research Laboratories. December 1963. AMRL-TDR-63-133.

The UDOLT (Universal Digital Operational Flight Trainer) system is the first application of a high speed, general purpose digital computer as the control device for flight simulation. The engineering problems of digital control of continuous events, computer speed, and computer programing are discussed. Application of the computer for selected problems is shown for two aircraft.

Benenati, A. T., Hull, R., Korobow, N., & Nienaltowski, W. Development of an automatic monitoring system for flight simulators. Wright-Patterson Air Force Base, Ohio: 6570th Aerospace Medical Research Laboratories. May 1962. MRL-TDR-62-47.

A design study for an automatic monitoring system for flight simulators is presented. System functions discussed are (1) recording and playback and (2) evaluation and scoring. Objective evaluation and scoring is accomplished by comparison of the monitoring parameters to the programmed criteria. Instructor access to this information is provided. The engineering requirements for the entire system are discussed.

Blanchard, R. E. Development of a job sample proficiency test for guided missilemen maintaining the Mark 15 Mod 0 TARTAR missile. Santa Monica, Calif.: Dunlap & Associates, Inc. June 1961. Research Report No. 2.

Brady, J. S. & Daily, A. Evaluation of personnel performance in complex systems. Los Angeles, Calif.: Space Technology Laboratories, Inc. April 1961. ACPL-TM-60-1.

Assessment of personnel performance required the development of a common additive metric for sub-tasks which bear a determinate relationship to over-all mission performance. The authors use this metric to assess total performance quality from behavior samples of specific parameters.

Briggs, L. J., Besnard, G. G., & Walker, E. S. An E-4 fire control system performance test: I. Functional description. Lowry Air Force Base, Colorado: Air Force Personnel and Training Research Center. March 1955. Technical Memorandum ASPRL-TM-55-8.

Brody, A. L. & Weinstock, S. Mathematical theories in performance decision making and learning. A literature review. Wright-Patterson Air Force Base, Ohio. July 1962. MRL-TDR-62-76.

Buckout, R. A bibliography on aircrew proficiency measurement. Wright-Patterson Air Force Base, Ohio: 6570th Aerospace Medical Laboratories. May 1962. MRL-TDR-62-49.

Buddenhagen, T. F. & Wolpin, M. P. A study of visual simulation techniques for astronautical flight training. Wright-Patterson Air Force Base, Ohio. March 1961. WADD Technical Report 60-756.

Costello, C. G. & Stephen, J. The tape recorder as an instrument for measuring time. Amer. J. Psychol., 1963, 76, 324-325.

Cotterman, T. E. Task classification: An approach to partially ordering information on human learning. Wright-Patterson Air Force Base, Ohio. January 1959. WADC Technical Note 58-374.

The author proposes that a taxonomy be developed which would describe task-by-task the nature of some basic learning situations. He suggests that the behavior categories would best be defined by describing the stimulus-response transformations which take place.

Coulson, J. E. (Ed.). Programed learning and computer-based instruction. New York: John Wiley & Sons, Inc. October 1961.

Danneskiold, R. D. Objective scoring procedure for operational flight trainer performance. Port Washington, N. Y.: U. S. Naval Special Devices Center. February 1955. SDC Report No. 999-2-4.

Demaree, R. G., et al. Proficiency of Q-24 radar mechanics: I. Purposes, instruments, and sample of the study. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center. November 1954. Research Bulletin AFPIRC-TR-54-50.

Eckstrand, G. A. & Rockway, M. R. Spacecrew training: A review of progress and prospects. Wright-Patterson Air Force Base, Ohio. December 1961. ASD Technical Report No. 61-721.

Descriptions of astronaut training programs are provided. Various systems utilizing man in space are described (including the Vostok program). Training requirements, present and future, are discussed.

Fattu, N. A. A catalog of trouble shooting tests (Survey of tests developed to December 1956). Bloomington, Indiana: Indiana University, Institute of Educational Research. December 1956. Research Report No. 1.

Feallock, J. B. & Briggs, G. E. A multiman-machine system simulation facility and related research on information-processing and decision-making. Wright-Patterson Air Force Base, Ohio. June 1963.

Fitzpatrick, R. The development of a research program on advanced synthetic electronic type flight simulators. Pittsburgh, Penna.: American Institute for Research. February 1950. Report No. R-GR-AIR-125-50-FR-13.

The purpose of this study was to investigate the use of flight simulators, and to develop research programs for more effective use of the simulators. Military and civilian use of the devices is described and interviews with subject matter experts are included. An outline of proposed research areas is provided.

Fitzpatrick, R., et al. Development of objective flight checks and proficiency measures for use with bomber, reconnaissance and cargo crews. Pittsburgh, Penna.: American Institute for Research. March 1954. R-AIR-A73-54-FR-65.

This report describes the construction of flight checks for normal maneuvers or procedures, emergency procedures, a handbook on the preparation of flight checks, and research related to these data. A discussion of optimum length of flight check, as determined by the reliability for different size check lists, is included.

Flanagan, J. C. The critical incident technique. Psychol. Bull., 1954, 51, 327-328.

Foley, J. P., Jr. Performance testing: Testing for what is real. Wright-Patterson Air Force Base, Ohio. June 1963. AMRL Memorandum P-42.

This report presents the difficulties involved in developing and administering performance examinations and the dangers of depending on written examinations as substitutes for performance examinations. Note is made of the lack of research information on valid substitution of written for performance examinations.

Folley, J. D., Jr., Altman, J. W., Glaser, R., Preston, H. O., & Weislogel, R. L. (Eds.). Human factors methods for system design. Pittsburgh, Penna.: American Institute for Research. 1960. AIR-B90-60-FR-225.

French, R. S. The K-system MAC-1 trouble shooting trainer: I. Developments, design and use. Lackland Air Force Base, Texas: USAF Personnel & Training Research Center. October 1956. Research Report No. 56-119.

This report describes a training device which simulates system failures. The device trains maintenance technicians in the use of logical system analysis techniques in tracking down the source of trouble to a replaceable unit. Later studies showed the effectiveness of the device, which trained as well as more costly bench and off-the-shelf equipment.

Gagné, R. M. Training devices and simulators: Some research issues. Amer. Psychologist, 1954, 9, 95-107.

Gagné, R. M. The analysis of instructional objectives. Paper read at National Education Association Conference, convened 31 March - 2 April 1963, in Washington, D. C.

The author presents his latest thoughts on classification and taxonomy of behaviors. Described is an education-oriented behavior taxonomy of seven categories which differ from one another in methods of optimal acquisition of the identified behavior. Implications of the analysis are found in preparation of syllabuses and other outlines of teaching materials in the fields of education and training.

Glanzer, M., Glaser, R., & Klaus, D. J. The team performance record: An aid for team analysis and team training. Pittsburgh, Penna.: American Institute for Research. December 1956. Technical Report AIR-B6-56-SR-67.

Observations of several hundred Navy teams were analyzed to determine the specific behavior critical in producing effective team performance. Thirteen behavioral categories were formed from these data and a team performance recording form developed. The construction and use of the recording instrument is discussed.

Glaser, R. & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

This chapter discusses the characteristics of proficiency measures and their uses. Definition of the to-be-assessed performance is explained, followed by discussion of the sampling and weighting of component aspects of the performance. Some applications of proficiency measurement are given.

Greer, G. D., Jr., Smith, W. D., & Hatfield, J. L. Improving flight proficiency evaluation in army helicopter pilot training. Fort Rucker, Alabama: U. S. Army Aviation Human Research Unit. May 1962. HumRRO Technical Report No. 77.

The need for standardized flights to afford more objective rater observation was recognized, and a new helicopter pilot proficiency check list was designed. The new form scaled critical performance of essential maneuvers in the standard training flight. Essentially, this did away with instructor bias due to "loose" rating forms and individually determined problems.

Gustafson, H. W. Research on methods of evaluating maintenance proficiency. Lackland Air Force Base, Texas: USAF Personnel Training Research Center. January 1958. Technical Report No. 58-6.

Four exploratory research efforts were initiated to improve performance testing techniques for maintenance personnel. It was concluded that objective performance evaluation of maintenance personnel is feasible. A recommendation for increased technical knowledge in measurement by the instructors is made.

Harris, D. & Mackie, R. R. Factors influencing the use of practical performance tests in the Navy. Los Angeles: Human Factors Research, Inc. 1962. Technical Report 703-1.

Surveys were conducted and analyzed to provide information on the status of practical performance testing in the Navy as it existed in the first part of 1962. Attitudes of and recommendations by training personnel were elicited, about the use of performance tests.

Harter, G. A. & Gain, P. An electronic target simulator for use with operational radar surveillance systems. Wright-Patterson Air Force Base, Ohio. May 1957. WADC Technical Report 57-277.

Highland, R. W. A guide for use in performance testing in Air Force technical schools. Lowry Air Force Base, Colorado. January 1955. ASPRL-TM-55-1.

Hixon, W. C., Harter, G. A., Warren, C. E., & Cowan, J. D., Jr. An electronic radar target simulator for air traffic control studies. Wright-Patterson Air Force Base, Ohio. December 1954. WADC Technical Report 54-569.

Horst, P. The logic of personnel selection and classification. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

Houston, R., Smith, J., & Flexman, R. E. Performance of student pilots flying the T-6 aircraft in primary pilot training. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center. 1954. AFPTRC-TR-54-109.

Human Resources Research Center. Selected measures of proficiency for B-29 mechanics: Studies 2 and 3. Bolling Air Force Base, Washington, D. C. March 1952.

Krasny, L. M. The functional design of a special-purpose digital computer for real-time flight simulation. Wright-Patterson Air Force Base, Ohio. April 1962. MRL Technical Documentary Report 62-39.

Lohrenz, C. A. & Zymet, B. L. Synthesized equipment for ground based radar systems: I. Radar operator training the man, the machine and the simulator. Wright-Patterson Air Force Base, Ohio. October 1961. ASD Technical Report 61-411(I).

McNulty, C. F. Simulation techniques for spacecrew training: State-of-the-art review. Wright-Patterson Air Force Base, Ohio. April 1962. MRL Technical Documentary Report 62-32.

Marks, M. R. Development of human proficiency and performance measures for weapon systems testing. Wright-Patterson Air Force Base, Ohio. December 1961. ASD Technical Report No. 61-733.

This report is primarily concerned with conveying information on the uses of elementary statistics and measurement theory in setting up a workable personnel subsystem test and evaluation program. Different evaluative instruments are examined in terms of applicability and appropriateness to testing situations in the personnel subsystem.

Miller, R. B. Task and part-task trainers and training. Wright-Patterson Air Force Base, Ohio. January 1956. WADD Technical Report No. 56-41.

This report describes the division of total performance requirements of a position into training segments that lend themselves to distinctive types of trainers. Principal variables in the division are phase of learning and time sharing of activities. Risks of improper part-task training are enumerated, and principles are proposed for reducing such risks.

Miller, R. B. Task description and analysis. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

The author introduces the reader to the techniques of task analysis including the system philosophy and the mechanics of task description. The relationship of task analysis to equipment design and training is presented. An order within which to present task information is listed which represents, in essence, a behavioral categorization scheme.

Miller, R. B. & Swain, A. D. Proposal for a new instrument flight training device. Pittsburgh, Penna.: American Institute for Research. August 1952. GR-AIR-A41-52-SR-13.

This document is an application of the senior author's work on training equipment and task analysis. It is a series of recommendations made in different task areas relating to a job of basic instrument flying.

Muckler, F. A., Nygaard, J. E., O'Kelly, L. I., & Williams, A. C., Jr. Psychological variables in the design of flight simulators for training. Wright-Patterson Air Force Base, Ohio. January 1959. WADC Technical Report 56-369.

Nichols, T. F., et al. Performance evaluation of light weapons infantryman (MOS 111.0) graduates of the advanced individual training course (ATP 7-17). Fort Benning, Georgia: U. S. Army Infantry Human Research Unit. December 1961. HumRRO Technical Report 81.

A realistic combat field exercise was developed and administered to recent infantry training graduates. Methods of scoring performance, both objective and subjective are all job specific. Examples of criteria used are: targets fired at, targets hit, choice of cover, choice of weapons, etc.

Operator Training Branch, Behavioral Sciences Laboratory. Considerations in the design of automatic proficiency measurement equipment in simulators. Wright-Patterson Air Force Base, Ohio: 6570th Aerospace Medical Research Laboratories. June 1963. AMRL Memorandum P-40.

This report was written to show the advantages to be gained in performance measurement, by utilizing automatic scoring equipment. Review of the present and future methods shows that conducting automatic proficiency evaluation in a ground training environment could result in more thorough evaluations which could be obtained more reliably, more accurately, more economically and in less time than the current in-flight subjective evaluation methods.

Parker, J. F., Jr. & Downs, Judith E. Selection of training media. Wright-Patterson Air Force Base, Ohio. September 1961. ASD Technical Report 61-473.

This report is designed to assist a training analyst faced with the problem of selecting specific training aids and devices to be used in the support and development of the personnel subsystem of a military system. The effectiveness of various training media in meeting specific training objectives is indicated and justified in terms of available objective evidence on the subject.

Parker, J. F., Jr. & Fleishman, E. A. Prediction of advanced levels of proficiency in a complex tracking task. Wright-Patterson Air Force Base, Ohio. December 1959. WADC Technical Report No. 59-255.

Fifty scores taken from apparatus and paper-and-pencil psychomotor tasks by AFROTC Ss were examined for predictability of terminal tracking proficiency. Fifteen ability factors were identified by factor analysis. These factors accounted for only a small part of the variance in tracking performance. Early proficiency was found to be unrelated to terminal proficiency in tracking tasks.

Reed, L. E., et al. A methodological approach to the analysis and automatic handling of task information for systems in the conceptual phase. Wright-Patterson Air Force Base, Ohio. August 1963. AMRL-TDR-63-78.

This report presents techniques for analyzing and automatically processing task and task-requirements data generated during the conceptual stage of system development. A tryout of the program indicates that the techniques described are useful to human factors specialists in isolating and processing tasks and task requirements for making personnel, training, and training-equipment recommendations.

Richlin, R., Siegel, A. I., & Schultz, D. G. Development and application of a technical behavior check list (TBCL) criterion to the selective emergency service rate (SESR) program for aviation electronics technicians. Applied Psychological Services for Office of Naval Research, Under Contract No. 2279(00). 1960.

Ronan, W. W. Emergency training procedures multi-engine aircraft. Pittsburgh, Penna.: American Institute for Research. March 1954. Interim Report.

Rose, L., Bogan, C. J., & Heaviside, J. B. Instrumentation of flight simulators. Wright-Patterson Air Force Base, Ohio. December 1958. WADC Technical Note 58-295.

Rulon, P. J., et al. Proficiency of Q-24 radar mechanics: II. The performance trouble shooting test. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center. November 1954. Research Bulletin AFPTRC-TR-54-51.

Ryans, D. G. & Fredericksen, N. Performance tests of educational achievement. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Background information is provided for anyone wishing a comprehensive non-technical summary, designed to give the reader a view of what is important in the area of performance evaluation. Included are descriptions of various types of performance tests, their uses, and an explanation of the steps involved in development of the tests.

Smith, J. F., Flexman, R. E., & Houston, R. C. Development of an objective method of recording flight performance. Lackland Air Force Base, Texas: Human Resources Research Center. December 1952. Technical Report 52-15.

Smode, A. F., Gruber, A., & Ely, J. H. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments. Wright-Patterson Air Force Base, Ohio: 6570th Aerospace Medical Research Laboratories. February 1962. MRL-TDR-62-2.

The authors discuss techniques of measurement and measurement theory. Application is made to a multi-man crew space mission, and a matrix of behavioral categories and appropriate measurement classes is developed. Types of devices to obtain the information required are noted.

Stave, A. M. Human factors in the design of automatic programming and recording for trainers: AN/ASG-15-T-1 fire control system trainer. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories. August 1960. WADD-TR-60-558.

Swanson, A. M. Notes on simulator instrumentation for measurement of pilot proficiency. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center. May 1957. AFPTRC Technical Memorandum OL-TM-57-3.

Thomas, R. E., Pritsker, A. B. A., Christner, Charlotte A., Byers, R. H., & Huebner, W. J. The effects of various levels of automation on human operators' performance in man-machine systems. Wright-Patterson Air Force Base, Ohio. February 1961. WADD Technical Report 60-618.

Trygg, Lavon E. & Kelsey, Lucille E. A bibliography of reports issued by the Behavioral Sciences Laboratory: Engineering psychology, training psychology, environmental stress, simulation techniques, and physical anthropology. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories. April 1964.

Van Buskirk, R. C. & Huebner, W. J. Human-initiated malfunctions and system performance evaluation. Wright-Patterson Air Force Base, Ohio. September 1962. AMRL Technical Documentary Report 62-105.

Webber, C. E. Survey study of the use of analog-digital conversion techniques and high-speed digital computer for the treatment of pilot performance in a flight simulator. Urbana, Illinois: University of Illinois, Aviation Psychology Laboratory. January 1958. Research Memorandum Report 58-1.

Webber, C. E. & Adams, J. A. Issues in the use of an analog-digital data system for the measurement of tracking behavior. Urbana, Illinois: University of Illinois, Aviation Psychology Laboratory. April 1960. AFOSR TN-59-528.

Weislogel, R. L. & Jacobs, T. O. A technique for displaying task analysis information. Kirtland Air Force Base, N. M.: Air Force Special Weapons Center. March 1956. AFSWC TN-56-11.

Willis, M. P. Deriving training device implications from learning theory principles, Volume I: Guidelines for training device design, development and use. Port Washington, New York: U. S. Naval Training Device Center, July 1961. NAVTRADEVCEEN 784-1.

Initially, a behavioral classification scheme was developed and within its context, discussions of the theories of Hull, Guthrie, Tolman, Miller, Harlow, Hebb, Estes, Skinner, and Spence were undertaken. Final product was the development and cataloguing of implications from the theories, to the design and use of training devices by the Navy.

Wilson, C. L. On-the-job and operational criteria. Chapter 12, in Robert Glaser (Ed.), Training research and education. Pittsburgh, Penna.: University of Pittsburgh Press, 1962.

Wilson, C. L., Mackie, R. R., Buckner, D. N., Siegel, A. I., & Courtney, D. A manual for use in the preparation and administration of practical performance tests. DDC Document No. AD 98240 (undated).

The document is meant for those non-technical people having need of a guide that will help them build, use, and understand practical performance tests. Elementary material and concepts of performance testing, statistics, the examiner's role and test conditions are presented.

Woolman, M. On-site training of guided missile operators. Washington, D. C.: Training Methods Division, Human Resources Research Office. August 1960. HumRRO Technical Report 64.

This study was concerned with developing, testing, and evaluating methods of training guided-missile operators at the operational site. The principal experimental group in the study was an operational-context training group. Evaluation showed this method of learning the tasks to be a superior one.

BLANK PAGE